Proceedings of the Linux Symposium

July 21st–24th, 2004
Ottawa, Ontario
Canada

Contents

Conference Organizers

Andrew J. Hutton, *Steamballoon, Inc.*
Stephanie Donovan, *Linux Symposium*
C. Craig Ross, *Linux Symposium*

Review Committee

Jes Sorensen, *Wild Open Source, Inc.*
Matt Domsch, *Dell*
Gerrit Huizenga, *IBM*
Matthew Wilcox, *Hewlett-Packard*
Dirk Hohndel, *Intel*
Val Henson, *Sun Microsystems*
Jamal Hadi Salimi, *Znyx*
Andrew Hutton, *Steamballoon, Inc.*

Proceedings Formatting Team

John W. Lockhart, *Red Hat, Inc.*

TCP Connection Passing

Werner Almesberger
werner@almesberger.net

Abstract

tcpcp is an experimental mechanism that allows cooperating applications to pass ownership of TCP connection endpoints from one Linux host to another one. tcpcp can be used between hosts using different architectures and does not need the other endpoint of the connection to cooperate (or even to know what's going on).

1 Introduction

When designing systems for load-balancing, process migration, or fail-over, there is eventually the point where one would like to be able to "move" a socket from one machine to another one, without losing the connection on that socket, similar to file descriptor passing on a single host. Such a move operation usually involves at least three elements:

1. Moving any application space state related to the connection to the new owner. E.g. in the case of a Web server serving large static files, the application state could simply be the file name and the current position in the file.

2. Making sure that packets belonging to the connection are sent to the new owner of the socket. Normally this also means that the previous owner should no longer receive them.

3. Last but not least, creating compatible network state in the kernel of the new connection owner, such that it can resume the communication where the previous owner left off.

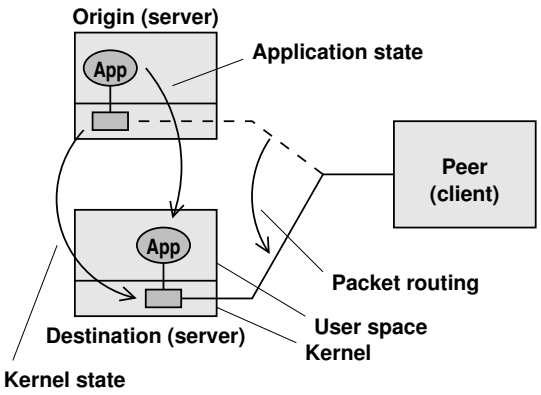


Figure 1: Passing one end of a TCP connection from one host to another.

Figure 1 illustrates this for the case of a client-server application, where one server passes ownership of a connection to another server. We shall call the host from which ownership of the connection endpoint is taken the *origin*, the host to which it is transferred the *destination*, and the host on the other end of the connection (which does not change) the *peer*.

Details of moving the application state are beyond the scope of this paper, and we will only sketch relatively simple examples. Similarly, we will mention a few ways for how the redirection in the network can be accomplished, but without going into too much detail.

The complexity of the kernel state of a network connection, and the difficulty of moving this state from one host to another, varies greatly with the transport protocol being used. Among the two major transport protocols of the Internet, UDP [1] and TCP [2], the latter clearly presents more of a challenge in this regard. Nevertheless, some issues also apply to UDP.

tcpcp (TCP Connection Passing) is a proof of concept implementation of a mechanism that allows applications to transport the kernel state of a TCP endpoint from one host to another, while the connection is established, and without requiring the peer to cooperate in any way. tcpcp is not a complete process migration or load-balancing solution, but rather a building block that can be integrated into such systems. tcpcp consists of a kernel patch (at the time of writing for version 2.6.4 of the Linux kernel) that implements the operations for dumping and restoring the TCP connection endpoint, a library with wrapper functions (see Section 3), and a few applications for debugging and demonstration.

The project's home page is at `http://tcpcp.sourceforge.net/`

The remainder of this paper is organized as follows: this section continues with a description of the context in which connection passing exists. Section 2 explains the connection passing operation in detail. Section 3 introduces the APIs tcpcp provides. The information that defines a TCP connection and its state is described in Section 4. Sections 5 and 6 discuss congestion control and the limitations TCP imposes on checkpointing. Security implications of the availability and use of tcpcp are examined in Section 7. We conclude with an outlook on future direction the work on tcpcp will take in Section 8, and the conclusions in Section 9.

The excellent "TCP/IP Illustrated" [3] is recommended for readers who wish to refresh their memory of TCP/IP concepts and terminology.

1.1 There is more than one way to do it

tcpcp is only one of several possible methods for passing TCP connections among hosts. Here are some alternatives:

In some cases, the solution is to avoid passing the "live" TCP connection, but to terminate the connection between the origin and the peer, and rely on higher protocol layers to re-establish a new connection between the destination and the peer. Drawbacks of this approach include that those higher layers need to know that they have to re-establish the connection, and that they need to do this within an acceptable amount of time. Furthermore, they may only be able to do this at a few specific points during a communication.

The use of HTTP redirection [4] is a simple example of connection passing above the transport layer.

Another approach is to introduce an intermediate layer between the application and the kernel, for the purpose of handling such redirection. This approach is fairly common in process migration solutions, such as Mosix [5], MIGSOCK [6], etc. It requires that the peer be equipped with the same intermediate layer.

1.2 Transparency

The key feature of tcpcp is that the peer can be left completely unaware that the connection is passed from one host to another. In detail, this means:

- The peer's networking stack can be used "as is," without modification and without requiring non-standard functionality
- The connection is not interrupted

- The peer does not have to stop sending
- No contradictory information is sent to the peer
- These properties apply to all protocol layers visible to the peer

Furthermore, tcpcp allows the connection to be passed at any time, without needing to synchronize the data stream with the peer.

The kernels of the hosts between which the connection is passed both need to support tcpcp, and the application(s) on these hosts will typically have to be modified to perform the connection passing.

1.3 Various uses

Application scenarios in which the functionality provided by tcpcp could be useful include load balancing, process migration, and fail-over.

In the case of load balancing, an application can send connections (and whatever processing is associated with them) to another host if the local one gets overloaded. Or, one could have a host acting as a dispatcher that may perform an initial dialog and then assigns the connection to a machine in a farm.

For process migration, tcpcp would be invoked when moving a file descriptor linked to a socket. If process migration is implemented in the kernel, an interface would have to be added to tcpcp to allow calling it in this way.

Fail-over is tricker, because there is normally no prior indication when the origin will become unavailable. We discuss the issues arising from this in more detail in Section 6.

2 Passing the connection

Figure 2 illustrates the connection passing procedure in detail.

1. The application at the origin initiates the procedure by requesting retrieval of what we call the *Internal Connection Information* (ICI) of a socket. The ICI contains all the information the kernel needs to recreate a TCP connection endpoint
2. As a side-effect of retrieving the ICI, tcpcp *isolates* the connection: all incoming packets are silently discarded, and no packets are sent. This is accomplished by setting up a per-socket filter, and by changing the output function. Isolating the socket ensures that the state of the connection being passed remains stable at either end.
3. The kernel copies all relevant variables, plus the contents of the out-of-order and send/retransmit buffers to the ICI. The out-of-order buffer contains TCP segments that have not been acknowledged yet, because an earlier segment is still missing.
4. After retrieving the ICI, the application empties the receive buffer. It can either process this data directly, or send it along with the other information, for the destination to process.
5. The origin sends the ICI and any relevant application state to the destination. The application at the origin keeps the socket open, to ensure that it stays isolated.
6. The destination opens a new socket. It may then bind it to a new port (there are other choices, described below).

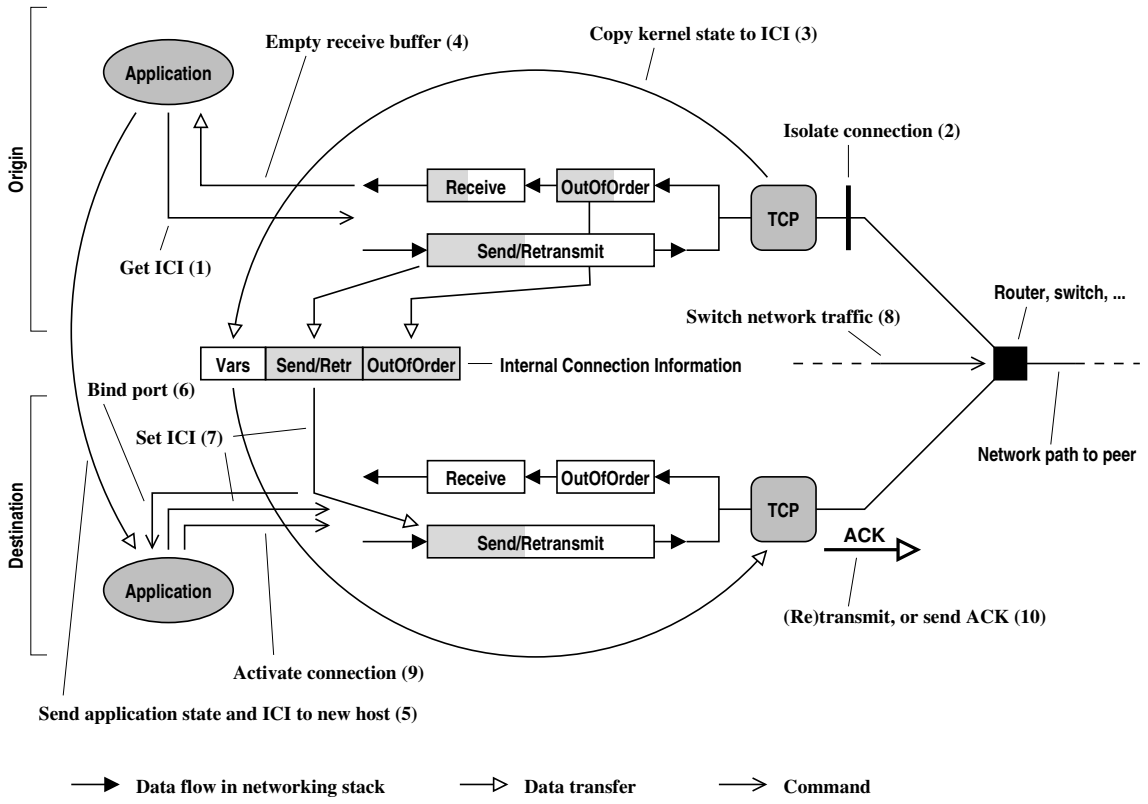


Figure 2: Passing a TCP connection endpoint in ten easy steps.

7. The application at the destination now sets the ICI on the socket. The kernel creates and populates the necessary data structures, but does not send any data yet. The current implementation makes no use of the out-of-order data.

8. Network traffic belonging to the connection is redirected from the origin to the destination host. Scenarios for this are described in more detail below. The application at the origin can now close the socket.

9. The application at the destination makes a call to *activate* the connection.

10. If there is data to transmit, the kernel will do so. If there is no data, an otherwise empty ACK segment (like a window probe) is sent to wake up the peer.

Note that, at the end of this procedure, the socket at the destination is a perfectly normal TCP endpoint. In particular, this endpoint can be passed to another host (or back to the original one) with tcpcp.

2.1 Local port selection

The local port at the destination can be selected in three ways:

- The destination can simply try to use the same port as the origin. This is necessary if no address translation is performed on the connection.

- The application can bind the socket before setting the ICI. In this case, the port in the ICI is ignored.

- The application can also clear the port information in the ICI, which will cause the socket to be bound to any available port. Compared to binding the socket before setting the ICI, this approach has the advantage of using the local port number space much more efficiently.

The choice of the port selection method depends on how the environment in which tcpcp operates is structured. Normally, either the first or the last method would be used.

2.2 Switching network traffic

There are countless ways for redirecting IP packets from one host to another, without help from the transport layer protocol. They include redirecting part of the link layer, ingenious modifications of how link and network layer interact [7], all kinds of tunnels, network address translation (NAT), etc.

Since many of the techniques are similar to network-based load balancing, the Linux Virtual Server Project [8] is a good starting point for exploring these issues.

While a comprehensive study of this topic if beyond the scope of this paper, we will briefly sketch an approach using a static route, because this is conceptually straightforward and relatively easy to implement.

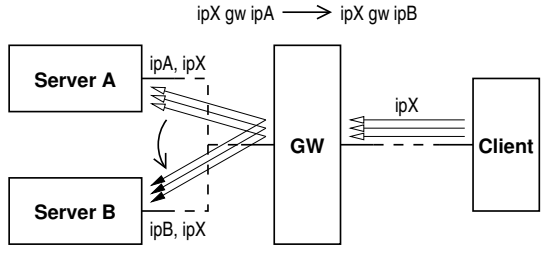


Figure 3: Redirecting network traffic using a static route.

The scenario shown in Figure 3 consists of two servers **A** and **B**, with interfaces with the IP addresses **ipA** and **ipB**, respectively. Each server also has a virtual interface with the address **ipX**. **ipA**, **ipB**, and **ipX** are on the same subnet, and also the gateway machine has an interface on this subnet.

At the gateway, we create a static route as follows:

`route add` *ipX* `gw` *ipA*

When the client connects to the address **ipX**, it reaches host **A**. We can now pass the connection to host **B**, as outlined in Section 2. In Step 8, we change the static route on the gateway as follows:

```
route del ipX
route add ipX gw ipB
```

One major limitation of this approach is of course that this routing change affects all connections to **ipX**, which is usually undesirable. Nevertheless, this simple setup can be used to demonstrate the operation of tcpcp.

3 APIs

The API for tcpcp consists of a low-level part that is based on getting and setting socket options, and a high-level library that provides convenient wrappers for the low-level API.

We mention only the most important aspects of both APIs here. They are described in more detail in the documentation that is included with tcpcp.

3.1 Low-level API

The ICI is retrieved by getting the `TCP_ICI` socket option. As a side-effect, the connection is isolated, as described in Section 2. The application can determine the maximum ICI size

for the connection in question by getting the `TCP_MAXICISIZE` socket option.

Example:

```
void *buf;
int ici_size;
size_t size = sizeof(int);

getsockopt(s,SOL_TCP,TCP_MAXICISIZE,
    &ici_size,&size);
buf = malloc(ici_size);
size = ici_size;
getsockopt(s,SOL_TCP,TCP_ICI,
    buf,&size);
```

The connection endpoint at the destination is created by setting the `TCP_ICI` socket option, and the connection is activated by "setting" the `TCP_CP_FN` socket option to the value `TCPCP_ACTIVATE`.[1]

Example:

```
int sub_function = TCPCP_ACTIVATE;

setsockopt(s,SOL_TCP,TCP_ICI,
    buf,size);
/* ... */
setsockopt(s,SOL_TCP,TCP_CP_FN,
    &sub_function,
    sizeof(sub_function));
```

3.2 High-level API

These are the most important functions provided by the high-level API:

```
void *tcpcp_get(int s);
int tcpcp_size(const void *ici);
int tcpcp_create(const void *ici);
int tcpcp_activate(int s);
```

[1] The use of a multiplexed socket option is admittedly ugly, although convenient during development.

`tcpcp_get` allocates a buffer for the ICI, and retrieves that ICI (isolating the connection as a side-effect). The amount of data in the ICI can be queried by calling `tcpcp_size` on it.

`tcpcp_create` sets an ICI on a socket, and `tcpcp_activate` activates the connection.

4 Describing a TCP endpoint

In this section, we describe the parameters that define a TCP connection and its state. tcpcp collects all the information it needs to re-create a TCP connection endpoint in a data structure we call *Internal Connection Information* (ICI).

The ICI is portable among systems supporting tcpcp, irrespective of their CPU architecture.

Besides this data, the kernel maintains a large number of additional variables that can either be reset to default values at the destination (such as congestion control state), or that are only rarely used and not essential for correct operation of TCP (such as statistics).

4.1 Connection identifier

Each TCP connection in the global Internet or any private internet [9] is uniquely identified by the IP addresses of the source and destination host, and the port numbers used at both ends.

tcpcp currently only supports IPv4, but can be extended to support IPv6, should the need arise.

4.2 Fixed data

A few parameters of a TCP connection are negotiated during the initial handshake, and remain unchanged during the life time of the connection. These parameters include whether window scaling, timestamps, or selective acknowledgments are used, the number of bits by

Connection identifier	
`ip.v4.ip_src`	IPv4 address of the host on which the ICI was recorded (source)
`ip.v4.ip_dst`	IPv4 address of the peer (destination)
`tcp_sport`	Port at the source host
`tcp_dport`	Port at the destination host
Fixed at connection setup	
`tcp_flags`	TCP flags (window scale, SACK, ECN, etc.)
`snd_wscale`	Send window scale
`rcv_wscale`	Receive window scale
`snd_mss`	Maximum Segment Size at the source host
`rcv_mss`	MSS at the destination host
Connection state	
`state`	TCP connection state (e.g. ESTABLISHED)
Sequence numbers	
`snd_nxt`	Sequence number of next new byte to send
`rcv_nxt`	Sequence number of next new byte expected to receive
Windows (flow-control)	
`snd_wnd`	Window received from peer
`rcv_wnd`	Window advertised to peer
Timestamps	
`ts_gen`	Current value of the timestamp generator
`ts_recent`	Most recently received timestamp

Table 1: TCP variables recorded in tcpcp's Internal Connection Information (ICI) structure.

which the window is shifted, and the maximum segment sizes (MSS).

These parameters are used mainly for sanity checks, and to determine whether the destination host is able to handle the connection. The received MSS continues of course to limit the segment size.

4.3 Sequence numbers

The sequence numbers are used to synchronize all aspects of a TCP connection.

Only the sequence numbers we expect to see in the network, in either direction, are needed when re-creating the endpoint. The kernel uses several variables that are derived from these sequence numbers. The values of these variables either coincide with `snd_nxt` and `rcv_nxt` in the state we set up, or they can be calculated by examining the send buffer.

4.4 Windows (flow-control)

The (flow-control) window determines how much more data can be sent or received without overrunning the receiver's buffer.

The window the origin received from the peer is also the window we can use after re-creating the endpoint.

The window the origin advertised to the peer defines the minimum receive buffer size at the destination.

4.5 Timestamps

TCP can use timestamps to detect old segments with wrapped sequence numbers [10]. This mechanism is called *Protect Against Wrapped Sequence numbers* (PAWS).

Linux uses a global counter (`tcp_time_stamp`) to generate local timestamps. If a moved connection were to use the counter at the new host, local round-trip-time calculation may be confused when receiving timestamp replies from the previous connection, and the peer's PAWS algorithm will discard segments if timestamps appear to have jumped back in time.

Just turning off timestamps when moving the connection is not an acceptable solution, even though [10] seems to allow TCP to just stop sending timestamps, because doing so would bring back the problem PAWS tries to solve in the first place, and it would also reduce the accuracy of round-trip-time estimates, possibly degrading the throughput of the connection.

A more satisfying solution is to synchronization the local timestamp generator. This is accomplished by introducing a per-connection timestamp offset that is added to the value of `tcp_time_stamp`. This calculation is hidden in the macro `tp_time_stamp(tp)`, which just becomes `tcp_time_stamp` if the kernel is configured without tcpcp.

The addition of the timestamp offset is the only major change tcpcp requires in the existing TCP/IP stack.

4.6 Receive buffers

There are two buffers at the receiving side: the buffer containing segments received out-of-order (see Section 2), and the buffer with data that is ready for retrieval by the application. tcpcp currently ignores both buffers: the out-of-order buffer is copied into the ICI, but not used when setting up the new socket. Any data in the receive buffer is left for the application to read and process.

4.7 Send buffer

The send and retransmit buffer contains data that is no longer accessible through the socket API, and that cannot be discarded. It is therefore placed in the ICI, and used to populate the send buffer at the destination.

4.8 Selective acknowledgments

In Section 5 of [11], the use of inbound SACK information is left optional. tcpcp takes advantage of this, and neither preserves SACK information collected from inbound segments, nor the history of SACK information sent to the peer.

Outbound SACKs convey information about the receiver's out-of-order queue. Fortunately, [11] declares this information as purely advisory. In particular, if reception of data has been acknowledged with a SACK, this does not imply that the receiver has to remember having done so. First, it can request retransmission of this data, and second, when constructing new SACKs, the receiver is encouraged to include information from previous SACKs, but is under no obligation to do so.

Therefore, while [11] discourages losing SACK information, doing so does not violate its requirements.

Losing SACK information may temporarily degrade the throughput of the TCP connection. This is currently of little concern, because tcpcp forces the connection into slow start, which has even more drastic performance implications.

SACK recovery may need to be reconsidered once tcpcp implements more sophisticated congestion control.

4.9 Other data

The TCP connection state is currently always ESTABLISHED. It may be useful to also allow passing connections in earlier states, e.g. SYN_RCVD. This is for further study.

Congestion control data and statistics are currently omitted. The new connection starts with slow-start, to allow TCP to discover the characteristics of the new path to the peer.

5 Congestion control

Most of the complexity of TCP is in its congestion control. tcpcp currently avoids touching congestion control almost entirely, by setting the destination to slow start.

This is a highly conservative approach that is appropriate if knowing the characteristics of the path between the origin and the peer does not give us any information on the characteristics of the path between the destination and the peer, as shown in the lower part of Figure 4.

However, if the characteristics of the two paths can be expected to be very similar, e.g. if the hosts passing the connection are on the same LAN, better performance could be achieved by allowing tcpcp to resume the connection at or nearly at full speed.

Re-establishing congestion control state is for further study. To avoid abuse, such an operation can be made available only to sufficiently trusted applications.

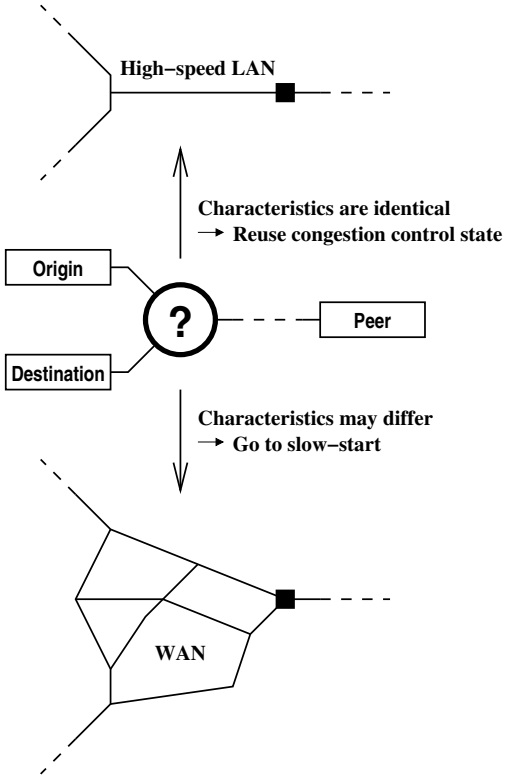


Figure 4: Depending on the structure of the network, the congestion control state of the original connection may or may not be reused.

6 Checkpointing

tcpcp is primarily designed for scenarios, where the old and the new connection owner are both functional during the process of connection passing.

A similar usage scenario would if the node owning the connection occasionally retrieves ("checkpoints") the momentary state of the connection, and after failure of the connection owner, another node would then use the checkpoint data to resurrect the connection.

While apparently similar to connection passing, checkpointing presents several problems which we discuss in this section. Note that this is speculative and that the current implementation of tcpcp does not support any of the exten-

sions discussed here.

We consider the send and receive flow of the TCP connection separately, and we assume that sequence numbers can be directly translated to application state (e.g. when transferring a file, application state consists only of the actual file position, which can be trivially mapped to and from TCP sequence numbers). Furthermore, we assume the connection to be in ESTABLISHED state at both ends.

6.1 Outbound data

One or more of the following events may occur between the last checkpoint and the moment the connection is resurrected:

- the sender may have enqueued more data
- the receiver may have acknowledged more data
- the receiver may have retrieved more data, thereby growing its window

Assuming that no additional data has been received from the peer, the new sender can simply re-transmit the last segment. (Alternatively, `tcp_xmit_probe_skb` might be useful for the same purpose.) In this case, the following protocol violations can occur:

- The sequence number may have wrapped. This can be avoided by making sure that a checkpoint is never older than the Maximum Segment Lifetime (MSL)[2], and that less than 2^{31} bytes are sent between checkpoints.
- If using PAWS, the timestamp may be below the last timestamp sent by the old sender. The best solution for avoiding this is probably to tightly synchronize clock on the old and the new connection owner, and to make a conservative estimate of the number of ticks of the local timestamp clock that have passed since taking the checkpoint. This assumes that the timestamp clock ticks roughly in real time.

Since new data in the segment sent after resurrecting the connection cannot exceed the receiver's window, the only possible outcomes are that the segment contains either new data, or only old data. In either case, the receiver will acknowledge the segment.

Upon reception of an acknowledgment, either in response to the retransmitted segment, or from a packet in flight at the time when the connection was resurrected, the sender knows how far the connection state has advanced since the checkpoint was taken.

If the sequence number from the acknowledgment is below `snd_nxt`, no special action is necessary. If the sequence number is above `snd_nxt`, the sender would exceptionally treat this as a valid acknowledgment.[3]

As a possible performance improvement, the sender may notify the application once a new sequence number has been received, and the application could then skip over unnecessary data.

6.2 Inbound data

The main problem with checkpointing of incoming data is that TCP will acknowledge data that has not yet been retrieved by the application. Therefore, checkpointing would have to delay outbound acknowledgments until the application has actually retrieved them, and has

[2] [2] specifies a MSL of two minutes.

[3] Note that this exceptional condition does not necessarily have to occur with the first acknowledgment received.

checkpointed the resulting state change.

To intercept all types of ACKs, `tcp_transmit_skb` would have to be changed to send `tp->copied_seq` instead of `tp->rcv_nxt`. Furthermore, a new API function would be needed to trigger an explicit acknowledgment after the data has been stored or processed.

Putting acknowledges under application control would change their timing. This may upset the round-trip time estimation of the peer, and it may also cause it to falsely assume changes in the congestion level along the path.

7 Security

tcpcp bypasses various sets of access and consistency checks normally performed when setting up TCP connections. This section analyzes the overall security impact of tcpcp.

7.1 Two lines of defense

When setting TCP_ICI, the kernel has no means of verifying that the connection information actually originates from a compatible system. Users may therefore manipulate connection state, copy connection state from arbitrary other systems, or even synthesize connection state according to their wishes. tcpcp provides two mechanisms to protect against intentional or accidental mis-uses:

1. tcpcp only takes as little information as possible from the user, and re-generates as much of the state related to the TCP connection (such as neighbour and destination data) as possible from local information. Furthermore, it performs a number of sanity checks on the ICI, to ensure its integrity, and compatibility with constraints of the local system (such as buffer size limits and kernel capabilities).

2. Many manipulations possible through tcpcp can be shown to be available through other means if the application has the `CAP_NET_RAW` capability. Therefore, establishing a new TCP connection with tcpcp also requires this capability. This can be relaxed on a host-wide basis.

7.2 Retrieval of sensitive kernel data

Getting `TCP_ICI` may retrieve information from the kernel that one would like to hide from unprivileged applications, e.g. details about the state of the TCP ISN generator. Since the equally unprivileged `TCP_INFO` already gives access to most TCP connection metadata, tcpcp does not create any new vulnerabilities.

7.3 Local denial of service

Setting `TCP_ICI` could be used to introduce inconsistent data in the TCP stack, or the kernel in general. Preventing this relies on the correctness and completeness of the sanity checks mentioned before.

tcpcp can be used to accumulate stale data in the kernel. However, this is not very different from e.g. creating a large number of unused sockets, or letting buffers fill up in TCP connections, and therefore poses no new security threat.

tcpcp can be used to shutdown connections belonging to third party applications, provided that the usual access restrictions grant access to copies of their socket descriptors. This is similar to executing `shutdown` on such sockets, and is therefore believed to pose no new threat.

7.4 Restricted state transitions

tcpcp could be used to advance TCP connection state past boundaries imposed by internal or external control mechanisms. In particular, conspiring applications may create TCP connections without ever exchanging SYN packets, bypassing SYN-filtering firewalls. Since SYN-filtering firewalls can already be avoided by privileged applications, sites depending on SYN-filtering firewalls should therefore use the default setting of tcpcp, which makes its use also a privileged operation.

7.5 Attacks on remote hosts

The ability to set `TCP_ICI` makes it easy to commit all kinds of of protocol violations. While tcpcp may simplify implementing such attacks, this type of abuses has always been possible for privileged users, and therefore, tcpcp poses no new security threat to systems properly resistant against network attacks.

However, if a site allows systems where only trusted users may be able to communicate with otherwise shielded systems with known remote TCP vulnerabilities, tcpcp could be used for attacks. Such sites should use the default setting, which makes setting `TCP_ICI` a privileged operation.

7.6 Security summary

To summarize, the author believes that the design of tcpcp does not open any new exploits if tcpcp is used in its default configuration.

Obviously, some subtleties have probably been overlooked, and there may be bugs inadvertently leading to vulnerabilities. Therefore, tcpcp should receive public scrutiny before being considered fit for regular use.

8 Future work

To allow faster connection passing among hosts that share the same, or a very similar path to the peer, tcpcp should try to avoid going to slow start. To do so, it will have to pass more congestion control information, and integrate it properly at the destination.

Although not strictly part of tcpcp, the redirection apparatus for the network should be further extended, in particular to allow individual connections to be redirected at that point too, and to include some middleware that coordinates the redirecting with the changes at the hosts passing the connection.

It would be very interesting if connection passing could also be used for checkpointing. The analysis in Section 6 suggests that at least limited checkpointing capabilities should be feasible without interfering with regular TCP operation.

The inner workings of TCP are complex and easily disturbed. It is therefore important to subject tcpcp to thorough testing, in particular in transient states, such as during recovery from lost segments. The umlsim simulator [12] allows to generate such conditions in a deterministic way, and will be used for these tests.

9 Conclusion

tcpcp is a proof of concept implementation that successfully demonstrates that an endpoint of a TCP connection can be passed from one host to another without involving the host at the opposite end of the TCP connection. tcpcp also shows that this can be accomplished with a relatively small amount of kernel changes.

tcpcp in its present form is suitable for experimental use as a building block for load balancing and process migration solutions. Future

work will focus on improving the performance of tcpcp, on validating its correctness, and on exploring checkpointing capabilities.

References

[1] RFC768; Postel, Jon. *User Datagram Protocol*, IETF, August 1980.

[2] RFC793; Postel, Jon. *Transmission Control Protocol*, IETF, September 1981.

[3] Stevens, W. Richard. *TCP/IP Illustrated, Volume 1 – The Protocols*, Addison-Wesley, 1994.

[4] RFC2616; Fielding, Roy T.; Gettys, James; Mogul, Jeffrey C.; Frystyk Nielsen, Henrik; Masinter, Larry; Leach, Paul J.; Berners-Lee, Tim. *Hypertext Transfer Protocol – HTTP/1.1*, IETF, June 1999.

[5] Bar, Moshe. *OpenMosix*, Proceedings of the 10th International Linux System Technology Conference (Linux-Kongress 2003), pp. 94–102, October 2003.

[6] Kuntz, Bryan; Rajan, Karthik. *MIGSOCK – Migratable TCP Socket in Linux*, CMU, M.Sc. Thesis, February 2002. `http://www-2.cs.cmu.edu/~softagents/migsock/MIGSOCK.pdf`

[7] Leite, Fábio Olivé. *Load-Balancing HA Clusters with No Single Point of Failure*, Proceedings of the 9th International Linux System Technology Conference (Linux-Kongress 2002), pp. 122–131, September 2002. `http://www.linux-kongress.org/2002/papers/lk2002-leite.html`

[8] *Linux Virtual Server Project*, `http://www.linuxvirtualserver.org/`

[9] RFC1918; Rekhter, Yakov; Moskowitz, Robert G.; Karrenberg, Daniel; de Groot, Geert Jan; Lear, Eliot. *Address Allocation for Private Internets*, IETF, February 1996.

[10] RFC1323; Jacobson, Van; Braden, Bob; Borman, Dave. *TCP Extensions for High Performance*, IETF, May 1992.

[11] RFC2018; Mathis, Matt; Mahdavi, Jamshid; Floyd, Sally; Romanow, Allyn. *TCP Selective Acknowledgement Options*, IETF, October 1996.

[12] Almesberger, Werner. *UML Simulator*, Proceedings of the Ottawa Linux Symposium 2003, July 2003. `http://archive.linuxsymposium.org/ols2003/Proceedings/All-Reprints/Reprint-Almesberger-OLS2003.pdf`

Cooperative Linux

Dan Aloni
da-x@colinux.org

Abstract

In this paper I'll describe Cooperative Linux, a port of the Linux kernel that allows it to run as an unprivileged lightweight virtual machine in kernel mode, on top of another OS kernel. It allows Linux to run under any operating system that supports loading drivers, such as Windows or Linux, after minimal porting efforts. The paper includes the present and future implementation details, its applications, and its comparison with other Linux virtualization methods. Among the technical details I'll present the CPU-complete context switch code, hardware interrupt forwarding, the interface between the host OS and Linux, and the management of the VM's pseudo physical RAM.

1 Introduction

Cooperative Linux utilizes the rather underused concept of a Cooperative Virtual Machine (CVM), in contrast to traditional VMs that are unprivileged and being under the complete control of the host machine.

The term **Cooperative** is used to describe two entities working in parallel, e.g. coroutines [2]. In that sense the most plain description of Cooperative Linux is turning two operating system kernels into two big coroutines. In that mode, each kernel has its own complete CPU context and address space, and each kernel decides when to give control back to its partner.

However, only one of the two kernels has control on the physical hardware, where the other is provided only with virtual hardware abstraction. From this point on in the paper I'll refer to these two kernels as the host operating system, and the guest Linux VM respectively. The host can be every OS kernel that exports basic primitives that provide the Cooperative Linux portable driver to run in CPL0 mode (ring 0) and allocate memory.

The special CPL0 approach in Cooperative Linux makes it significantly different than traditional virtualization solutions such as VMware, plex86, Virtual PC, and other methods such as Xen. All of these approaches work by running the guest OS in a less privileged mode than of the host kernel. This approach allowed for the extensive simplification of Cooperative Linux's design and its short early-beta development cycle which lasted only one month, starting from scratch by modifying the vanilla Linux 2.4.23-pre9 release until reaching to the point where KDE could run.

The only downsides to the CPL0 approach is stability and security. If it's unstable, it has the potential to crash the system. However, measures can be taken, such as cleanly shutting it down on the first internal Oops or panic. Another disadvantage is security. Acquiring root user access on a Cooperative Linux machine can potentially lead to root on the host machine if the attacker loads specially crafted kernel module or uses some very elaborated exploit in case which the Cooperative Linux kernel was compiled without module support.

Most of the changes in the Cooperative Linux patch are on the i386 tree—the only supported architecture for Cooperative at the time of this writing. The other changes are mostly additions of virtual drivers: cobd (block device), conet (network), and cocon (console). Most of the changes in the i386 tree involve the initialization and setup code. It is a goal of the Cooperative Linux kernel design to remain as close as possible to the standalone i386 kernel, so all changes are localized and minimized as much as possible.

2 Uses

Cooperative Linux in its current early state can already provide some of the uses that User Mode Linux[1] provides, such as virtual hosting, kernel development environment, research, and testing of new distributions or buggy software. It also enabled new uses:

- **Relatively effortless migration path from Windows.** In the process of switching to another OS, there is the choice between installing another computer, dual-booting, or using a virtualization software. The first option costs money, the second is tiresome in terms of operation, but the third can be the most quick and easy method—especially if it's free. This is where Cooperative Linux comes in. It is already used in workplaces to convert Windows users to Linux.
- **Adding Windows machines to Linux clusters.** The Cooperative Linux patch is minimal and can be easily combined with others such as the MOSIX or OpenMOSIX patches that add clustering capabilities to the kernel. This work in progress allows to add Windows machines to super-computer clusters, where one illustration could tell about a secretary workstation computer that runs Cooperative Linux as a screen saver—when the secretary goes home at the end of the day and leaves the computer unattended, the office's cluster gets more CPU cycles for free.
- **Running an otherwise-dual-booted Linux system from the other OS.** The Windows port of Cooperative Linux allows it to mount real disk partitions as block devices. Numerous people are using this in order to access, rescue, or just run their Linux system from their ext3 or reiserfs file systems.
- **Using Linux as a Windows firewall on the same machine.** As a likely competitor to other out-of-the-box Windows firewalls, iptables along with a stripped-down Cooperative Linux system can potentially serve as a network firewall.
- **Linux kernel development / debugging / research and study on another operating systems.**

  Digging inside a running Cooperative Linux kernel, you can hardly tell the difference between it and a standalone Linux. All virtual addresses are the same—Oops reports look familiar and the architecture dependent code works in the same manner, excepts some transparent conversions, which are described in the next section in this paper.
- **Development environment for porting to and from Linux.**

3 Design Overview

In this section I'll describe the basic methods behind Cooperative Linux, which include

complete context switches, handling of hardware interrupts by forwarding, physical address translation and the pseudo physical memory RAM.

3.1 Minimum Changes

To illustrate the minimal effect of the Cooperative Linux patch on the source tree, here is a diffstat listing of the patch on Linux 2.4.26 as of May 10, 2004:

```
CREDITS                                |    6
Documentation/devices.txt              |    7
Makefile                               |    8
arch/i386/config.in                    |   30
arch/i386/kernel/Makefile              |    2
arch/i386/kernel/cooperative.c         |  181 +++++
arch/i386/kernel/head.S                |    4
arch/i386/kernel/i387.c                |    8
arch/i386/kernel/i8259.c               |  153 ++++
arch/i386/kernel/ioport.c              |   10
arch/i386/kernel/process.c             |   28
arch/i386/kernel/setup.c               |   61 +
arch/i386/kernel/time.c                |  104 +++
arch/i386/kernel/traps.c               |    9
arch/i386/mm/fault.c                   |    4
arch/i386/mm/init.c                    |   37 +
arch/i386/vmlinux.lds                  |   82 +-
drivers/block/Config.in                |    4
drivers/block/Makefile                 |    1
drivers/block/cobd.c                   |  334 ++++++++++
drivers/block/ll_rw_blk.c              |    2
drivers/char/Makefile                  |    4
drivers/char/colx_keyb.c               | 1221 ++++++++++++*
drivers/char/mem.c                     |    8
drivers/char/vt.c                      |    8
drivers/net/Config.in                  |    4
drivers/net/Makefile                   |    1
drivers/net/conet.c                    |  205 ++++++
drivers/video/Makefile                 |    4
drivers/video/cocon.c                  |  484 ++++++++++++++++
include/asm-i386/cooperative.h         |  175 +++++
include/asm-i386/dma.h                 |    4
include/asm-i386/io.h                  |   27
include/asm-i386/irq.h                 |    6
include/asm-i386/mc146818rtc.h         |    7
include/asm-i386/page.h                |   30
include/asm-i386/pgalloc.h             |    7
include/asm-i386/pgtable-2level.h      |    8
include/asm-i386/pgtable.h             |    7
include/asm-i386/processor.h           |   12
include/asm-i386/system.h              |    8
include/linux/console.h                |    1
include/linux/cooperative.h            |  317 +++++++++
include/linux/major.h                  |    1
init/do_mounts.c                       |    3
init/main.c                            |    9
kernel/Makefile                        |    2
kernel/cooperative.c                   |  254 +++++++
kernel/panic.c                         |    4
kernel/printk.c                        |    6
50 files changed, 3828 insertions(+), 74 deletions(-)
```

3.2 Device Driver

The device driver port of Cooperative Linux is used for accessing kernel mode and using the kernel primitives that are exported by the host OS kernel. Most of the driver is OS-independent code that interfaces with the OS dependent primitives that include page allocations, debug printing, and interfacing with user space.

When a Cooperative Linux VM is created, the driver loads a kernel image from a vmlinux file that was compiled from the patched kernel with `CONFIG_COOPERATIVE`. The vmlinux file doesn't need any cross platform tools in order to be generated, and the same vmlinux file can be used to run a Cooperative Linux VM on several OSes of the same architecture.

The VM is associated with a per-process resource—a file descriptor in Linux, or a device handle in Windows. The purpose of this association makes sense: if the process running the VM ends abnormally in any way, all resources are cleaned up automatically from a callback when the system frees the per-process resource.

3.3 Pseudo Physical RAM

In Cooperative Linux, we had to work around the Linux MM design assumption that the entire physical RAM is bestowed upon the kernel on startup, and instead, only give Cooperative Linux a fixed set of physical pages, and then only do the translations needed for it to work transparently in that set. All the memory which Cooperative Linux considers as physical is in that allocated set, which we call the Pseudo Physical RAM.

The memory is allocated in the host OS using the appropriate kernel function—`alloc_pages()` in Linux and `MmAllocatePagesForMdl()` in Windows—so it is not mapped in any address space on the host for not wasting PTEs. The allocated pages are always resident and not freed until the VM is downed. Page tables

```
--- linux/include/asm-i386/pgtable-2level.h     2004-04-20 08:04:01.000000000 +0300
+++ linux/include/asm-i386/pgtable-2level.h     2004-05-09 16:54:09.000000000 +0300
@@ -58,8 +58,14 @@
 }
 #define ptep_get_and_clear(xp) __pte(xchg(&(xp)->pte_low, 0))
 #define pte_same(a, b)         ((a).pte_low == (b).pte_low)
-#define pte_page(x)            (mem_map+((unsigned long)(((x).pte_low >> PAGE_SHIFT))))
 #define pte_none(x)            (!(x).pte_low)
+
+#ifndef CONFIG_COOPERATIVE
+#define pte_page(x)            (mem_map+((unsigned long)(((x).pte_low >> PAGE_SHIFT))))
 #define __mk_pte(page_nr,pgprot) __pte(((page_nr) << PAGE_SHIFT) | pgprot_val(pgprot))
+#else
+#define pte_page(x)            CO_VA_PAGE((x).pte_low)
+#define __mk_pte(page_nr,pgprot) __pte((CO_PA(page_nr) & PAGE_MASK) | pgprot_val(pgprot))
+#endif

 #endif /* _I386_PGTABLE_2LEVEL_H */
```

Table 1: Example of MM architecture dependent changes

are created for mapping the allocated pages in the VM's kernel virtual address space. The VM's address space resembles the address space of a regular kernel—the normal RAM zone is mapped contiguously at 0xc0000000.

The VM address space also has its own special fixmaps—the page tables themselves are mapped at 0xfef00000 in order to provide an O(1) ability for translating PPRAM (Psuedo-Physical RAM) addresses to physical addresses when creating PTEs for user space and `vmalloc()` space. On the other way around, a special physical-to-PPRAM map is allocated and mapped at 0xff000000, to speed up handling of events such as pages faults which require translation of physical addresses to PPRAM address. This bi-directional memory address mapping allows for a negligible overhead in page faults and user space mapping operations.

Very few changes in the i386 MMU macros were needed to facilitate the PPRAM. An example is shown in Table 1. Around an #ifdef of `CONFIG_COOPERATIVE` the `__mk_pte()` low level MM macro translates a PPRAM struct page to a PTE that maps the real physical page. Respectively, `pte_page()` takes a PTE that was generated by `__mk_pte()` and returns the corresponding struct page for it. Other macros such as `pmd_page()` and `load_cr3()` were also changed.

3.4 Context Switching

The Cooperative Linux VM uses only one host OS process in order to provide a context for itself and its processes. That one process, named colinux-daemon, can be called a Super Process since it frequently calls the kernel driver to perform a context switch from the host kernel to the guest Linux kernel and back. With the frequent (HZ times a second) host kernel entries, it is able able to completely control the CPU and MMU without affecting anything else in the host OS kernel.

On the Intel 386 architecture, a complete context switch requires that the top page directory table pointer register—CR3—is changed. However, it is not possible to easily change both the instruction pointer (EIP) and CR3 in one instruction, so it implies that the code that changes CR3 must be mapped in both contexts for the change to be possible. It's problematic to map that code at the same virtual address in both contexts due to design limitations—the two contexts can divide the kernel and user ad-

dress space differently, such that one virtual address can contain a kernel mapped page in one OS and a user mapped page in another.

In Cooperative Linux the problem was solved by using an intermediate address space during the switch (referred to as the 'passage page,' see Figure 1). The intermediate address space is defined by a specially created page tables in both the guest and host contexts and maps the same code that is used for the switch (passage code) at both of the virtual addresses that are involved. When a switch occurs, first CR3 is changed to point to the intermediate address space. Then, EIP is relocated to the other mapping of the passage code using a jump. Finally, CR3 is changed to point to the top page table directory of the other OS.

The single MMU page that contains the passage page code, also contains the saved state of one OS while the other is executing. Upon the beginning of a switch, interrupts are turned off, and a current state is saved to the passage page by the passage page code. The state includes all the general purpose registers, the segment registers, the interrupt descriptor table register (IDTR), the global descriptor table (GDTR), the local descriptor register (LTR), the task register (TR), and the state of the FPU / MMX / SSE registers. In the middle of the passage page code, it restores the state of the other OS and interrupts are turned back on. This process is akin to a "normal" process to process context switch.

Since control is returned to the host OS on every hardware interrupt (described in the following section), it is the responsibility of the host OS scheduler to give time slices to the Cooperative Linux VM just as if it was a regular process.

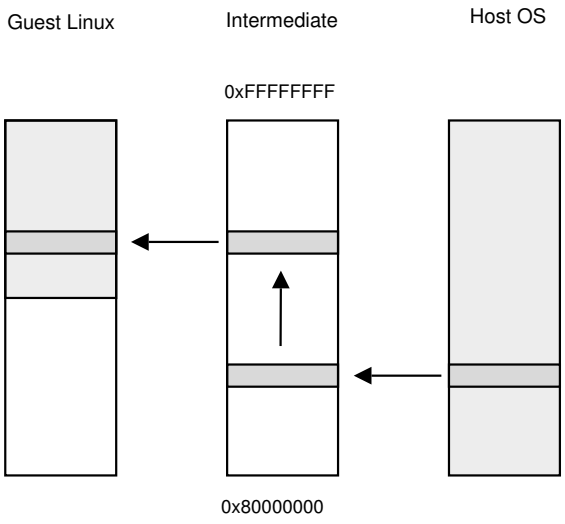


Figure 1: Address space transition during an OS cooperative kernel switch, using an inter-mapped page

3.5 Interrupt Handling and Forwarding

Since a complete MMU context switch also involves the IDTR, Cooperative Linux must set an interrupt vector table in order to handle the hardware interrupts that occur in the system during its running state. However, Cooperative Linux only forwards the invocations of interrupts to the host OS, because the latter needs to know about these interrupts in order to keep functioning and support the colinux-daemon process itself, regardless to the fact that external hardware interrupts are meaningless to the Cooperative Linux virtual machine.

The interrupt vectors for the internal processor exceptions (0x0–0x1f) and the system call vector (0x80) are kept like they are so that Cooperative Linux handles its own page faults and other exceptions, but the other interrupt vectors point to special proxy ISRs (interrupt service routines). When such an ISR is invoked during the Cooperative Linux context by an external hardware interrupt, a context switch is made to the host OS using the passage code. On the

other side, the address of the relevant ISR of the host OS is determined by looking at its IDT. An interrupt call stack is forged and a jump occurs to that address. Between the invocation of the ISR in the Linux side and the handling of the interrupt in the host side, the interrupt flag is disabled.

The operation adds a tiny latency to interrupt handling in the host OS, but it is quite neglectable. Considering that this interrupt forwarding technique also involves the hardware timer interrupt, the host OS cannot detect that its CR3 was hijacked for a moment and therefore no exceptions in the host side would occur as a result of the context switch.

To provide interrupts for the virtual device drivers of the guest Linux, the changes in the arch code include a virtual interrupt controller which receives messages from the host OS on the occasion of a switch and invokes `do_IRQ()` with a forged `struct pt_args`. The interrupt numbers are virtual and allocated on a per-device basis.

4 Benchmarks And Performance

4.1 Dbench results

This section shows a comparison between User Mode Linux and Cooperative Linux. The machine which the following results were generated on is a 2.8GHz Pentium 4 with HT enabled, 512GB RAM, and a 120GB SATA Maxtor hard-drive that hosts ext3 partitions. The comparison was performed using the dbench 1.3-2 package of Debian on all setups.

The host machine runs the Linux 2.6.6 kernel patched with SKAS support. The UML kernel is Linux 2.6.4 that runs with 32MB of RAM, and is configured to use SKAS mode. The Cooperative Linux kernel is a Linux 2.4.26 kernel and it is configured to run with 32MB of RAM, same as the UML system. The root file-system of both UML and Cooperative Linux machines is the same host Linux file that contains an ext3 image of a 0.5GB minimized Debian system.

The commands ‘dbench 1’, ‘dbench 3’, and ‘dbench 10’ were run in 3 consecutive runs for each command, on the host Linux, on UML, and on Cooperative Linux setups. The results are shown in Table 2, Table 3, and Table 4.

System	Throughput	Netbench
Host	43.813	54.766
	50.117	62.647
	44.128	55.160
UML	10.418	13.022
	9.408	11.760
	9.309	11.636
coLinux	10.418	13.023
	12.574	15.718
	12.075	15.094

Table 2: output of dbench 10 (units are in MB/sec)

System	Throughput	Netbench
Host	43.287	54.109
	41.383	51.729
	59.965	74.956
UML	11.857	14.821
	15.143	18.929
	14.602	18.252
coLinux	24.095	30.119
	32.527	40.659
	36.423	45.528

Table 3: output of dbench 3 (units are in MB/sec)

4.2 Understanding the results

From the results in these runs, ‘dbench 10’, ‘dbench 3’, and ‘dbench 1’ show 20%, 123%, and 303% increase respectively, compared to UML. These numbers relate to the number

System	Throughput	Netbench
Host	158.205	197.756
	182.191	227.739
	179.047	223.809
UML	15.351	19.189
	16.691	20.864
	16.180	20.226
coLinux	45.592	56.990
	72.452	90.565
	106.952	133.691

Table 4: output of dbench 1 (units are in MB/sec)

of dbench threads, which is a result of the synchronous implementation of cobd[1]. Yet, neglecting the versions of the kernels compared, Cooperative Linux achieves much better probably because of low overhead with regard to context switching and page faulting in the guest Linux VM.

The current implementation of the cobd driver is synchronous file reading and writing directly from the kernel of the host Linux—No user space of the host Linux is involved, therefore less context switching and copying. About copying, the specific implementation of cobd in the host Linux side benefits from the fact that `filp->f_op->read()` is called directly on the cobd driver's request buffer after mapping it using `kmap()`. Reimplementing this driver as asynchronous on both the host and guest—can improve performance.

Unlike UML, Cooperative Linux can benefit in the terms of performance from the implementation of kernel-to-kernel driver bridges such as cobd. For example, currently virtual Ethernet in Cooperative Linux is done similar to UML—i.e., using user space daemons with tuntap on the host. If instead we create a kernel-to-kernel implementation with no user space daemons in between, Cooperative Linux has the potential to achieve much better in benchmarking.

[1] ubd UML equivalent

5 Planned Features

Since Cooperative Linux is a new project (2004–), most of its features are still waiting to be implemented.

5.1 Suspension

Software-suspending Linux is a challenge on standalone Linux systems, considering the entire state of the hardware needs to be saved and restored, along with the space that needs to be found for storing the suspended image. On User Mode Linux suspending [3] is easier—only the state of a few processes needs saving, and no hardware is involved.

However, in Cooperative Linux, it will be even easier to implement suspension, because it will involve its internal state almost entirely. The procedure will involve serializing the pseudo physical RAM by enumerating all the page table entries that are used in Cooperative Linux, either by itself (for user space and vmalloc page tables) or for itself (the page tables of the pseudo physical RAM), and change them to contain the pseudo value instead of the real value.

The purpose of this suspension procedure is to allow no notion of the real physical memory to be contained in any of the pages allocated for the Cooperative Linux VM, since Cooperative Linux will be given a different set of pages when it will resume at a later time. At the suspended state, the pages can be saved to a file and the VM could be resumed later. Resuming will involve loading that file, allocating the memory, and fix-enumerate all the page tables again so that the values in the page table entries point to the newly allocated memory.

Another implementation strategy will be to just dump everything on suspension as it is, but on resume—enumerate all the page table entries and adjust between the values of the old RPPFNs[2] and new RPPFNs.

Note that a suspended image could be created under one host OS and be resumed in another host OS of the same architecture. One could carry a suspended Linux on a USB memory device and resume/suspend it on almost any computer.

5.2 User Mode Linux[1] inside Cooperative Linux

The possibility of running UML inside Cooperative Linux is not far from being immediately possible. It will allow to bring UML with all its glory to operating systems that cannot support it otherwise because of their user space APIs. Combining UML and Cooperative Linux cancels the security downside that running Cooperative Linux could incur.

5.3 Live Cooperative Distributions

Live-CD distributions like KNOPPIX can be used to boot on top of another operating system and not only as standalone, reaching a larger sector of computer users considering the host operating system to be Windows NT/2000/XP.

5.4 Integration with ReactOS

ReactOS, the free Windows NT clone, will be incorporating Cooperative Linux as a POSIX subsystem.

5.5 Miscellaneous

- Virtual frame buffer support.
- Incorporating features from User Mode Linux, e.g. humfs [3].
- Support for more host operating systems such as FreeBSD.

6 Conclusions

We have discussed how Cooperative Linux works and its benefits—apart from being a BSKH[4], Cooperative Linux has the potential to become an alternative to User Mode Linux that enhances on portability and performance, rather than on security.

Moreover, the implications that Cooperative Linux has on what is the media defines as ‘Linux on the Desktop’—are massive, as the world’s most dominant albeit proprietary desktop OS supports running Linux distributions for free, as another software, with the aimed-for possibility that the Linux newbie would switch to the standalone Linux. As user-friendliness of the Windows port will improve, the exposure that Linux gets by the average computer user can increase tremendously.

7 Thanks

Muli Ben Yehuda, IBM

Jun Okajima, Digital Infra

Kuniyasu Suzaki, AIST

References

[1] Jeff Dike. User Mode Linux. `http://user-mode-linux.sf.net`.

[2] real physical page frame numbers

[3] A recent addition to UML that provides an host FS implementation that uses files in order to store its VFS metadata

[4] Big Scary Kernel Hack

[2] Donald E. Knuth. *The Art of Computer Programming*, volume 1. Addison-Wesley, Reading, Massachusetts, 1997. Describes coroutines in their pure sense.

[3] Richard Potter. Scrapbook for User Mode Linux. `http://sbuml.sourceforge.net/`.

Build your own Wireless Access Point

Erik Andersen
Codepoet Consulting
andersen@codepoet.org

Abstract

This presentation will cover the software, tools, libraries, and configuration files needed to construct an embedded Linux wireless access point. Some of the software available for constructing embedded Linux systems will be discussed, and selection criteria for which tools to use for differing embedded applications will be presented. During the presentation, an embedded Linux wireless access point will be constructed using the Linux kernel, the uClibc C library, BusyBox, the syslinux bootloader, iptables, etc. Emphasis will be placed on the more generic aspects of building an embedded Linux system using BusyBox and uClibc. At the conclusion of the presentation, the presenter will (with luck) boot up the newly constructed wireless access point and demonstrate that it is working perfectly. Source code, build system, cross compilers, and detailed instructions will be made available.

1 Introduction

When I began working on embedded Linux, the question of whether or not Linux was small enough to fit inside a particular device was a difficult problem. Linux distributions[1] have historically been designed for server and desktop systems. As such, they deliver a full-featured, comprehensive set of tools for just about every purpose imaginable. Most Linux distributions, such as Red Hat, Debian, or SuSE, provide hundreds of separate software packages adding up to several gigabytes of software. The goal of server or desktop Linux distributions has been to provide as much value as possible to the user; therefore, the large size is quite understandable. However, this has caused the Linux operating system to be much larger then is desirable for building an embedded Linux system such as a wireless access point. Since embedded devices represent a fundamentally different target for Linux, it became apparent to me that embedded devices would need different software than what is commonly used on desktop systems. I knew that Linux has a number of strengths which make it extremely attractive for the next generation of embedded devices, yet I could see that developers would need new tools to take advantage of Linux within small, embedded spaces.

I began working on embedded Linux in the middle of 1999. At the time, building an 'embedded Linux' system basically involved copying binaries from an existing Linux distribution to a target device. If the needed software did not fit into the required amount of flash memory, there was really nothing to be done about it except to add more flash or give up on the project. Very little effort had been made to develop smaller application programs and li-

[1] The term "distribution" is used by the Linux community to refer to a collection of software, including the Linux kernel, application programs, and needed library code, which makes up a complete running system. Sometimes, the term "Linux" or "GNU/Linux" is also used to refer to this collection of software.

braries designed for use in embedded Linux.

As I began to analyze how I could save space, I decided that there were three main areas that could be attacked to shrink the footprint of an embedded Linux system: the kernel, the set of common application programs included in the system, and the shared libraries. Many people doing Linux kernel development were at least talking about shrinking the footprint of the kernel. For the past five years, I have focused on the latter two areas: shrinking the footprint of the application programs and libraries required to produce a working embedded Linux system. This paper will describe some of the software tools I've worked on and maintained, which are now available for building very small embedded Linux systems.

2 The C Library

Let's take a look at an embedded Linux system, the Linux Router Project, which was available in 1999. `http://www.linuxrouter.org/` The Linux Router Project, begun by Dave Cinege, was and continues to be a very commonly used embedded Linux system. Its self-described tagline reads "A networking-centric micro-distribution of Linux" which is "small enough to fit on a single 1.44MB floppy disk, and makes building and maintaining routers, access servers, thin servers, thin clients, network appliances, and typically embedded systems next to trivial." First, let's download a copy of one of the Linux Router Project's "idiot images." I grabbed my copy from the mirror site at `ftp://sunsite.unc.edu/pub/Linux/distributions/linux-router/dists/current/idiot-image_1440KB_FAT_2.9.8_Linux_2.2.gz`.

Opening up the idiot-image there are several very interesting things to be seen.

```
# gunzip \
  idiot-image_1440KB_FAT_2.9.8_Linux_2.2.gz
# mount \
  idiot-image_1440KB_FAT_2.9.8_Linux_2.2 \
  /mnt -o loop

# du -ch /mnt/*
34K     /mnt/etc.lrp
6.0K    /mnt/ldlinux.sys
512K    /mnt/linux
512     /mnt/local.lrp
1.0K    /mnt/log.lrp
17K     /mnt/modules.lrp
809K    /mnt/root.lrp
512     /mnt/syslinux.cfg
1.0K    /mnt/syslinux.dpy
1.4M    total

# mkdir test
# cd test
# tar -xzf /mnt/root.lrp

# du -hs
2.2M    .
2.2M    total

# du -ch bin root sbin usr var
460K    bin
8.0K    root
264K    sbin
12K     usr/bin
304K    usr/sbin
36K     usr/lib/ipmasqadm
40K     usr/lib
360K    usr
56K     var/lib/lrpkg
60K     var/lib
4.0K    var/spool/cron/crontabs
8.0K    var/spool/cron
12K     var/spool
76K     var
1.2M    total

# du -ch lib
24K     lib/POSIXness
1.1M    lib
1.1M    total

# du -h lib/libc-2.0.7.so
644K    lib/libc-2.0.7.so
```

Taking a look at the software contained in this embedded Linux system, we quickly notice that in a software image totaling 2.2 Megabytes, the libraries take up over half the space. If we look even closer at the set of libraries, we quickly find that the largest single component in the entire system is the GNU C library, in this case occupying nearly 650k. What is more, this is a very old version of the C library; newer versions of GNU glibc,

such as version 2.3.2, are over 1.2 Megabytes all by themselves! There are tools available from Linux vendors and in the Open Source community which can reduce the footprint of the GNU C library considerably by stripping unwanted symbols; however, using such tools precludes adding additional software at a later date. Even when these tools are appropriate, there are limits to the amount of size which can be reclaimed from the GNU C library in this way.

The prospect of shrinking a single library that takes up so much space certainly looked like low hanging fruit. In practice, however, replacing the GNU C library for embedded Linux systems was not easy task.

3 The origins of uClibc

As I despaired over the large size of the GNU C library, I decided that the best thing to do would be to find another C library for Linux that would be better suited for embedded systems. I spent quite a bit of time looking around, and after carefully evaluating the various Open Source C libraries that I knew of[2], I sadly found that none of them were suitable replacements for glibc. Of all the Open Source C libraries, the library closest to what I imagined an embedded C library should be was called `uC-libc` and was being used for uClinux systems. However, it also had many problems at the time—not the least of which was that uC-libc had no central maintainer. The only mechanism being used to support multiple architectures was a complete source tree fork, and there had already been a few such forks with plenty of divergant code. In short, uC-libc was a mess of twisty versions, all different. After spending some time with the code, I decided to fix it, and in the process changed the name to `uClibc` (no hyphen).

With the help of D. Jeff Dionne, one of the creators of uClinux[3], I ported uClibc to run on Intel compatible x86 CPUs. I then grafted in the header files from glibc 2.1.3 to simplify software ports, and I cleaned up the resulting breakage. The header files were later updated again to generally match glibc 2.3.2. This effort has made porting software from glibc to uClibc extremely easy. There were, however, many functions in uClibc that were either broken or missing and which had to be re-written or created from scratch. When appropriate, I sometimes grafted in bits of code from the current GNU C library and libc5. Once the core of the library was reasonably solid, I began adding a platform abstraction layer to allow uClibc to compile and run on different types of CPUs. Once I had both the ARM and x86 platforms basically running, I made a few small announcements to the Linux community. At that point, several people began to make regular contributions. Most notably was Manuel Novoa III, who began contributing at that time. He has continued working on uClibc and is responsible for significant portions of uClibc such as the stdio and internationalization code.

After a great deal of effort, we were able to build the first shared library version of uClibc in January 2001. And earlier this year we were able to compile a Debian Woody system using uClibc[4], demonstrating the library is now able

[2] The Open Source C libraries I evaluated at the time included Al's Free C RunTime library (no longer on the Internet); dietlibc available from `http://www.fefe.de/dietlibc/`; the minix C library available from `http://www.cs.vu.nl/cgi-bin/raw/pub/minix/`; the newlib library available from `http://sources.redhat.com/newlib/`; and the eCos C library available from `ftp://ecos.sourceware.org/pub/ecos/`.

[3] uClinux is a port of Linux designed to run on microcontrollers which lack Memory Management Units (MMUs) such as the Motorolla DragonBall or the ARM7TDMI. The uClinux web site is found at `http://www.uclinux.org/`.

[4] `http://www.uclibc.org/dists/`

to support a complete Linux distribution. People now use uClibc to build versions of Gentoo, Slackware, Linux from Scratch, rescue disks, and even live Linux CDs[5]. A number of commercial products have also been released using uClibc, such as wireless routers, network attached storage devices, DVD players, etc.

4 Compiling uClibc

Before we can compile uClibc, we must first grab a copy of the source code and unpack it so it is ready to use. For this paper, we will just grab a copy of the daily uClibc snapshot.

```
# SITE=http://www.uclibc.org/downloads
# wget -q $SITE/uClibc-snapshot.tar.bz2

# tar -xjf uClibc-snapshot.tar.bz2
# cd uClibc
```

uClibc requires a configuration file, `.config`, that can be edited to change the way the library is compiled, such as to enable or disable features (i.e. whether debugging support is enabled or not), to select a cross-compiler, etc. The preferred method when starting from scratch is to run `make defconfig` followed by `make menuconfig`. Since we are going to be targeting a standard Intel compatible x86 system, no changes to the default configuration file are necessary.

5 The Origins of BusyBox

As I mentioned earlier, the two components of an embedded Linux that I chose to work towards reducing in size were the shared libraries and the set common application programs. A typical Linux system contains a variety of command-line utilities from numerous different organizations and independent programmers. Among the most prominent of these utilities were GNU shellutils, fileutils, textutils (now combined to form GNU coreutils), and similar programs that can be run within a shell (commands such as `sed`, `grep`, `ls`, etc.). The GNU utilities are generally very high-quality programs, and are almost without exception very, very feature-rich. The large feature set comes at the cost of being quite large—prohibitively large for an embedded Linux system. After some investigation, I determined that it would be more efficient to replace them rather than try to strip them down, so I began looking at alternatives.

Just as with alternative C libraries, there were several choices for small shell utilities: BSD has a number of utilities which could be used. The Minix operating system, which had recently released under a free software license, also had many useful utilities. Sash, the stand alone shell, was also a possibility. After quite a lot of research, the one that seemed to be the best fit was BusyBox. It also appealed to me because I was already familiar with BusyBox from its use on the Debian boot floppies, and because I was acquainted with Bruce Perens, who was the maintainer. Starting approximately in October 1999, I began enhancing BusyBox and fixing the most obvious problems. Since Bruce was otherwise occupied and was no longer actively maintaining BusyBox, Bruce eventually consented to let me take over maintainership.

Since that time, BusyBox has gained a large following and attracted development talent from literally the whole world. It has been used in commercial products such as the IBM Linux wristwatch, the Sharp Zaurus PDA, and Linksys wireless routers such as the WRT54G, with many more products being released all the time. So many new features and applets have been added to BusyBox, that the biggest chal-

[5] Puppy Linux available from `http://www.goosee.com/puppy/` is a live linux CD system built with uClibc that includes such favorites as XFree86 and Mozilla.

lenge I now face is simply keeping up with all of the patches that get submitted!

6 So, How Does It Work?

BusyBox is a multi-call binary that combines many common Unix utilities into a single executable. When it is run, BusyBox checks if it was invoked via a symbolic link (a `symlink`), and if the name of the symlink matches the name of an applet that was compiled into BusyBox, it runs that applet. If BusyBox is invoked as `busybox`, then it will read the command line and try to execute the applet name passed as the first argument. For example:

```
# ./busybox date
Wed Jun 2 15:01:03 MDT 2004

# ./busybox echo "hello there"
hello there

# ln -s ./busybox uname
# ./uname
Linux
```

BusyBox is designed such that the developer compiling it for an embedded system can select exactly which applets to include in the final binary. Thus, it is possible to strip out support for unneeded and unwanted functionality, resulting in a smaller binary with a carefully selected set of commands. The customization granularity for BusyBox even goes one step further: each applet may contain multiple features that can be turned on or off. Thus, for example, if you do not wish to include large file support, or you do not need to mount NFS filesystems, you can simply turn these features off, further reducing the size of the final BusyBox binary.

7 Compiling Busybox

Let's walk through a normal compile of BusyBox. First, we must grab a copy of the BusyBox source code and unpack it so it is ready to use. For this paper, we will just grab a copy of the daily BusyBox snapshot.

```
# SITE=http://www.busybox.net/downloads
# wget -q $SITE/busybox-snapshot.tar.bz2
# tar -xjf busybox-snapshot.tar.bz2
# cd busybox
```

Now that we are in the BusyBox source directory we can configure BusyBox so that it meets the needs of our embedded Linux system. This is done by editing the file `.config` to change the set of applets that are compiled into BusyBox, to enable or disable features (i.e. whether debugging support is enabled or not), and to select a cross-compiler. The preferred method when starting from scratch is to run `make defconfig` followed by `make menuconfig`. Once BusyBox has been configured to taste, you just need to run `make` to compile it.

8 Installing Busybox to a Target

If you then want to install BusyBox onto a target device, this is most easily done by typing: `make install`. The installation script automatically creates all the required directories (such as `/bin`, `/sbin`, and the like) and creates appropriate symlinks in those directories for each applet that was compiled into the BusyBox binary.

If we wanted to install BusyBox to the directory /mnt, we would simply run:

```
# make PREFIX=/mnt install

[--installation text omitted--]
```

9 Let's build something that works!

Now that I have certainly bored you to death, we finally get to the fun part, building our own embedded Linux system. For hardware, I will be using a Soekris 4521 system[6] with an 133 Mhz AMD Elan CPU, 64 MB main memory, and a generic Intersil Prism based 802.11b card that can be driven using the `hostap`[7] driver. The root filesystem will be installed on a compact flash card.

To begin with, we need to create toolchain with which to compile the software for our wireless access point. This requires we first compile GNU binutils[8], then compile the GNU compiler collection—gcc[9], and then compile uClibc using the newly created gcc compiler. With all those steps completed, we must finally recompile gcc using using the newly built uClibc library so that `libgcc_s` and `libstdc++` can be linked with uClibc.

Fortunately, the process of creating a uClibc toolchain can be automated. First we will go to the uClibc website and obtain a copy of the uClibc `buildroot` by going here:

`http://www.uclibc.org/cgi-bin/cvsweb/buildroot/`

and clicking on the "Download tarball" link[10]. This is a simple GNU make based build system which first builds a uClibc toolchain, and then builds a root filesystem using the newly built uClibc toolchain.

For the root filesystem of our wireless access point, we will need a Linux kernel, uClibc, BusyBox, pcmcia-cs, iptables, hostap, wtools, bridgeutils, and the dropbear ssh server. To compile these programs, we will first edit the buildroot Makefile to enable each of these items. Figure 1 shows the changes I made to the buildroot Makefile:

Running `make` at this point will download the needed software packages, build a toolchain, and create a minimal root filesystem with the specified software installed.

On my system, with all the software packages previously downloaded and cached locally, a complete build took 17 minutes, 19 seconds. Depending on the speed of your network connection and the speed of your build system, now might be an excellent time to take a lunch break, take a walk, or watch a movie.

10 Checking out the new Root Filesystem

We now have our root filesystem finished and ready to go. But we still need to do a little more work before we can boot up our newly built embedded Linux system. First, we need to compress our root filesystem so it can be loaded as an initrd.

```
# gzip -9 root_fs_i386
# ls -sh root_fs_i386.gz
1.1M root_fs_i386.gz
```

Now that our root filesystem has been compressed, it is ready to install on the boot media. To make things simple, I will install the Compact Flash boot media into a USB card reader device, and copy files using the card reader.

```
# ms-sys -s /dev/sda
Public domain master boot record
successfully written to /dev/sda
```

[6] `http://www.soekris.com/net4521.htm`
[7] `http://hostap.epitest.fi/`
[8] `http://sources.redhat.com/binutils/`
[9] `http://gcc.gnu.org/`
[10] `http://www.uclibc.org/cgi-bin/cvsweb/buildroot.tar.gz?view=tar`

```
--- Makefile
+++ Makefile
@@ -140,6 +140,6 @@
 # Unless you want to build a kernel, I recommend just using
 # that...
-TARGETS+=kernel-headers
-#TARGETS+=linux
+#TARGETS+=kernel-headers
+TARGETS+=linux
 #TARGETS+=system-linux

@@ -150,5 +150,5 @@
 #TARGETS+=zlib openssl openssh
 # Dropbear sshd is much smaller than openssl + openssh
-#TARGETS+=dropbear_sshd
+TARGETS+=dropbear_sshd

 # Everything needed to build a full uClibc development system!
@@ -175,5 +175,5 @@

 # Some stuff for access points and firewalls
-#TARGETS+=iptables hostap wtools dhcp_relay bridge
+TARGETS+=iptables hostap wtools dhcp_relay bridge
 #TARGETS+=iproute2 netsnmp
```

Figure 1: Changes to the buildroot Makefile

```
# mkdosfs /dev/sda1
mkdosfs 2.10 (22 Sep 2003)

# syslinux /dev/sda1

# cp root_fs_i386.gz /mnt/root_fs.gz

# cp build_i386/buildroot-kernel /mnt/linux
```

So we now have a copy of our root filesystem and Linux kernel on the compact flash disk. Finally, we need to configure the bootloader. In case you missed it a few steps ago, we are using the syslinux bootloader for this example. I happen to have a ready to use syslinux configuration file, so I will now install that to the compact flash disk as well:

```
# cat  syslinux.cfg
TIMEOUT 0
PROMPT 0
DEFAULT linux
LABEL linux
 KERNEL linux
 APPEND initrd=root_fs.gz \
        console=ttyS0,57600 \
        root=/dev/ram0 boot=/dev/hda1,msdos rw

# cp syslinux.cfg /mnt
```

And now, finally, we are done. Our embedded Linux system is complete and ready to boot. And you know what? It is very, very small. Take a look at Table 1.

With a carefully optimized Linux kernel (which this kernel unfortunately isn't) we could expect to have even more free space. And remember, every bit of space we save is money that embedded Linux developers don't have to spend on expensive flash memory. So now comes the final test; it is now time to boot from our compact flash disk. Here is what you should see.

```
[----kernel boot messages snipped--]
```

```
# ll /mnt
total 1.9M
drwxr-r-     2  root  root   16K  Jun  2  16:39  ./
drwxr-xr-x  22  root  root  4.0K  Feb  6  07:40  ../
-r-xr-r-     1  root  root  7.7K  Jun  2  16:36  ldlinux.sys*
-rwxr-r-     1  root  root  795K  Jun  2  16:36  linux*
-rwxr-r-     1  root  root  1.1M  Jun  2  16:36  root_fs.gz*
-rwxr-r-     1  root  root   170  Jun  2  16:39  syslinux.cfg*
```

Table 1: Output of `ls -lh /mnt`.

```
Freeing unused kernel memory: 64k freed

Welcome to the Erik's wireless access point.

uclibc login: root

BusyBox v1.00-pre10 (2004.06.02-21:54+0000)
Built-in shell (ash)
Enter 'help' for a list of built-in commands.

# du -h / | tail -n 1
2.6M

#
```

And there you have it—your very own wireless access point. Some additional configuration will be necessary to start up the wireless interface, which will be demonstrated during my presentation.

11 Conclusion

The two largest components of a standard Linux system are the utilities and the libraries. By replacing these with smaller equivalents a much more compact system can be built. Using BusyBox and uClibc allows you to customize your embedded distribution by stripping out unneeded applets and features, thus further reducing the final image size. This space savings translates directly into decreased cost per unit as less flash memory will be required. Combine this with the cost savings of using Linux, rather than a more expensive proprietary OS, and the reasons for using Linux become very compelling. The example Wireless Access point we created is a simple but useful example. There are thousands of other potential applications that are only waiting for you to create them.

Run-time testing of LSB Applications

Stuart Anderson
Free Standards Group
anderson@freestandards.org

Matt Elder
University of South Caroilina
happymutant@sc.rr.com

Abstract

The dynamic application test tool is capable of checking API usage at run-time. The LSB defines only a subset of all possible parameter values to be valid. This tool is capable of checking these value while the application is running.

This paper will explain how this tool works, and highlight some of the more interesting implementation details such as how we managed to generate most of the code automatically, based on the interface descriptions contained in the LSB database.

Results to date will be presented, along with future plans and possible uses for this tool.

1 Introduction

The Linux Standard Base (LSB) Project began in 1998, when the Linux community came together and decided to take action to prevent GNU/Linux based operating systems from fragmenting in the same way UNIX operating systems did in the 1980s and 1990s. The LSB defines the Application Binary Interface (ABI) for the core part of a GNU/Linux system. As an ABI, the LSB defines the interface between the operating system and the applications. A complete set of tests for an ABI must be capable of measuring the interface from both sides.

Almost from the beginning, testing has been a cornerstone of the project. The LSB was originally organized around 3 components: the written specification, a sample implementation, and the test suites. The written specification is the ultimate definition of the LSB. Both the sample implementation, and the test suites yield to the authority of the written specification.

The sample implementation (SI) is a minimal subset of a GNU/Linux system that provides a runtime that implements the LSB, and as little else as possible. The SI is neither intended to be a minimal distribution, nor the basis for a distribution. Instead, it is used as both a proof of concept and a testing tool. Applications which are seeking certification are required to prove they execute correctly using the SI and two other distributions. The SI is also used to validate the runtime test suites.

The third component is testing. One of the things that strengthens the LSB is its ability to measure, and thus prove, conformance to the standard. Testing is achieved with an array of different test suites, each of which measures a different aspect of the specification.

LSB Runtime

- `cmdchk`

  This test suite is a simple existence test that ensures the required LSB commands and utilities are found on an LSB conforming system.

- `libchk`

  This test suite checks the libraries required by the LSB to ensure they contain the interfaces and symbol versions as specified by the LSB.

- `runtimetests`

  This test suite measures the behavior of the interfaces provided by the GNU/Linux system. This is the largest of the test suites, and is actually broken down into several components, which are referred to collectively as the runtime tests. These tests are derived from the test suites used by the Open Group for UNIX branding.

LSB Packaging

- `pkgchk`

  This test examines an RPM format package to ensure it conforms to the LSB.

- `pkginstchk`

  This test suite is used to ensure that the package management tool provided by a GNU/Linux system will correctly install LSB conforming packages. This suite is still in early stages of development.

LSB Application

- `appchk`

  This test performs a static analysis of an application to ensure that it only uses libraries and interfaces specified by the LSB.

- `dynchk`

  This test is used to measure an applications use of the LSB interfaces during its execution, and is the subject of this paper.

2 The database

The LSB Specification contains over 6600 interfaces, each of which is associated with a library and a header file, and may have parameters. Because of the size and complexity of the data describing these interfaces, a database is used to maintain this information.

It is impractical to try and keep the specification, test suites and development libraries and headers synchronized for this much data. Instead, portions of the specification and tests, and all of the development headers and libraries are generated from the database. This ensures that as changes are made to the database, the changes are propagated to the other parts of the project as well.

Some of the relevant data components in this DB are Libraries, Headers, Interfaces, and Types. There are also secondary components and relations between all of the components. A short description of some of these is needed before moving on to how the dynchk test is constructed.

2.1 Library

The LSB specifies 17 shared libraries, which contains the 6600 interfaces. The interfaces in each library are grouped into logical units called a LibGroup. The LibGroups help to organize the interfaces, which is very useful in the written specification, but isn't used much elsewhere.

2.2 Interface

An Interface represents a globally visible symbol, such as a function, or piece of data. Interfaces have a Type, which is either the type of the global data or the return type of the function. If the Interface is a function, then it will have zero or more Parameters, which form a

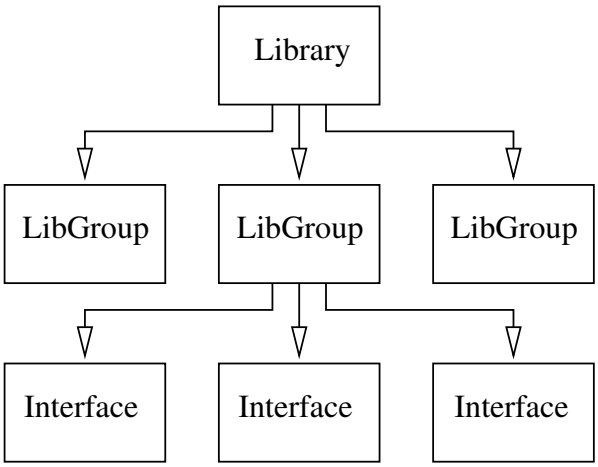


Figure 1: Relationship between Library, LibGroup and Interface

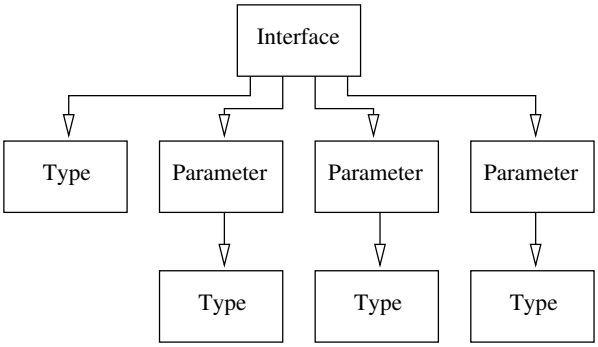


Figure 2: Relationship between Interface, Type and Parameter

set of Types ordered by their position in the parameter list.

2.3 Type

As mentioned above, the database contains enough information to be able to generate header files which are a part of the LSB development tools. This means that the database must be able to represent Clanguage types. The Type and TypeMember tables provide these. These tables are used recursively. If a Type is defined in terms of another type, then it will have a base type that points to that other type.

For structs and unions, the TypeMemeber table is used to hold the ordered list of members. Entries in the TypeMember table point back to the Type table to describe the type of each member. For enums, the TypeMember table is also used to hold the ordered list of values.

Tid	Ttype	Tname	Tbasetype
1	Intrinsic	int	0
2	Pointer	int *	1

Table 1: Example of recursion in Type table for int *

```
struct foo {
  int    a;
  int   *b;
}
```

Figure 3: Sample struct

Tid	Ttype	Tname	Tbasetype
1	Intrinsic	int	0
2	Pointer	int *	1
3	Struct	foo	0

Table 2: Contents of Type table

The structure shown in Figure 3 is represented by the entries in the Type table in Table 2 and the TypeMember table in Table 3.

2.4 Header

Headers, like Libraries, have their contents arranged into logical groupings known a HeaderGroups. Unlike Libraries, these HeaderGroups are ordered so that the proper sequence of definitions within a header file can be maintained. HeaderGroups contain Constant definitions (i.e. #define statements) and Type definitions. If you examine a few well designed header files, you will notice a pattern of a comment followed by related constant definitions and type definitions. The entire header file can be viewed as a repeating sequence of this pat-

Tmid	TMname	TMtypeid	TMposition	TMmemberof
10	a	1	0	3
11	b	2	1	3

Table 3: Contents of TypeMember

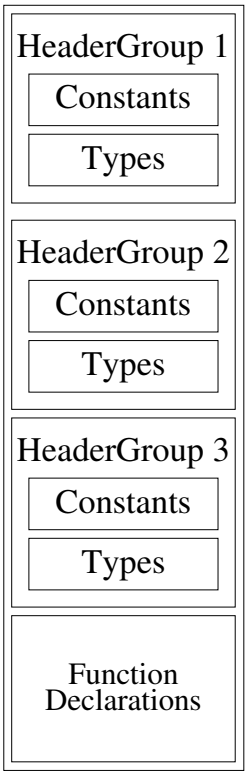


Figure 4: Organization of Headers

tern. This pattern is the basis for the HeaderGroup concept.

2.5 TypeType

One last construct in our database should be mentioned. While we are able to represent a syntactic description of interfaces and types in the database, this is not enough to automatically generate meaningful test cases. We need to add some semantic information that better describes how the types in structures and parameters are used. As an example, `struct sockaddr` contains a member, `sa_family`, of type unsigned short. The compiler will of course ensure that only values between 0 and $2^{16} - 1$ will be used, but only a few of those values have any meaning in this context. By adding the semantic information that this member holds a socket family value, the test generator can cause the value found in `sa_family` to be tested against the legal socket families values (`AF_INET`, `AF_INET6`, etc), instead of just ensuring the value falls between 0 and $2^{16} - 1$, which is really just a noop test.

Example TypeType entries

- `RWaddress`

  An address from the process space that must be both readable and writable.

- `Rdaddress`

  An address from the process space that must be at least readable.

- `filedescriptor`

  A small integer value greater than or equal to 0, and less than the maximum file descriptor for the process.

- `pathname`

  The name of a file or directory that should be compared against the Filesystem Hierarchy Standard.

2.6 Using this data

As mentioned above, the data in the database is used to generate different portions of the LSB project. This strategy was adopted to ensure

these different parts would always be in sync, without having to depend on human intervention.

The written specification contains tables of interfaces, and data definitions (constants and types). These are all generated from the database.

The LSB development environment[1] consists of stub libraries and header files that contain only the interfaces defined by the LSB. This development environment helps catch the use of non-LSB interfaces during the development or porting of an application instead of being surprised by later test results. Both the stub libraries and headers are produced by scripts pulling data from the database.

Some of the test suites described previously have components which are generated from the database. `Cmdchk` and `libchk` have lists of commands and interfaces respectively which are extracted from the database. The static application test tool, `appchk`, also has a list of interfaces that comes from the database. The dynamic application test tool, `dynchk`, has the majority of its code generated from information in the database.

3 The Dynamic Checker

The static application checker simply examines an executable file to determine if it is using interfaces beyond those allowed by the LSB. This is very useful to determine if an application has been built correctly. However, is unable to determine if the interfaces are used correctly when the application is executed. A different kind of test is required to be able to perform this level of checking. This new test must interact with the application while it is running, without interfering with the execution of the application.

This new test has two major components: a mechanism for hooking itself into an application, and a collection of functions to perform the tests for all of the interfaces. These components can mostly be developed independently of each other.

3.1 The Mechanism

The mechanism for interacting with the application must be transparent and noninterfering to the application. We considered the approach used by 3 different tools: abc, ltrace, and fakeroot.

- `abc`—This tool was the inspiration for our new dynamic checker. `abc` was developed as part of the SVR4 ABI test tools. `abc` works by modifying the target application. The application's executable is modified to load a different version of the shared libraries and to call a different version of each interface. This is accomplished by changing the strings in the symbol table and `DT_NEEDED` records. For example, `libc.so.1` is changed to `LiBc.So.1`, and `fread()` is changed to `FrEaD()`. The test set is then located in `/usr/lib/LiBc.So.1`, which in turns loads the original `/usr/lib/libc.so.1`. This mechanism works, but the requirement to modify the executable file is undesirable.
- `ltrace`—This tool is similar to `strace`, except that it traces calls into shared libraries instead of calls into the kernel. `ltrace` uses the ptrace interface to control the application's process. With this approach, the test sets are located in a separate program and are invoked by stopping the application upon

[1] See the May Issue of *Linux Journal* for more information on the LSB Development Environment.

entry to the interface being tested. This approach has two drawbacks: first, the code required to decode the process stack and extract the parameters is unique to each architecture, and second, the tests themselves are more complicated to write since the parameters have to be fetched from the application's process.

- `fakeroot`—This tool is used to create an environment where an unprivileged process appears to have root privileges. `fakeroot` uses `LD_PRELOAD` to load an additional shared library before any of the shared libraries specified by the `DT_NEEDED` records in the executable. This extra library contains a replacement function for each file manipulation function. The functions in this library will be selected by the dynamic linker instead of the normal functions found in the regular libraries. The test sets themselves will perform tests of the parameters, and then call the original version of the functions.

We chose to use the `LD_PRELOAD` mechanism because we felt it was the simplest to use. Based on this mechanism, a sample test case looks like Figure 5.

One problem that must be avoided when using this mechanism is recursion. If the above function just called `read()` at the end, it would end up calling itself again. Instead, the `RTLD_NEXT` flag passed to `dlsym()` tells the dynamic linker to look up the symbol on one of the libraries loaded after the current library. This will get the original version of the function.

3.2 Test set organization

The test set functions are organized into 3 layers. The top layer contains the functions that are test stubs for the LSB interfaces. These functions are implemented by calling the functions in layers 2 and 3. An example of a function in the first layer was given in Figure 5.

The second layer contains the functions that test data structures and types which are passed in as parameters. These functions are also implemented by calling the functions in layer 3 and other functions in layer 2. A function in the second layer looks like Figure 6.

The third layer contains functions that test the types which have been annotated with additional semantic information. These functions often have to perform nontrivial operations to test the assertion required for these supplemental types. Figure 7 is an example of a layer 3 function.

Presently, there are 3056 functions in layer 1 (tests for `libstdc++` are not yet being generated), 106 functions in layer 2, and just a few in layer 3. We estimate that the total number of functions in layer 3 upon completion of the test tool will be on the order of several dozen. The functions in the first two layers are automatically generated based on the information in the database. Functions in layer 3 are hand coded.

3.3 Automatic generation of the tests

In Table 4, is a summary of the size of the test tool so far. As work progresses, these numbers will only get larger. Most of the code in the test is very repetitive, and prone to errors when edited manually. The ability to automate the process of creating this code is highly desirable.

Let's take another look at the sample function from layer 1. This time, however, lets replace some of the code with a description of the information it represents. See Figure 8 for this parameterized version.

All of the occurrences of the string *read* are

```
  ssize_t read (int arg0, void *arg1, size_t arg2) {
    if (!funcptr)
        funcptr = dlsym(RTLD_NEXT, "read");
    validate_filedescriptor(arg0, "read");
    validate_RWaddress(arg1, "read");
    validate_size_t(arg2, "read");
    return funcptr(arg0, arg1, arg2);
  }
```

Figure 5: Test case for read() function

```
void validate_struct_sockaddr_in(struct sockaddr_in *input,
                                 char *name) {
  validate_socketfamily(input->sin_family,name);
  validate_socketport(input->sin_port,name);
  validate_IPv4Address((input->sin_addr), name);
}
```

Figure 6: Test case for validating `struct sockaddr_in`

Module	Files	Lines of Code
libc	752	19305
libdl	5	125
libgcc_s	13	262
libGL	450	11046
libICE	49	1135
libm	281	6568
libncurses	266	6609
libpam	13	335
libpthread	82	2060
libSM	37	865
libX11	668	16112
libXext	113	2673
libXt	288	7213
libz	39	973
structs	106	1581

Table 4: Summary of generated code

actually just the function name, and could have been replaced also.

The same thing can be done for the sample function from layer 2 as is seen in Figure 9.

These two examples, now represent templates that can be used to create the functions for layers 1 and 2. From the previous description of the database, you can see that there is enough information available to be able to instantiate these templates for each interfaces, and structure used by the LSB.

The automation is implemented by 2 perl scripts: `gen_lib.pl` and `gen_tests.pl`. These scripts generate the code for layers 1 and 2 respectively.

Overall, these scripts work well, but we have run into a few interesting situations along the way.

3.4 Handling the exceptions

So far, we have come up with an overall architecture for the test tool, selected a mechanism that allows us to hook the tests into the running application, discovered the pattern in the test functions so that we could create a template for

```
void validate_filedescriptor(const int fd, const char *name) {
   if (fd >= lsb_sysconf(_SC_OPEN_MAX))
       ERROR("fd too big");
   else if (fd < 0)
       ERROR("fd negative");
   }
```

Figure 7: Test case for validating a filedescriptor

```
 return-type read (list of parameters) {
     if (!funcptr)
         funcptr = dlsym(RTLD_NEXT, "read");
     validate_parameter1 type(arg0, "read");
     validate_parameter2 type(arg1, "read");
     validate_parameter3 type(arg2, "read");
     return funcptr(arg0, arg1, arg2);
 }
```

Figure 8: Parameterized test case for a function

automatically generating the code, and implemented the scripts to generate all of the tests cases. The only problem is that now we run into the real world, where things don't always follow the rules.

Here are a few of the interesting situations we have encountered

- `Variadic Functions`

  Of the 725 functions in libc, 25 of them take a variable number of parameters. This causes problems in the generation of the code for the test case, but most importantly it affects our ability to know how to process the arguments. These function have to be written by hand to handle the special needs of these functions. For the functions in the `exec`, `printf` and `scanf` families, the test cases can be implemented by calling the varargs form of the function (`execl()` can be implemented using `execv()`).

- `open()`

  In addition to the problems of being a variadic function, the third parameter to `open()` and `open64()` is only valid if the `O_CREAT` flag is set in the second parameter to these functions. This simple exception requires a small amount of manual intervention, so these function have to be maintained by hand.

- `memory allocation`

  One of the recursion problems we ran into is that memory will be allocated within the `dlsym()` function call, so the implementation of one test case ends up invoking the test case for one of the memory allocation routines, which by default would call `dlsym()`, creating the recursion. This cycle had to be broken by having the test cases for these routines call libc private interfaces to memory allocation.

- `changing memory map`

```
void validate_struct_structure name(struct structure name
   *input, char *name) {
     validate_type of member 1(input->name of member 1, name);
     validate_type of member 2(input->name of member 2, name);
     validate_type of member 3((input->name of member 3), name);
}
```

Figure 9: Parameterized test case for a struct

Pointers are validated by making sure they contain an address that is valid for the process. `/proc/self/maps` is read to obtain the memory map of the current process. These results are cached, for performance reasons, but usually, the memory map of the process will change over time. Both the stack and the heap will grow, resulting in valid pointers being checked against a cached copy of the memory map. In the event a pointer is found to be invalid, the memory map is re-read, and the pointer checked again. The `mmap()` and `munmap()` test cases are also maintained by hand so that they can also cause the memory map to be re-read.

- `hidden ioctl()s`

  By design, the LSB specifies interfaces at the highest possible level. One example of this, is the use of the termio functions, instead of specifying the underlying `ioctl()` interface. It turns out that this tool catches the underlying `ioctl()` calls anyway, and flags it as an error. The solution is for the termio functions the set a flag indicating that the `ioctl()` test case should skip its tests.

- `Optionally NULL parameters`

  Many interfaces have parameters which may be NULL. This triggerred lots of warnings for many programs. The solution was to add a flag that indicated that the Parameter may be NULL, and to not try to validate the pointer, or the data being pointed to.

No doubt, there will be more interesting situations to have to deal with before this tool is completed.

4 Results

As of the deadline for this paper, results are preliminary, but encouraging. The tool is initially being tested against simple commands such as ls and vi, and some X Windows clients such as xclock and xterm. The tool is correctly inserting itself into the application under test, and we are getting some interesting results that will be examined more closely.

One example is vi passes a NULL to `__strtol_internal` several times during startup.

The tool was designed to work across all architectures. At present, it has been built and tested on only the IA32 and IA64 architectures. No significant problems are anticipate on other architectures.

Additional results and experience will be presented at the conference.

5 Future Work

There is still much work to be done. Some of the outstanding tasks are highlighted here.

- Additional `TypeTypes`

  Semantic information needs to be added for additional parameters and structures. The additional layer 3 tests that correspond to this information must also be implemented.

- Architecture-specific interfaces

  As we found in the LSB, there are some interfaces, and types that are unique to one or more architectures. These need to be handled properly so they are not part of the tests when built on an architecture for which they don't apply.

- Unions

  Although Unions are represented in the database in the same way as structures, the database does not contain enough information to describe how to interpret or test the contents of a union. Test cases that involve unions may have to be written by hand.

- Additional libraries

  The information in the database for the graphics libraries and for `libstdc++` is incomplete, therefore, it is not possible to generate all of the test cases for those libraries. Once the data is complete, the test cases will also be complete.

Linux Block IO—present and future

Jens Axboe
SuSE
axboe@suse.de

Abstract

One of the primary focus points of 2.5 was fixing up the bit rotting block layer, and as a result 2.6 now sports a brand new implementation of basically anything that has to do with passing IO around in the kernel, from producer to disk driver. The talk will feature an in-depth look at the IO core system in 2.6 comparing to 2.4, looking at performance, flexibility, and added functionality. The rewrite of the IO scheduler API and the new IO schedulers will get a fair treatment as well.

No 2.6 talk would be complete without 2.7 speculations, so I shall try to predict what changes the future holds for the world of Linux block I/O.

1 2.4 Problems

One of the most widely criticized pieces of code in the 2.4 kernels is, without a doubt, the block layer. It's bit rotted heavily and lacks various features or facilities that modern hardware craves. This has led to many evils, ranging from code duplication in drivers to massive patching of block layer internals in vendor kernels. As a result, vendor trees can easily be considered forks of the 2.4 kernel with respect to the block layer code, with all of the problems that this fact brings with it: 2.4 block layer code base may as well be considered dead, no one develops against it. Hardware vendor drivers include many nasty hacks and `#ifdef's` to work in all of the various 2.4 kernels that are out there, which doesn't exactly enhance code coverage or peer review.

The block layer fork didn't just happen for the fun of it of course, it was a direct result of the various problem observed. Some of these are added features, others are deeper rewrites attempting to solve scalability problems with the block layer core or IO scheduler. In the next sections I will attempt to highlight specific problems in these areas.

1.1 IO Scheduler

The main 2.4 IO scheduler is called `elevator_linus`, named after the benevolent kernel dictator to credit him for some of the ideas used. `elevator_linus` is a one-way scan elevator that always scans in the direction of increasing LBA. It manages latency problems by assigning sequence numbers to new requests, denoting how many new requests (either merges or inserts) may pass this one. The latency value is dependent on data direction, smaller for reads than for writes. Internally, `elevator_linus` uses a double linked list structure (the kernels `struct list_head`) to manage the request structures. When queuing a new IO unit with the IO scheduler, the list is walked to find a suitable insertion (or merge) point yielding an **O(N)** runtime. That in itself is suboptimal in presence of large amounts of IO and to make matters even worse, we repeat this scan if the request free list was empty when we entered

the IO scheduler. The latter is not an error condition, it will happen all the time for even moderate amounts of write back against a queue.

1.2 struct buffer_head

The main IO unit in the 2.4 kernel is the `struct buffer_head`. It's a fairly unwieldy structure, used at various kernel layers for different things: caching entity, file system block, and IO unit. As a result, it's suboptimal for either of them.

From the block layer point of view, the two biggest problems is the size of the structure and the limitation in how big a data region it can describe. Being limited by the file system *one block* semantics, it can at most describe a `PAGE_CACHE_SIZE` amount of data. In Linux on x86 hardware that means 4KiB of data. Often it can be even worse: raw io typically uses the soft sector size of a queue (default 1KiB) for submitting io, which means that queuing eg 32KiB of IO will enter the io scheduler 32 times. To work around this limitation and get at least to a page at the time, a 2.4 hack was introduced. This is called `vary_io`. A driver advertising this capability acknowledges that it can manage `buffer_head's` of varying sizes at the same time. File system read-ahead, another frequent user of submitting larger sized io, has no option but to submit the read-ahead window in units of the page size.

1.3 Scalability

With the limit on `buffer_head` IO size and `elevator_linus` runtime, it doesn't take a lot of thinking to discover obvious scalability problems in the Linux 2.4 IO path. To add insult to injury, the entire IO path is guarded by a single, global lock: `io_request_lock`. This lock is held during the entire IO queuing operation, and typically also from the other end when a driver subtracts requests for IO submission. A single global lock is a big enough problem on its own (bigger SMP systems will suffer immensely because of cache line bouncing), but add to that long runtimes and you have a really huge IO scalability problem.

Linux vendors have long shipped lock scalability patches for quite some time to get around this problem. The adopted solution is typically to make the queue lock a pointer to a driver local lock, so the driver has full control of the granularity and scope of the lock. This solution was adopted from the 2.5 kernel, as we'll see later. But this is another case where driver writers often need to differentiate between vendor and vanilla kernels.

1.4 API problems

Looking at the block layer as a whole (including both ends of the spectrum, the producers and consumers of the IO units going through the block layer), it is a typical example of code that has been hacked into existence without much thought to design. When things broke or new features were needed, they had been grafted into the existing mess. No well defined interface exists between file system and block layer, except a few scattered functions. Controlling IO unit flow from IO scheduler to driver was impossible: 2.4 exposes the IO scheduler data structures (the `->queue_head` linked list used for queuing) directly to the driver. This fact alone makes it virtually impossible to implement more clever IO scheduling in 2.4. Even the recently (in the 2.4.20's) added lower latency work was horrible to work with because of this lack of boundaries. Verifying correctness of the code is extremely difficult; peer review of the code likewise, since a reviewer must be intimate with the block layer structures to follow the code.

Another example on lack of clear direction is

the partition remapping. In 2.4, it's the driver's responsibility to resolve partition mappings. A given request contains a device and sector offset (i.e. `/dev/hda4`, sector 128) and the driver must map this to an absolute device offset before sending it to the hardware. Not only does this cause duplicate code in the drivers, it also means the IO scheduler has no knowledge of the real device mapping of a particular request. This adversely impacts IO scheduling whenever partitions aren't laid out in strict ascending disk order, since it causes the io scheduler to make the wrong decisions when ordering io.

2 2.6 Block layer

The above observations were the initial kick off for the 2.5 block layer patches. To solve some of these issues the block layer needed to be turned inside out, breaking basically anything-io along the way.

2.1 bio

Given that `struct buffer_head` was one of the problems, it made sense to start from scratch with an IO unit that would be agreeable to the upper layers as well as the drivers. The main criteria for such an IO unit would be something along the lines of:

1. Must be able to contain an arbitrary amount of data, as much as the hardware allows. Or as much that makes *sense* at least, with the option of easily pushing this boundary later.
2. Must work equally well for pages that have a virtual mapping as well as ones that do not.
3. When entering the IO scheduler and driver, IO unit must point to an absolute location on disk.
4. Must be able to stack easily for IO stacks such as raid and device mappers. This includes full redirect stacking like in 2.4, as well as partial redirections.

Once the primary goals for the IO structure were laid out, the `struct bio` was born. It was decided to base the layout on a scatter-gather type setup, with the `bio` containing a map of pages. If the map count was made flexible, items 1 and 2 on the above list were already solved. The actual implementation involved splitting the data container from the `bio` itself into a `struct bio_vec` structure. This was mainly done to ease allocation of the structures so that `sizeof(struct bio)` was always constant. The `bio_vec` structure is simply a tuple of `{page, length, offset}`, and the `bio` can be allocated with room for anything from 1 to `BIO_MAX_PAGES`. Currently Linux defines that as 256 pages, meaning we can support up to 1MiB of data in a single `bio` for a system with 4KiB page size. At the time of implementation, 1MiB was a good deal beyond the point where increasing the IO size further didn't yield better performance or lower CPU usage. It also has the added bonus of making the `bio_vec` fit inside a single page, so we avoid higher order memory allocations (`sizeof(struct bio_vec)` == 12 on 32-bit, 16 on 64-bit) in the IO path. This is an important point, as it eases the pressure on the memory allocator. For swapping or other low memory situations, we ideally want to stress the allocator as little as possible.

Different hardware can support different sizes of io. Traditional parallel ATA can do a maximum of 128KiB per request, qlogicfc SCSI doesn't like more than 32KiB, and lots of high end controllers don't impose a significant limit on max IO size but may restrict the maximum number of segments that one IO may be composed of. Additionally, software raid or de-

vice mapper stacks may like special alignment of IO or the guarantee that IO won't cross stripe boundaries. All of this knowledge is either impractical or impossible to statically advertise to submitters of io, so an easy interface for populating a `bio` with pages was essential if supporting large IO was to become practical. The current solution is `int bio_add_page()` which attempts to add a single page (full or partial) to a `bio`. It returns the amount of bytes successfully added. Typical users of this function continue adding pages to a `bio` until it fails—then it is submitted for IO through `submit_bio()`, a new `bio` is allocated and populated until all data has gone out. `int bio_add_page()` uses statically defined parameters inside the request queue to determine how many pages can be added, and attempts to query a registered `merge_bvec_fn` for dynamic limits that the block layer cannot know about.

Drivers hooking into the block layer before the IO scheduler[1] deal with `struct bio` directly, as opposed to the `struct request` that are output after the IO scheduler. Even though the page addition API guarantees that they never need to be able to deal with a `bio` that is too big, they still have to manage local splits at sub-page granularity. The API was defined that way to make it easier for IO submitters to manage, so they don't have to deal with sub-page splits. 2.6 block layer defines two ways to deal with this situation—the first is the general clone interface. `bio_clone()` returns a clone of a `bio`. A clone is defined as a private copy of the `bio` itself, but with a shared `bio_vec` page map list. Drivers can modify the cloned `bio` and submit it to a different device without duplicating the data. The second interface is tailored specifically to single page splits and was written by kernel raid maintainer Neil Brown. The main function is `bio_split()` which returns a `struct bio_pair` describing the two parts of the original `bio`. The two `bio`'s can then be submitted separately by the driver.

[1] Also known as at `make_request` time.

2.2 Partition remapping

Partition remapping is handled inside the IO stack before going to the driver, so that both drivers and IO schedulers have immediate full knowledge of precisely where data should end up. The device unfolding is done automatically by the same piece of code that resolves full `bio` redirects. The worker function is `blk_partition_remap()`.

2.3 Barriers

Another feature that found its way to some vendor kernels is IO barriers. A barrier is defined as a piece of IO that is guaranteed to:

- Be on platter (or safe storage at least) when completion is signaled.
- Not proceed any previously submitted io.
- Not be proceeded by later submitted io.

The feature is handy for journalled file systems, fsync, and any sort of cache bypassing IO[2] where you want to provide guarantees on data order and correctness. The 2.6 code isn't even complete yet or in the Linus kernels, but it has made its way to Andrew Morton's -mm tree which is generally considered a staging area for features. This section describes the code so far.

The first type of barrier supported is a soft barrier. It isn't of much use for data integrity applications, since it merely implies ordering inside the IO scheduler. It is signaled with the `REQ_SOFTBARRIER` flag inside `struct request`. A stronger barrier is the

[2] Such types of IO include `O_DIRECT` or raw.

hard barrier. From the block layer and IO scheduler point of view, it is identical to the soft variant. Drivers need to know about it though, so they can take appropriate measures to correctly honor the barrier. So far the ide driver is the only one supporting a full, hard barrier. The issue was deemed most important for journalled desktop systems, where the lack of barriers and risk of crashes / power loss coupled with ide drives generally always defaulting to write back caching caused significant problems. Since the ATA command set isn't very intelligent in this regard, the ide solution adopted was to issue pre- and post flushes when encountering a barrier.

The hard and soft barrier share the feature that they are both tied to a piece of data (a `bio`, really) and cannot exist outside of data context. Certain applications of barriers would really like to issue a disk flush, where finding out which piece of data to attach it to is hard or impossible. To solve this problem, the 2.6 barrier code added the `blkdev_issue_flush()` function. The block layer part of the code is basically tied to a queue hook, so the driver issues the flush on its own. A helper function is provided for SCSI type devices, using the generic SCSI command transport that the block layer provides in 2.6 (more on this later). Unlike the queued data barriers, a barrier issued with `blkdev_issue_flush()` works on all interesting drivers in 2.6 (IDE, SCSI, SATA). The only missing bits are drivers that don't belong to one of these classes—things like **CISS** and **DAC960**.

2.4 IO Schedulers

As mentioned in section 1.1, there are a number of known problems with the default 2.4 IO scheduler and IO scheduler interface (or lack thereof). The idea to base latency on a unit of data (sectors) rather than a time based unit is hard to tune, or requires auto-tuning at runtime and this never really worked out. Fixing the runtime problems with `elevator_linus` is next to impossible due to the data structure exposing problem. So before being able to tackle any problems in that area, a neat API to the IO scheduler had to be defined.

2.4.1 Defined API

In the spirit of avoiding over-design[3], the API was based on initial adaption of `elevator_linus`, but has since grown quite a bit as newer IO schedulers required more entry points to exploit their features.

The core function of an IO scheduler is, naturally, insertion of new io units and extraction of ditto from drivers. So the first 2 API functions are defined, `next_req_fn` and `add_req_fn`. If you recall from section 1.1, a new IO unit is first attempted merged into an existing request in the IO scheduler queue. And if this fails and the newly allocated request has raced with someone else adding an adjacent IO unit to the queue in the mean time, we also attempt to merge `struct requests`. So 2 more functions were added to cater to these needs, `merge_fn` and `merge_req_fn`. Cleaning up after a successful merge is done through `merge_cleanup_fn`. Finally, a defined IO scheduler can provide init and exit functions, should it need to perform any duties during queue init or shutdown.

The above described the IO scheduler API as of 2.5.1, later on more functions were added to further abstract the IO scheduler away from the block layer core. More details may be found in the `struct elevator_s` in `<linux/elevator.h>` kernel include file.

[3] Some might, rightfully, claim that this is worse than no design

2.4.2 deadline

In kernel 2.5.39, `elevator_linus` was finally replaced by something more appropriate, the *deadline* IO scheduler. The principles behind it are pretty straight forward — new requests are assigned an expiry time in milliseconds, based on data direction. Internally, requests are managed on two different data structures. The sort list, used for inserts and front merge lookups, is based on a red-black tree. This provides **O(log n)** runtime for both insertion and lookups, clearly superior to the doubly linked list. Two FIFO lists exist for tracking request expiry times, using a double linked list. Since strict FIFO behavior is maintained on these two lists, they run in **O(1)** time. For back merges it is important to maintain good performance as well, as they dominate the total merge count due to the layout of files on disk. So **deadline** added a merge hash for back merges, ideally providing **O(1)** runtime for merges. Additionally, **deadline** adds a one-hit merge cache that is checked even before going to the hash. This gets surprisingly good hit rates, serving as much as 90% of the merges even for heavily threaded io.

Implementation details aside, **deadline** continues to build on the fact that the fastest way to access a single drive, is by scanning in the direction of ascending sector. With its superior runtime performance, **deadline** is able to support very large queue depths without suffering a performance loss or spending large amounts of time in the kernel. It also doesn't suffer from latency problems due to increased queue sizes. When a request expires in the FIFO, **deadline** jumps to that disk location and starts serving IO from there. To prevent accidental seek storms (which would further cause us to miss deadlines), **deadline** attempts to serve a number of requests from that location before jumping to the next expired request. This means that the assigned request deadlines are soft, not a specific hard target that must be met.

2.4.3 Anticipatory IO scheduler

While **deadline** works very well for most workloads, it fails to observe the natural dependencies that often exist between synchronous reads. Say you want to list the contents of a directory—that operation isn't merely a single sync read, it consists of a number of reads where only the completion of the final request will give you the directory listing. With **deadline**, you could get decent performance from such a workload in presence of other IO activities by assigning very tight read deadlines. But that isn't very optimal, since the disk will be serving other requests in between the dependent reads causing a potentially disk wide seek every time. On top of that, the tight deadlines will decrease performance on other io streams in the system.

Nick Piggin implemented an anticipatory IO scheduler [Iyer] during 2.5 to explore some interesting research in this area. The main idea behind the anticipatory IO scheduler is a concept called *deceptive idleness*. When a process issues a request and it completes, it might be ready to issue a new request (possibly close by) immediately. Take the directory listing example from above—it might require 3–4 IO operations to complete. When each of them completes, the process[4] is ready to issue the next one almost instantly. But the traditional io scheduler doesn't pay any attention to this fact, the new request must go through the IO scheduler and wait its turn. With **deadline**, you would have to typically wait 500 milliseconds for each read, if the queue is held busy by other processes. The result is poor interactive performance for each process, even though overall throughput might be acceptable or even good.

[4] Or the kernel, on behalf of the process.

Instead of moving on to the next request from an unrelated process immediately, the anticipatory IO scheduler (hence forth known as **AS**) opens a small window of opportunity for that process to submit a new IO request. If that happens, **AS** gives it a new chance and so on. Internally it keeps a decaying histogram of IO *think times* to help the anticipation be as accurate as possible.

Internally, **AS** is quite like **deadline**. It uses the same data structures and algorithms for sorting, lookups, and FIFO. If the think time is set to 0, it is very close to **deadline** in behavior. The only differences are various optimizations that have been applied to either scheduler allowing them to diverge a little. If **AS** is able to reliably predict when waiting for a new request is worthwhile, it gets phenomenal performance with excellent interactiveness. Often the system throughput is sacrificed a little bit, so depending on the workload **AS** might not be the best choice always. The IO storage hardware used, also plays a role in this—a non-queuing ATA hard drive is a much better fit than a SCSI drive with a large queuing depth. The SCSI firmware reorders requests internally, thus often destroying any accounting that **AS** is trying to do.

2.4.4 CFQ

The third new IO scheduler in 2.6 is called **CFQ**. It's loosely based on the ideas on stochastic fair queuing (SFQ [McKenney]). **SFQ** is fair as long as its hashing doesn't collide, and to avoid that, it uses a continually changing hashing function. Collisions can't be completely avoided though, frequency will depend entirely on workload and timing. **CFQ** is an acronym for completely fair queuing, attempting to get around the collision problem that **SFQ** suffers from. To do so, **CFQ** does away with the fixed number of buckets that processes can be placed in. And using regular hashing technique to find the appropriate bucket in case of collisions, fatal collisions are avoided.

CFQ deviates radically from the concepts that **deadline** and **AS** is based on. It doesn't assign deadlines to incoming requests to maintain fairness, instead it attempts to divide bandwidth equally among classes of processes based on some correlation between them. The default is to hash on thread group id, tgid. This means that bandwidth is attempted distributed equally among the processes in the system. Each class has its own request sort and hash list, using red-black trees again for sorting and regular hashing for back merges. When dealing with writes, there is a little catch. A process will almost never be performing its own writes—data is marked dirty in context of the process, but write back usually takes place from the *pdflush* kernel threads. So **CFQ** is actually dividing read bandwidth among processes, while treating each pdflush thread as a separate process. Usually this has very minor impact on write back performance. Latency is much less of an issue with writes, and good throughput is very easy to achieve due to their inherent asynchronous nature.

2.5 Request allocation

Each block driver in the system has at least one `request_queue_t` request queue structure associated with it. The recommended setup is to assign a queue to each logical spindle. In turn, each request queue has a `struct request_list` embedded which holds free `struct request` structures used for queuing io. 2.4 improved on this situation from 2.2, where a single global free list was available to add one per queue instead. This free list was split into two sections of equal size, for reads and writes, to prevent either

direction from starving the other[5]. 2.4 statically allocated a big chunk of requests for each queue, all residing in the precious low memory of a machine. The combination of **O(N)** runtime and statically allocated request structures firmly prevented any real world experimentation with large queue depths on 2.4 kernels.

2.6 improves on this situation by dynamically allocating request structures on the fly instead. Each queue still maintains its request free list like in 2.4. However it's also backed by a memory pool[6] to provide deadlock free allocations even during swapping. The more advanced io schedulers in 2.6 usually back each request by its own private request structure, further increasing the memory pressure of each request. Dynamic request allocation lifts some of this pressure as well by pushing that allocation inside two hooks in the IO scheduler API—`set_req_fn` and `put_req_fn`. The latter handles the later freeing of that data structure.

2.6 Plugging

For the longest time, the Linux block layer has used a technique dubbed *plugging* to increase IO throughput. In its simplicity, plugging works sort of like the plug in your tub drain—when IO is queued on an initially empty queue, the queue is plugged. Only when someone asks for the completion of some of the queued IO is the plug yanked out, and io is allowed to drain from the queue. So instead of submitting the first immediately to the driver, the block layer allows a small buildup of requests. There's nothing wrong with the principle of plugging, and it has been shown to work well for a number of workloads. However, the block layer maintains a global list of plugged queues inside the `tq_disk` task queue. There are three main problems with this approach:

1. It's impossible to go backwards from the file system and find the specific queue to unplug.
2. Unplugging one queue through `tq_disk` unplugs all plugged queues.
3. The act of plugging and unplugging touches a global lock.

All of these adversely impact performance. These problems weren't really solved until late in 2.6, when Intel reported a huge scalability problem related to unplugging [Chen] on a 32 processor system. 93% of system time was spent due to contention on `blk_plug_lock`, which is the 2.6 direct equivalent of the 2.4 `tq_disk` embedded lock. The proposed solution was to move the plug lists to a per-CMU structure. While this would solve the contention problems, it still leaves the other 2 items on the above list unsolved.

So work was started to find a solution that would fix all problems at once, and just generally Feel Right. 2.6 contains a link between the block layer and write out paths which is embedded inside the queue, a `struct backing_dev_info`. This structure holds information on read-ahead and queue congestion state. It's also possible to go from a `struct page` to the backing device, which may or may not be a block device. So it would seem an obvious idea to move to a backing device unplugging scheme instead, getting rid of the global `blk_run_queues()` unplugging. That solution would fix all three issues at once—there would be no global way to unplug all devices, only target specific unplugs, and the backing device gives us a mapping from page to queue. The code was rewritten to do just that, and provide unplug functionality going from a specific `struct block_device`, page, or backing device. Code and interface was much superior to the existing code base,

[5] In reality, to prevent writes for consuming all requests.

[6] `mempool_t` interface from Ingo Molnar.

and results were truly amazing. Jeremy Higdon tested on an 8-way IA64 box [Higdon] and got 75–80 thousand IOPS on the stock kernel at 100% CPU utilization, 110 thousand IOPS with the per-CPU Intel patch also at full CPU utilization, and finally 200 thousand IOPS at merely 65% CPU utilization with the backing device unplugging. So not only did the new code provide a huge speed increase on this machine, it also went from being CPU to IO bound.

2.6 also contains some additional logic to unplug a given queue once it reaches the point where waiting longer doesn't make much sense. So where 2.4 will always wait for an explicit unplug, 2.6 can trigger an unplug when one of two conditions are met:

1. The number of queued requests reach a certain limit, `q->unplug_thresh`. This is device tweak able and defaults to 4.
2. When the queue has been idle for `q->unplug_delay`. Also device tweak able, and defaults to 3 milliseconds.

The idea is that once a certain number of requests have accumulated in the queue, it doesn't make much sense to continue waiting for more—there is already an adequate number available to keep the disk happy. The time limit is really a last resort, and should rarely trigger in real life. Observations on various work loads have verified this. More than a handful or two timer unplugs per minute usually indicates a kernel bug.

2.7 SCSI command transport

An annoying aspect of CD writing applications in 2.4 has been the need to use ide-scsi, necessitating the inclusion of the entire SCSI stack for only that application. With the clear majority of the market being ATAPI hardware, this becomes even more silly. ide-scsi isn't without its own class of problems either—it lacks the ability to use DMA on certain writing types. CDDA audio ripping is another application that thrives with ide-scsi, since the native uniform cdrom layer interface is less than optimal (put mildly). It doesn't have DMA capabilities at all.

2.7.1 Enhancing struct request

The problem with 2.4 was the lack of ability to generically send SCSI "like" commands to devices that understand them. Historically, only file system read/write requests could be submitted to a driver. Some drivers made up faked requests for other purposes themselves and put then on the queue for their own consumption, but no defined way of doing this existed. 2.6 adds a new request type, marked by the `REQ_BLOCK_PC` bit. Such a request can be either backed by a `bio` like a file system request, or simply has a data and length field set. For both types, a SCSI command data block is filled inside the request. With this infrastructure in place and appropriate update to drivers to understand these requests, it's a cinch to support a much better direct-to-device interface for burning.

Most applications use the SCSI sg API for talking to devices. Some of them talk directly to the `/dev/sg*` special files, while (most) others use the `SG_IO` ioctl interface. The former requires a yet unfinished driver to transform them into block layer requests, but the latter can be readily intercepted in the kernel and routed directly to the device instead of through the SCSI layer. Helper functions were added to make burning and ripping even faster, providing DMA for all applications and without copying data between kernel and user space at all. So the zero-copy DMA burning was possible, and this even without changing most ap-

plications.

3 Linux-2.7

The 2.5 development cycle saw the most massively changed block layer in the history of Linux. Before 2.5 was opened, Linus had clearly expressed that one of the most important things that needed doing, was the block layer update. And indeed, the very first thing merged was the complete bio patch into 2.5.1-pre2. At that time, no more than a handful drivers compiled (let alone worked). The 2.7 changes will be nowhere as severe or drastic. A few of the possible directions will follow in the next few sections.

3.1 IO Priorities

Prioritized IO is a very interesting area that is sure to generate lots of discussion and development. It's one of the missing pieces of the complete resource management puzzle that several groups of people would very much like to solve. People running systems with many users, or machines hosting virtual hosts (or completed virtualized environments) are dying to be able to provide some QOS guarantees. Some work was already done in this area, so far nothing complete has materialized. The CKRM [CKRM] project spear headed by IBM is an attempt to define global resource management, including io. They applied a little work to the **CFQ** IO scheduler to provide equal bandwidth between resource management classes, but at no specific priorities. Currently I have a **CFQ** patch that is 99% complete that provides full priority support, using the IO contexts introduced by **AS** to manage fair sharing over the full time span that a process exists[7]. This works well enough, but only works for that specific IO scheduler. A nicer solution would be to create a scheme that works independently of the io scheduler used. That would require a rethinking of the IO scheduler API.

[7] CFQ currently tears down class structures as soon as it is empty, it doesn't persist over process life time.

3.2 IO Scheduler switching

Currently Linux provides no less than 4 IO schedulers—the 3 mentioned, plus a forth dubbed **noop**. The latter is a simple IO scheduler that does no request reordering, no latency management, and always merges whenever it can. Its area of application is mainly highly intelligent hardware with huge queue depths, where regular request reordering doesn't make sense. Selecting a specific IO scheduler can either be done by modifying the source of a driver and putting the appropriate calls in there at queue init time, or globally for any queue by passing the `elevator=xxx` boot parameter. This makes it impossible, or at least very impractical, to benchmark different IO schedulers without many reboots or recompiles. Some way to switch IO schedulers per queue and on the fly is desperately needed. Freezing a queue and letting IO drain from it until it's empty (pinning new IO along the way), and then shutting down the old io scheduler and moving to the new scheduler would not be so hard to do. The queues expose various sysfs variables already, so the logical approach would simply be to:

```
# echo deadline > \
    /sys/block/hda/queue/io_scheduler
```

A simple but effective interface. At least two patches doing something like this were already proposed, but nothing was merged at that time.

4 Final comments

The block layer code in 2.6 has come a long way from the rotted 2.4 code. New features

bring it more up-to-date with modern hardware, and completely rewritten from scratch core provides much better scalability, performance, and memory usage benefiting any machine from small to really huge. Going back a few years, I heard constant complaints about the block layer and how much it sucked and how outdated it was. These days I rarely hear anything about the current state of affairs, which usually means that it's doing pretty well indeed. 2.7 work will mainly focus on feature additions and driver layer abstractions (our concept of IDE layer, SCSI layer etc will be severely shook up). Nothing that will wreak havoc and turn everything inside out like 2.5 did. Most of the 2.7 work mentioned above is pretty light, and could easily be back ported to 2.6 once it has been completed and tested. Which is also a good sign that nothing really radical or risky is missing. So things are settling down, a sign of stability.

References

[Iyer] Sitaram Iyer and Peter Druschel, *Anticipatory scheduling: A disk scheduling framework to overcome deceptive idleness in synchronous I/O*, 18th ACM Symposium on Operating Systems Principles, `http://www.cs.rice.edu/~ssiyer/r/antsched/antsched.ps.gz`, 2001

[McKenney] Paul E. McKenney, *Stochastic Fairness Queuing*, INFOCOM `http://rdrop.com/users/paulmck/paper/sfq.2002.06.04.pdf`, 1990

[Chen] Kenneth W. Chen, *per-cpu blk_plug_list*, Linux kernel mailing list `http://www.ussg.iu.edu/hypermail/linux/kernel/0403.0/0179.html`, 2004

[Higdon] Jeremy Higdon, *Re: [PATCH] per-backing dev unplugging #2*, Linux kernel mailing list `http://marc.theaimsgroup.com/?l=linux-kernel&m=107941470424309&w=2`, 2004

[CKRM] IBM, *Class-based Kernel Resource Management (CKRM)*, `http://ckrm.sf.net`, 2004

[Bhattacharya] Suparna Bhattacharya, *Notes on the Generic Block Layer Rewrite in Linux 2.5*, General discussion, `Documentation/block/biodoc.txt`, 2002

Linux AIO Performance and Robustness for Enterprise Workloads

Suparna Bhattacharya, IBM (suparna@in.ibm.com)
John Tran, IBM (jbtran@ca.ibm.com)
Mike Sullivan, IBM (mksully@us.ibm.com)
Chris Mason, SUSE (mason@suse.com)

1 Abstract

In this paper we address some of the issues identified during the development and stabilization of Asynchronous I/O (AIO) on Linux 2.6.

We start by describing improvements made to optimize the throughput of streaming buffered filesystem AIO for microbenchmark runs. Next, we discuss certain tricky issues in ensuring data integrity between AIO Direct I/O (DIO) and buffered I/O, and take a deeper look at synchronized I/O guarantees, concurrent I/O, write-ordering issues and the improvements resulting from radix-tree based write-back changes in the Linux VFS.

We then investigate the results of using Linux 2.6 filesystem AIO on the performance metrics for certain enterprise database workloads which are expected to benefit from AIO, and mention a few tips on optimizing AIO for such workloads. Finally, we briefly discuss the issues around workloads that need to combine asynchronous disk I/O and network I/O.

2 Introduction

AIO enables a single application thread to overlap processing with I/O operations for better utilization of CPU and devices. AIO can improve the performance of certain kinds of I/O intensive applications like databases, web-servers and streaming-content servers. The use of AIO also tends to help such applications adapt and scale more smoothly to varying loads.

2.1 Overview of kernel AIO in Linux 2.6

The Linux 2.6 kernel implements in-kernel support for AIO. A low-level native AIO system call interface is provided that can be invoked directly by applications or used by library implementations to build POSIX/SUS semantics. All discussion hereafter in this paper pertains to the native kernel AIO interfaces.

Applications can submit one or more I/O requests asynchronously using the `io_submit()` system call, and obtain completion notification using the `io_getevents()` system call. Each I/O request specifies the operation (typically read/write), the file descriptor and the parameters for the operation (e.g., file offset, buffer). I/O requests are associated with the completion queue (ioctx) they were submitted against. The results of I/O are reported as completion events on this queue, and reaped using `io_getevents()`.

The design of AIO for the Linux 2.6 kernel has been discussed in [1], including the motivation

behind certain architectural choices, for example:

- Sharing a common code path for AIO and regular I/O
- A retry-based model for AIO continuations across blocking points in the case of buffered filesystem AIO (currently implemented as a set of patches to the Linux 2.6 kernel) where worker threads take on the caller's address space for executing retries involving access to user-space buffers.

2.2 Background on retry-based AIO

The retry-based model allows an AIO request to be executed as a series of non-blocking iterations. Each iteration retries the remaining part of the request from where the last iteration left off, re-issuing the corresponding AIO filesystem operation with modified arguments representing the remaining I/O. The retries are "kicked" via a special AIO waitqueue callback routine, `aio_wake_function()`, which replaces the default waitqueue entry used for blocking waits.

The high-level retry infrastructure is responsible for running the iterations in the address space context of the caller, and ensures that only one retry instance is active at a given time. This relieves the fops themselves from having to deal with potential races of that sort.

2.3 Overview of the rest of the paper

In subsequent sections of this paper, we describe our experiences in addressing several issues identified during the optimization and stabilization efforts related to the kernel AIO implementation for Linux 2.6, mainly in the area of disk- or filesystem-based AIO.

We observe, for example, how I/O patterns generated by the common VFS code paths used by regular and retry-based AIO could be non-optimal for streaming AIO requests, and we describe the modifications that address this finding. A different set of problems that has seen some development activity are the races, exposures and potential data-integrity concerns between direct and buffered I/O, which become especially tricky in the presence of AIO. Some of these issues motivated Andrew Morton's modified page-writeback design for the VFS using tagged radix-tree lookups, and we discuss the implications for the AIO `O_SYNC` write implementation. In general, disk-based filesystem AIO requirements for database workloads have been a guiding consideration in resolving some of the trade-offs encountered, and we present some initial performance results for such workloads. Lastly, we touch upon potential approaches to allow processing of disk-based AIO and communications I/O within a single event loop.

3 Streaming AIO reads

3.1 Basic retry pattern for single AIO read

The retry-based design for buffered filesystem AIO read works by converting each blocking wait for read completion on a page into a *retry exit*. The design queues an asynchronous notification callback and returns the number of bytes for which the read has completed so far without blocking. Then, when the page becomes up-to-date, the callback kicks off a retry continuation in task context. This retry continuation invokes the same filesystem read operation again using the caller's address space, but this time with arguments modified to reflect the remaining part of the read request.

For example, given a 16KB read request starting at offset 0, where the first 4KB is already in cache, one might see the following sequence of retries (in the absence of readahead):

```
first time:
   fop->aio_read(fd, 0, 16384) = 4096
and when read completes for the second page:
   fop->aio_read(fd, 4096, 12288) = 4096
and when read completes for the third page:
   fop->aio_read(fd, 8192, 8192) = 4096
and when read completes for the fourth page:
   fop->aio_read(fd, 12288, 4096) = 4096
```

3.2 Impact of readahead on single AIO read

Usually, however, the readahead logic attempts to batch read requests in advance. Hence, more I/O would be seen to have completed at each retry. The logic attempts to predict the optimal readahead window based on state it maintains about the sequentiality of past read requests on the same file descriptor. Thus, given a maximum readahead window size of 128KB, the sequence of retries would appear to be more like the following example, which results in significantly improved throughput:

```
first time:
   fop->aio_read(fd, 0, 16384) = 4096,
     after issuing readahead
     for 128KB/2 = 64KB
and when read completes for the above I/O:
   fop->aio_read(fd, 4096, 12288) = 12288
```

Notice that care is taken to ensure that readaheads are not repeated during retries.

3.3 Impact of readahead on streaming AIO reads

In the case of streaming AIO reads, a sequence of AIO read requests is issued on the same file descriptor, where subsequent reads are submitted without waiting for previous requests to complete (contrast this with a sequence of synchronous reads).

Interestingly, we encountered a significant throughput degradation as a result of the interplay of readahead and streaming AIO reads. To see why, consider the retry sequence for streaming random AIO read requests of 16KB, where `o1, o2, o3, ...` refer to the random offsets where these reads are issued:

```
first time:
  fop->aio_read(fd, o1, 16384) = -EIOCBRETRY,
    after issuing readahead for 64KB
    as the readahead logic sees the first page
    of the read
  fop->aio_read(fd, o2, 16384) = -EIOCBRETRY,
    after issuing readahead for 8KB (notice
    the shrinkage of the readahead window
    because of non-sequentiality seen by the
    readahead logic)
  fop->aio_read(fd, o3, 16384) = -EIOCBRETRY,
    after maximally shrinking the readahead
    window, turning off readahead and issuing
    4KB read in the slow path
  fop->aio_read(fd, o4, 16384) = -EIOCBRETRY,
    after issuing 4KB read in the slow path
.
.
and when read completes for o1
  fop->aio_read(fd, o1, 16384) = 16384
and when read completes for o2
  fop->aio_read(fd, o2, 16384) = 8192
and when read completes for o3
  fop->aio_read(fd, o3, 16384) = 4096
and when read completes for o4
  fop->aio_read(fd, o3, 16384) = 4096
.
.
```

In steady state, this amounts to a maximally-shrunk readahead window with 4KB reads at random offsets being issued serially one at a time on a slow path, causing seek storms and driving throughputs down severely.

3.4 Upfront readahead for improved streaming AIO read throughputs

To address this issue, we made the readahead logic aware of the sequentiality of all pages in a single read request upfront—before submitting the next read request. This resulted in a more desirable outcome as follows:

```
  fop->aio_read(fd, o1, 16384) = -EIOCBRETRY,
    after issuing readahead for 64KB
    as the readahead logic sees all the 4
    pages for the read
  fop->aio_read(fd, o2, 16384) = -EIOCBRETRY,
    after issuing readahead for 20KB, as the
    readahead logic sees all 4 pages of the
    read (the readahead window shrinks to
    4+1=5 pages)
```

```
  fop->aio_read(fd, o3, 16384) = -EIOCBRETRY,
    after issuing readahead for 20KB, as the
    readahead logic sees all 4 pages of the
    read (the readahead window is maintained
    at 4+1=5 pages)
        .
        .
and when read completes for o1
  fop->aio_read(fd, o1, 16384) = 16384
and when read completes for o2
  fop->aio_read(fd, o2, 16384) = 16384
and when read completes for o3
  fop->aio_read(fd, o3, 16384) = 16384
        .
        .
```

3.5 Upfront readahead and sendfile regressions

At first sight it appears that upfront readahead is a reasonable change for all situations, since it immediately passes to the readahead logic the entire size of the request. However, it has the unintended, potential side-effect of losing pipelining benefits for really large reads, or operations like sendfile which involve post processing I/O on the contents just read. One way to address this is to clip the maximum size of upfront readahead to the maximum readahead setting for the device. To see why even that may not suffice for certain situations, let us take a look at the following sequence for a webserver that uses non-blocking sendfile to serve a large (2GB) file.

```
  sendfile(fd, 0, 2GB, fd2) = 8192,
    tells readahead about up to 128KB
    of the read
  sendfile(fd, 8192, 2GB - 8192, fd2) = 8192,
    tells readahead about 8KB - 132KB
    of the read
  sendfile(fd, 16384, 2GB - 16384, fd2) = 8192,
    tells readahead about 16KB-140KB
    of the read
       ...
```

This confuses the readahead logic about the I/O pattern which appears to be 0–128K, 8K–132K, 16K–140K instead of clear sequentiality from 0–2GB that is really appropriate.

To avoid such unanticipated issues, upfront readahead required a special case for AIO alone, limited to the maximum readahead setting for the device.

3.6 Streaming AIO read microbenchmark comparisons

We explored streaming AIO throughput improvements with the retry-based AIO implementation and optimizations discussed above, using a custom microbenchmark called aio-stress [2]. aio-stress issues a stream of AIO requests to one or more files, where one can vary several parameters including I/O unit size, total I/O size, depth of iocbs submitted at a time, number of concurrent threads, and type and pattern of I/O operations, and reports the overall throughput attained.

The hardware included a 4-way 700MHz Pentium® III machine with 512MB of RAM and a 1MB L2 cache. The disk subsystem used for the I/O tests consisted of an Adaptec AIC7896/97 Ultra2 SCSI controller connected to a disk enclosure with six 9GB disks, one of which was configured as an ext3 filesystem with a block size of 4KB for testing.

The runs compared aio-stress throughputs for streaming random buffered I/O reads (i.e., without `O_DIRECT`), with and without the previously described changes. All the runs were for the case where the file was not already cached in memory. The above graph summarizes how the results varied across individual request sizes of 4KB to 64KB, where I/O was targeted to a single file of size 1GB, the depth of iocbs outstanding at a time being 64KB. A third run was performed to find out how the results compared with equivalent runs using AIO-DIO.

With the changes applied, the results showed an approximate 2x improvement across all block sizes, bringing throughputs to levels that match the corresponding results using AIO-DIO.

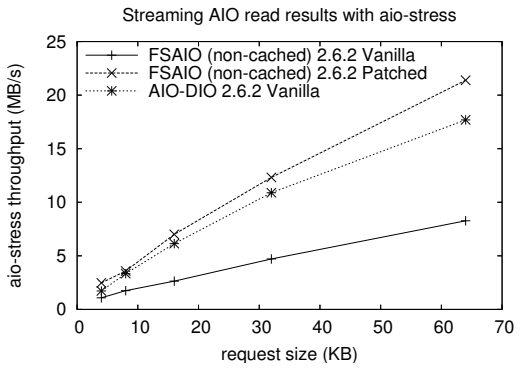


Figure 1: Comparisons of streaming random AIO read throughputs

4 AIO DIO vs cached I/O integrity issues

4.1 DIO vs buffered races

Stephen Tweedie discovered several races between DIO and buffered I/O to the same file [3]. These races could lead to potential stale-data exposures and even data-integrity issues. Most instances were related to situations when in-core meta-data updates were visible before actual instantiation or resetting of corresponding data blocks on disk. Problems could also arise when meta-data updates were not visible to other code paths that could simultaneously update meta-data as well. The races mainly affected sparse files due to the lack of atomicity between the file flush in the DIO paths and actual data block accesses.

The solution that Stephen Tweedie came up with, and which Badari Pulavarty reported to Linux 2.6, involved protecting block lookups and meta-data updates with the inode semaphore (`i_sem`) in DIO paths for both read and write, atomically with the file flush. Overwriting of sparse blocks in the DIO write path was modified to fall back to buffered writes. Finally, an additional semaphore (`i_alloc_sem`) was introduced to lock out deallocation of blocks by a truncate while DIO was in progress. The semaphore was implemented held in shared mode by DIO and in exclusive mode by truncate.

Note that handling the new locking rules (i.e., lock ordering of `i_sem` first and then `i_alloc_sem`) while allowing for filesystem-specific implementations of the DIO and file-write interfaces had to be handled with some care.

4.2 AIO-DIO specific races

The inclusion of AIO in Linux 2.6 added some tricky scenarios to the above-described problems because of the potential races inherent in returning without waiting for I/O completion. The interplay of AIO-DIO writes and truncate was a particular worry as it could lead to corruption of file data; for example, blocks could get deallocated and reallocated to a new file while an AIO-DIO write to the file was still in progress. To avoid this, AIO-DIO had to return with `i_alloc_sem` held, and only release it as part of I/O completion post-processing. Notice that this also had implications for AIO cancellation.

File size updates for AIO-DIO file extends could expose unwritten blocks if they happened before I/O completed asynchronously. The case involving fallback to buffered I/O was particularly non-trivial if a single request spanned allocated and sparse regions of a file. Specifically, part of the I/O could have been initiated via DIO then continued asynchronously, while the fallback to buffered I/O occurred and signaled I/O completion to the application. The application may thus have reused its I/O buffer, overwriting it with other data and potentially causing file data corruption if writeout to disk had still been pending.

It might appear that some of these problems

could be avoided if I/O schedulers guaranteed the ordering of I/O requests issued to the same disk block. However, this isn't a simple proposition in the current architecture, especially in generalizing the design to all possible cases, including network block devices. The use of I/O barriers would be necessary and the costs may not be justified for these special-case situations.

Instead, a pragmatic approach was taken in order to address this based on the assumptions that true asynchronous behaviour was really meaningful in practice, mainly when performing I/O to already-allocated file blocks. For example, databases typically preallocate files at the time of creation, so that AIO writes during normal operation and in performance-critical paths do not extend the file or encounter sparse regions. Thus, for the sake of correctness, synchronous behaviour may be tolerable for AIO writes involving sparse regions or file extends. This compromise simplified the handling of the scenarios described earlier. AIO-DIO file extends now wait for I/O to complete and update the file size. AIO-DIO writes spanning allocated and sparse regions now wait for previously- issued DIO for that request to complete before falling back to buffered I/O.

5 Concurrent I/O with synchronized write guarantees

An application opts for synchronized writes (by using the `O_SYNC` option on file open) when the I/O must be committed to disk before the write request completes. In the case of DIO, writes directly go to disk anyway. For buffered I/O, data is first copied into the page cache and later written out to disk; if synchronized I/O is specified then the request returns only after the writeout is complete.

An application might also choose to synchronize previously-issued writes to disk by invoking fsync(), which writes back data from the page cache to disk and waits for writeout to complete before returning.

5.1 Concurrent DIO writes

DIO writes formerly held the inode semaphore in exclusive mode until write completion. This helped ensure atomicity of DIO writes and protected against potential file data corruption races with truncate. However, it also meant that multiple threads or processes submitting parallel DIOs to different parts of the same file effectively became serialized synchronously. If the same behaviour were extended to AIO (i.e., having the `i_sem` held through I/O completion for AIO-DIO writes), it would significantly degrade throughput of streaming AIO writes as subsequent write submissions would block until completion of the previous request.

With the fixes described in the previous section, such synchronous serialization is avoidable without loss of correctness, as the inode semaphore needs to be held only when looking up the blocks to write, and not while actual I/O is in progress on the data blocks. This could allow concurrent DIO writes on different parts of a file to proceed simultaneously, and efficient throughputs for streaming AIO-DIO writes.

5.2 Concurrent `O_SYNC` buffered writes

In the original writeback design in the Linux VFS, per-address space lists were maintained for dirty pages and pages under writeback for a given file. Synchronized write was implemented by traversing these lists to issue writeouts for the dirty pages and waiting for writeback to complete on the pages on the writeback list. The inode semaphore had to be held all through to avoid possibilities of livelocking on these lists as further writes streamed into the same file. While this helped maintain atomicity

of writes, it meant that parallel O_SYNC writes to different parts of the file were effectively serialized synchronously. Further, dependence on i_sem-protected state in the address space lists across I/O waits made it difficult to retry-enable this code path for AIO support.

In order to allow concurrent O_SYNC writes to be active on a file, the range of pages to be written back and waited on could instead be obtained directly through a radix-tree lookup for the range of offsets in the file that was being written out by the request [4]. This would avoid traversal of the page lists and hence the need to hold i_sem across the I/O waits. Such an approach would also make it possible to complete O_SYNC writes as a sequence of non-blocking retry iterations across the range of bytes in a given request.

5.3 Data-integrity guarantees

Background writeout threads cannot block on the inode semaphore like O_SYNC/fsync writers. Hence, with the per-address space lists writeback model, some juggling involving movement across multiple lists was required to avoid livelocks. The implementation had to make sure that pages which by chance got picked up for processing by background writeouts didn't slip from consideration when waiting for writeback to complete for a synchronized write request. The latter would be particularly relevant for ensuring synchronized-write guarantees that impacted data integrity for applications. However, as Daniel McNeil's analysis would indicate [5], getting this right required the writeback code to write and wait upon I/O and dirty pages which were initiated by other processes, and that turned out to be fairly tricky.

One solution that was explored was per-address space serialization of writeback to ensure exclusivity to synchronous writers and shared mode for background writers. It involved navigating issues with busy-waits in background writers and the code was beginning to get complicated and potentially fragile.

This was one of the problems that finally prompted Andrew Morton to change the entire VFS writeback code to use radix-tree walks instead of the per-address space pagelists. The main advantage was that avoiding the need for movement across lists during state changes (e.g., when re-dirtying a page if its buffers were locked for I/O by another process) reduced the chances of pages getting missed from consideration without the added serialization of entire writebacks.

6 Tagged radix-tree based writeback

For the radix-tree walk writeback design to perform as well as the address space lists-based approach, an efficient way to get to the pages of interest in the radix trees is required. This is especially so when there are many pages in the pagecache but only a few are dirty or under writeback. Andrew Morton solved this problem by implementing tagged radix-tree lookup support to enable lookup of dirty or writeback pages in O(log64(n)) time [6].

This was achieved by adding tag bits for each slot to each radix-tree node. If a node is tagged, then the corresponding slots on all the nodes above it in the tree are tagged. Thus, to search for a particular tag, one would keep going down sub-trees under slots which have the tag bit set until the tagged leaf nodes are accessed. A tagged gang lookup function is used for in-order searches for dirty or writeback pages within a specified range. These lookups are used to replace the per-address-space page lists altogether.

To synchronize writes to disk, a tagged radix-tree gang lookup of dirty pages in the byte-range corresponding to the write request is performed and the resulting pages are written out. Next, pages under writeback in the byte-range are obtained through a tagged radix-tree gang lookup of writeback pages, and we wait for writeback to complete on these pages (without having to hold the inode semaphore across the waits). Observe how this logic lends itself to be broken up into a series of non-blocking retry iterations proceeding in-order through the range.

The same logic can also be used for a whole file sync, by specifying a byte-range that spans the entire file.

Background writers also use tagged radix-tree gang lookups of dirty pages. Instead of always scanning a file from its first dirty page, the index where the last batch of writeout terminated is tracked so the next batch of writeouts can be started after that point.

7 Streaming AIO writes

The tagged radix-tree walk writeback approach greatly simplifies the design of AIO support for synchronized writes, as mentioned in the previous section,

7.1 Basic retry pattern for synchronized AIO writes

The retry-based design for buffered AIO `O_SYNC` writes works by converting each blocking wait for writeback completion of a page into a *retry exit*. The conversion point queues an asynchronous notification callback and returns to the caller of the filesystem's AIO write operation the number of bytes for which writeback has completed so far without blocking. Then, when writeback completes for that page, the callback kicks off a retry continuation in task context which invokes the same AIO write operation again using the caller's address space, but this time with arguments modified to reflect the remaining part of the write request.

As writeouts for the range would have already been issued the first time before the loop to wait for writeback completion, the implementation takes care not to re-dirty pages or re-issue writeouts during subsequent retries of AIO write. Instead, when the code detects that it is being called in a retry context, it simply falls through directly to the step involving wait-on-writeback for the remaining range as specified by the modified arguments.

7.2 Filtered waitqueues to avoid retry storms with hashed wait queues

Code that is in a retry-exit path (i.e., the return path following a blocking point where a retry is queued) should in general take care not to call routines that could wakeup the newly-queued retry.

One thing that we had to watch for was calls to `unlock_page()` in the retry-exit path. This could cause a redundant wakeup if an async wait-on-page writeback was just queued for that page. The redundant wakeup would arise if the kernel used the same waitqueue on unlock as well as writeback completion for a page, with the expectation that the waiter would check for the condition it was waiting for and go back to sleep if it hadn't occurred. In the AIO case, however, a wakeup of the newly-queued callback in the same code path could potentially trigger a retry storm, as retries kept triggering themselves over and over again for the wrong condition.

The interplay of `unlock_page()` and `wait_on_page_writeback()` with hashed waitqueues can get quite tricky for retries. For example, consider what happens when the following sequence in retryable code is executed at the same time for 2 pages, *px*

and *py*, which happen to hash to the same waitqueue (Table 1).

```
lock_page(p)
check condition and process
unlock_page(p)
if (wait_on_page_writeback_wq(p)
  == -EIOCBQUEUED)
    return bytes_done
```

The above code could keep cycling between spurious retries on *px* and *py* until I/O is done, wasting precious CPU time!

If we can ensure specificity of the wakeup with hashed waitqueues then this problem can be avoided. William Lee Irwin's implementation of filtered wakeup support in the recent Linux 2.6 kernels [7] achieves just that. The wakeup routine specifies a key to match before invoking the wakeup function for an entry in the waitqueue, thereby limiting wakeups to those entries which have a matching key. For page waitqueues, the key is computed as a function of the page and the condition (unlock or writeback completion) for the wakeup.

7.3 Streaming AIO write microbenchmark comparisons

The following graph compares aio-stress throughputs for streaming random buffered I/O `O_SYNC` writes, with and without the previously-described changes. The comparison was performed on the same setup used for the streaming AIO read results discussed earlier. The graph summarizes how the results varied across individual request sizes of 4KB to 64KB, where I/O was targeted to a single file of size 1GB and the depth of iocbs outstanding at a time was 64KB. A third run was performed to determine how the results compared with equivalent runs using AIO-DIO.

With the changes applied, the results showed an approximate 2x improvement across all block sizes, bringing throughputs to levels that match the corresponding results using AIO-DIO.

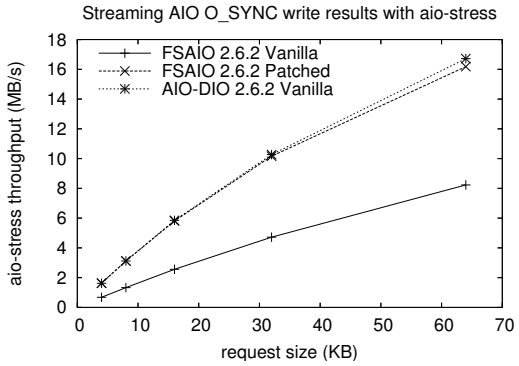


Figure 2: Comparisons of streaming random AIO write throughputs.

8 AIO performance analysis for database workloads

Large database systems leveraging AIO can show marked performance improvements compared to those systems that use synchronous I/O alone. We use IBM® DB2® Universal Database™ V8 running an online transaction processing (OLTP) workload to illustrate the performance improvement of AIO on raw devices and on filesystems.

8.1 DB2 page cleaners

A DB2 page cleaner is a process responsible for flushing dirty buffer pool pages to disk. It simulates AIO by executing asynchronously with respect to the agent processes. The number of page cleaners and their behavior can be tuned according to the demands of the system. The agents, freed from cleaning pages themselves, can dedicate their resources (e.g., processor cycles) towards processing transactions, thereby improving throughput.

```
 CPU1                                  CPU2
 lock_page(px)
 ...
 unlock_page(px)
                                       lock_page(py)
 wait_on_page_writeback_wq(px)         ...
                                       unlock_page(py) -> wakes up p1
                                                          triggering
               <-----------------------------------       a retry
lock_page(px)                         wait_on_page_writeback_wq(py)
...
unlock_page(py)  ---- wakes up py --- causes retry ---->
```

Table 1: Retry storm livelock with redundant wakeups on hashed wait queues

8.2 AIO performance analysis for raw devices

Two experiments were conducted to measure the performance benefits of AIO on raw devices for an update-intensive OLTP database workload. The workload used was derived from a TPC[8] benchmark, but is in no way comparable to any TPC results. For the first experiment, the database was configured with one page cleaner using the native Linux AIO interface. For the second experiment, the database was configured with 55 page cleaners all using the synchronous I/O interface. These experiments showed that a database, properly configured in terms of the number of page cleaners with AIO, can out-perform a properly configured database using synchronous I/O page cleaning.

For both experiments, the system configuration consisted of DB2 V8 running on a 2-way AMD Opteron system with Linux 2.6.1 installed. The disk subsystem consisted of two FAStT 700 storage servers, each with eight disk enclosures. The disks were configured as RAID-0 arrays with a stripe size of 256KB.

Table 2 shows the relative database performance with and without AIO. Higher numbers are better. The results show that the database performed 9% better when configured with one page cleaner using AIO, than when it was configured with 55 page cleaners using synchronous I/O.

Configuration	Relative Throughput
1 page cleaner with AIO	133
55 page cleaners without AIO	122

Table 2: Database performance with and without AIO.

Analyzing the I/O write patterns (see Table 3), we see that one page cleaner using AIO was sufficient to keep the buffer pools clean under a very heavy load, but that 55 page cleaners using synchronous I/O were not, as indicated by the 30% agent writes. This data suggests that more page cleaners should have been configured to improve the performance of the case with synchronous I/O. However, additional page cleaners consumed more memory, requiring a reduction in bufferpool size and thereby decreasing throughput. For the test configuration, 55 cleaners was the optimal number before memory constraints arose.

8.3 AIO performance analysis for filesystems

This section examines the performance improvements of AIO when used in conjunction with filesystems. This experiment was per-

Configuration	Page cleaner writes (%)	Agent writes (%)
1 page cleaner with AIO	100	0
55 page cleaners without AIO	70	30

Table 3: DB2 write patterns for raw device configurations.

formed using the same OLTP benchmark as in the previous section.

The test system consisted of two 1GHz AMD Opteron processors, 4GB of RAM and two QLogic 2310 FC controllers. Attached to the server was a single FAStT900 storage server and two disk enclosures with a total of 28 15K RPM 18GB drives. The Linux kernel used for the examination was 2.6.0+mm1, which includes the AIO filesystem support patches [9] discussed in this paper.

The database tables were spread across multiple ext2 filesystem partitions. Database logs were stored on a single raw partition.

Three separate tests were performed, utilizing different I/O methods for the database page cleaners.

Test 1. Synchronous (Buffered) I/O.

Test 2. Asynchronous (Buffered) I/O.

Test 3. Direct I/O.

The results are shown in Table 4 as relative commercial processing scores using synchronous I/O as the baseline (i.e., higher is better).

Looking at the efficiency of the page cleaners (see Table 5), we see that the use of AIO is more successful in keeping the buffer pools clean. In the synchronous I/O and DIO cases, the agents needed to spend more time cleaning

Configuration	Commercial Processing Scores
Synchronous I/O	100
AIO (Buffered)	113.7
DIO	111.9

Table 4: Database performance on filesystems with and without AIO.

buffer pool pages, resulting in less time processing transactions.

Configuration	Page cleaner writes (%)	Agent writes (%)
Synchronous I/O	37	63
AIO (buffered)	100	0
DIO	49	51

Table 5: DB2 write patterns for filesystem configurations.

8.4 Optimizing AIO for database workloads

Databases typically use AIO for streaming batches of random, synchronized write requests to disk (where the writes are directed to preallocated disk blocks). This has been found to improve the performance of OLTP workloads, as it helps bring down the number of dedicated threads or processes needed for flushing updated pages, and results in reduced memory footprint and better CPU utilization and scaling.

The size of individual write requests is determined by the page size used by the database. For example, a DB2 UDB installation might use a database page size of 8KB.

As observed in previous sections, the use of AIO helps reduce the number of database page cleaner processes required to keep the bufferpool clean. To keep the disk queues maximally utilized and limit contention, it may be preferable to have requests to a given disk streamed out from a single page cleaner. Typically a set of of disks could be serviced by each page

cleaner if and when multiple page cleaners need to be used.

Databases might also use AIO for reads, for example, for prefetching data to service queries. This usually helps improve the performance of decision support workloads. The I/O pattern generated in these cases is that of streaming batches of large AIO reads, with sizes typically determined by the file allocation extent size used by the database (e.g., a DB2 installation might use a database extent size of 256KB). For installations using buffered AIO reads, tuning the readahead setting for the corresponding devices to be more than the extent size would help improve performance of streaming AIO reads (recall the discussion in Section 3.5).

9 Addressing AIO workloads involving both disk and communications I/O

Certain applications need to handle both disk-based AIO and communications I/O. For communications I/O, the epoll interface—which provides support for efficient scalable event polling in Linux 2.6—could be used as appropriate, possibly in conjunction with `O_NONBLOCK` socket I/O. Disk-based AIO on the other hand, uses the native AIO API `io_getevents` for completion notification. This makes it difficult to combine both types of I/O processing within a single event loop, even when such a model is a natural way to program the application, as in implementations of the application on other operating systems.

How do we address this issue? One option is to extend epoll to enable it to poll for notification of AIO completion events, so that AIO completion status can then be reaped in a non-blocking manner. This involves mixing both epoll and AIO API programming models, which is not ideal.

9.1 AIO poll interface

Another alternative is to add support for polling an event on a given file descriptor through the AIO interfaces. This function, referred to as AIO poll, can be issued through `io_submit()` just like other AIO operations, and specifies the file descriptor and the eventset to wait for. When the event occurs, notification is reported through `io_getevents()`.

The retry-based design of AIO poll works by converting the blocking wait for the event into a *retry exit*.

The generic synchronous polling code fits nicely into the AIO retry design, so most of the original polling code can be used unchanged. The private data area of the iocb can be used to hold polling-specific data structures, and a few special cases can be added to the generic polling entry points. This allows the AIO poll case to proceed without additional memory allocations.

9.2 AIO operations for communications I/O

A third option is to add support for AIO operations for communications I/O. For example, AIO support for pipes has been implemented by converting the blocking wait for I/O on pipes to a *retry exit*. The generic pipe code was also structured such that conversion to AIO retries was quite simple, the only significant change was using the current `io_wait` context instead of a locally defined waitqueue, and returning early if no data was available.

However, AIO pipe testing did show significantly more context switches then the 2.4 AIO pipe implementation, and this was coupled with much lower performance. The AIO core functions were relying on workqueues to do most of the retries, and this resulted in constant

switching between the workqueue threads and user processes.

The solution was to change the AIO core to do retries in `io_submit()` and in `io_getevents()`. This allowed the process to do some of its own work while it is scheduled in. Also, retries were switched to a delayed workqueue, so that bursts of retries would trigger fewer context switches.

While delayed wakeups helped with pipe workloads, it also caused I/O stalls in filesystem AIO workloads. This was because a delayed wakeup was being used even when a user process was waiting in `io_getevents()`. When user processes are actively waiting for events, it proved best to trigger the worker thread immediately.

General AIO support for network operations has been considered but not implemented so far because of lack of supporting study that predicts a significant benefit over what epoll and non-blocking I/O can provide, except for the scope for enabling potential zero-copy implementations. This is a potential area for future research.

10 Conclusions

Our experience over the last year with AIO development, stabilization and performance improvements brought us to design and implementation issues that went far beyond the initial concern of converting key I/O blocking points to be asynchronous.

AIO uncovered scenarios and I/O patterns that were unlikely or less significant with synchronous I/O alone. For example, the issues we discussed around streaming AIO performance with readahead and concurrent synchronized writes, as well as DIO vs buffered I/O complexities in the presence of AIO. In retrospect, this was the hardest part of supporting AIO—modifiying code that was originally designed only for synchronous I/O.

Interestingly, this also meant that AIO appeared to magnify some problems early. For example, issues with hashed waitqueues that led to the filtered wakeup patches, and readahead window collapses with large random reads which precipitated improvements to the readahead code from Ramachandra Pai. Ultimately, many of the core improvements that helped AIO have had positive benefits in allowing improved concurrency for some of the synchronous I/O paths.

In terms of benchmarking and optimizing Linux AIO performance, there is room for more exhaustive work. Requirements for AIO fsync support are currently under consideration. There is also a need for more widely used AIO applications, especially those that take advantaged of AIO support for buffered I/O or bring out additional requirements like network I/O beyond epoll or AIO poll. Finally, investigations into API changes to help enable more efficient POSIX AIO implementations based on kernel AIO support may be a worthwhile endeavor.

11 Acknowledgements

We would like to thank the many people on the `linux-aio@kvack.org` and `linux-kernel@vger.kernel.org` mailing lists who provided us with valuable comments and suggestions during our development efforts.

We would especially like to acknowledge the important contributions of Andrew Morton, Daniel McNeil, Badari Pulavarty, Stephen Tweedie, and William Lee Irwin towards several pieces of work discussed in this paper.

This paper and the work it describes wouldn't have been possible without the efforts of Janet Morgan in many different ways, starting from review, test and debugging feedback to joining the midnight oil camp to help with modifications and improvements to the text during the final stages of the paper.

We also thank Brian Twitchell, Steve Pratt, Gerrit Huizenga, Wayne Young, and John Lumby from IBM for their help and discussions along the way.

This work was a part of the Linux Scalability Effort (LSE) on SourceForge, and further information about Linux 2.6 AIO is available at the LSE AIO web page [10]. All the external AIO patches including AIO support for buffered filesystem I/O, AIO poll and AIO support for pipes are available at [9].

12 Legal Statement

13 Disclaimer

The benchmarks discussed in this paper were conducted for research purposes only, under laboratory conditions. Results will not be realized in all computing environments.

References

[1] Suparna Bhattacharya, Badari Pulavarthy, Steven Pratt, and Janet Morgan. Asynchronous i/o support for linux 2.5. In *Proceedings of the Linux Symposium*. Linux Symposium, Ottawa, July 2003. `http://archive.linuxsymposium.org/ols2003/Proceedings/All-Reprints/Reprint-Pulavarty-OLS2003.pdf`.

[2] Chris Mason. aio-stress microbenchmark. `ftp://ftp.suse.com/pub/people/mason/utils/aio-stress.c`.

[3] Stephen C. Tweedie. Posting on dio races in 2.4. `http://marc.theaimsgroup.com/?l=linux-fsdevel&m=105597840711609&w=2`.

[4] Andrew Morton. O_sync speedup patch. `http://www.kernel.org/pub/linux/kernel/people/akpm/patches/2.6/2.6.0/2.6.0-mm1/broken-out/O_SYNC-speedup-2.patch`.

[5] Daniel McNeil. Posting on synchronized writeback races. `http://marc.theaimsgroup.com/?l=linux-aio&m=107671729611002&w=2`.

[6] Andrew Morton. Posting on in-order tagged radix tree walk based vfs writeback. `http://marc.theaimsgroup.com/?l=bk-commits-head&m=108184544016117&w=2`.

[7] William Lee Irwin. Filtered wakeup patch. `http://marc.theaimsgroup.com/?l=bk-commits-head&m=108459430513660&w=2`.

[8] Transaction processing performance council. http://www.tpc.org.

[9] Suparna Bhattacharya (with contributions from Andrew Morton & Chris Mason). Additional 2.6 Linux Kernel Asynchronous I/O patches. http://www.kernel.org/pub/linux/kernel/people/suparna/aio.

[10] LSE team. Kernel Asynchronous I/O (AIO) Support for Linux. http://lse.sf.net/io/aio.html.

Methods to Improve Bootup Time in Linux

Tim R. Bird
Sony Electronics
tim.bird@am.sony.com

Abstract

This paper presents several techniques for reducing the bootup time of the Linux kernel, including Execute-In-Place (XIP), avoidance of `calibrate_delay()`, and reduced probing by certain drivers and subsystems. Using a variety of techniques, the Linux kernel can be booted on embedded hardware in under 500 milliseconds. Current efforts and future directions of work to improve bootup time are described.

1 Introduction

Users of consumer electronics products expect their devices to be available for use very soon after being turned on. Configurations of Linux for desktop and server markets exhibit boot times in the range of 20 seconds to a few minutes, which is unacceptable for many consumer products.

No single item is responsible for overall poor boot time performance. Therefore a number of techniques must be employed to reduce the boot up time of a Linux system. This paper presents several techniques which have been found to be useful for embedded configurations of Linux.

2 Overview of Boot Process

The entire boot process of Linux can be roughly divided into 3 main areas: firmware, kernel, and user space. The following is a list of events during a typical boot sequence:

1. power on
2. firmware (bootloader) starts
3. kernel decompression starts
4. kernel start
5. user space start
6. RC script start
7. application start
8. first available use

This paper focuses on techniques for reducing the bootup time up until the start of user space. That is, techniques are described which reduce the firmware time, and the kernel start time. This includes activities through the completion of event 4 in the list above.

The actual kernel execution begins with the routine `start_kernel()`, in the file `init/main.c`.

An overview of major steps in the initialization sequence of the kernel is as follows:

- `start_kernel()`
  - init architecture
  - init interrupts
  - init memory
  - start idle thread
  - call `rest_init()`
    - start 'init' kernel thread

The `init` kernel thread performs a few other tasks, then calls `do_basic_setup()`, which calls `do_initcalls()`, to run through the array of initialization routines for drivers statically linked in the kernel. Finally, this thread switches to user space by `execve`-ing to the first user space program, usually `/sbin/init`.

- `init` (kernel thread)
  - call `do_basic_setup()`
    - call `do_initcalls()`
      - init buses and drivers
  - prepare and mount root filesystem
  - call `run_init_process()`
    - call `execve()` to start user space process

3 Typical Desktop Boot Time

The boot times for a typical desktop system were measured and the results are presented below, to give an indication of the major areas in the kernel where time is spent. While the numbers in these tests differ somewhat from those for a typical embedded system, it is useful to see these to get an idea of where some of the trouble spots are for kernel booting.

3.1 System

An HP XW4100 Linux workstation system was used for these tests, with the following characteristics:

- Pentium 4 HT processor, running at 3GHz
- 512 MB RAM
- Western Digital 40G hard drive on hda
- Generic CDROM drive on hdc

3.2 Measurement method

The kernel used was 2.6.6, with the KFI patch applied. KFI stands for "Kernel Function Instrumentation". This is an in-kernel system to measure the duration of each function executed during a particular profiling run. It uses the `-finstrument-functions` option of `gcc` to instrument kernel functions with callouts on each function entry and exit. This code was authored by developers at MontaVista Software, and a patch for 2.6.6 is available, although the code is not ready (as of the time of this writing) for general publication. Information about KFI and the patch are available at:

`http://tree.celinuxforum.org/pubwiki/moin.cgi/KernelFunctionInstrumentation`

3.3 Key delays

The average time for kernel startup of the test system was about 7 seconds. This was the amount of time for just the kernel and NOT the firmware or user space. It corresponds to the period of time between events 4 and 5 in the boot sequence listed in Section 2.

Some key delays were found in the kernel startup on the test system. Table 1 shows some of the key routines where time was spent during bootup. These are the low-level routines where significant time was spent inside the functions themselves, rather than in subroutines called by the functions.

Kernel Function	No. of calls	Avg. call time	Total time
delay_tsc	5153	1	5537
default_idle	312	1	325
get_cmos_time	1	500	500
psmouse_sendbyte	44	2.4	109
pci_bios_find_device	25	1.7	44
atkbd_sendbyte	7	3.7	26
calibrate_delay	1	24	24

Note: Times are in milliseconds.

Table 1: Functions consuming lots of time during a typical desktop Linux kernel startup.

Note that over 80% of the total time of the bootup (almost 6 seconds out of 7) was spent busywaiting in `delay_tsc()` or spinning in the routine `default_idle()`. It appears that great reductions in total bootup time could be achieved if these delays could be reduced, or if it were possible to run some initialization tasks concurrently.

Another interesting point is that the routine `get_cmos_time()` was extremely variable in the length of time it took. Measurements of its duration ranged from under 100 milliseconds to almost one second. This routine, and methods to avoid this delay and variability, are discussed in section 9.

3.4 High-level delay areas

Since `delay_tsc()` is used (via various `delay` mechanisms) for busywaiting by a number of different subsystems, it is helpful to identify the higher-level routines which end up invoking this function.

Table 2 shows some high-level routines called during kernel initialization, and the amount of time they took to complete on the test machine. Duration times marked with a tilde denote functions which were highly variable in duration.

Kernel Function	Duration time
ide_init	3327
time_init	~500
isapnp_init	383
i8042_init	139
prepare_namespace	~50
calibrate_delay	24

Note: Times are in milliseconds.

Table 2: High-level delays during a typical startup.

For a few of these, it is interesting to examine the call sequences underneath the high-level routines. This shows the connection between the high-level routines that are taking a long time to complete and the functions where the time is actually being spent.

Figures 1 and 2 show some call sequences for high-level calls which take a long time to complete.

In each call tree, the number in parentheses is the number of times that the routine was called by the parent in this chain. Indentation shows the call nesting level.

For example, in Figure 1, `do_probe()` is called a total of 31 times by `probe_hwif()`, and it calls `ide_delay_50ms()` 78 times, and `try_to_identify()` 8 times.

The timing data for the test system showed that IDE initialization was a significant contributor to overall bootup time. The call sequence underneath `ide_init()` shows that a large number of calls are made to the routine `ide_delay_50ms()`, which in turn calls

```
ide_init->
 probe_for_hwifs(1)->
  ide_scan_pcibus(1)->
   ide_scan_pci_dev(2)->
    piix_init_one(2)->
     init_setup_piix(2)->
      ide_setup_pci_device(2)->
       probe_hwif_init(2)->
        probe_hwif(4)->
         do_probe(31)->
          ide_delay_50ms(78)->
           __const_udelay(3900)->
            __delay(3900)->
             delay_tsc(3900)
          try_to_identify(8)->
           actual_try_to_identify(8)->
            ide_delay_50ms(24)->
             __const_udelay(1200)->
              __delay(1200)->
               delay_tsc(1200)
```

Figure 1: IDE init call tree

```
isapnp_init->
 isapnp_isolate(1)->
  isapnp_isolate_rdp_select(1)->
   __const_udelay(25)->
    __delay(25)->
     delay_tsc(25)
   isapnp_key(18)->
    __const_udelay(18)->
     __delay(18)->
      delay_tsc(18)
```

Figure 2: ISAPnP init call tree

`__const_udelay()` very many times. The busywaits in `ide_delay_50ms()` alone accounted for over 5 seconds, or about 70% of the total boot up time.

Another significant area of delay was the initialization of the ISAPnP system. This took about 380 milliseconds on the test machine.

Both the mouse and the keyboard drivers used crude busywaits to wait for acknowledgements from their respective hardware.

Finally, the routine `calibrate_delay()` took about 25 milliseconds to run, to compute the value of `loops_per_jiffy` and print (the related) `BogoMips` for the machine.

The remaining sections of this paper discuss various specific methods for reducing bootup time for embedded and desktop systems. Some of these methods are directly related to some of the delay areas identified in this test configuration.

4 Kernel Execute-In-Place

A typical sequence of events during bootup is for the bootloader to load a compressed kernel image from either disk or Flash, placing it into RAM. The kernel is decompressed, either during or just after the copy operation. Then the kernel is executed by jumping to the function `start_kernel()`.

Kernel Execute-In-Place (XIP) is a mechanism where the kernel instructions are executed directly from ROM or Flash.

In a kernel XIP configuration, the step of copying the kernel code segment into RAM is omitted, as well as any decompression step. Instead, the kernel image is stored uncompressed in ROM or Flash. The kernel data segments still need to be initialized in RAM, but by eliminating the text segment copy and decompression, the overall effect is a reduction in the time required for the firmware phase of the bootup.

Table 3 shows the differences in time duration for various parts of the boot stage for a system booted with and without use of kernel XIP. The times in the table are shown in milliseconds. The table shows that using XIP in this configuration significantly reduced the time to copy the kernel to RAM (because only the data segments were copied), and completely eliminated the time to decompress the kernel (453 milliseconds). However, the kernel initialization time increased slightly in the XIP configuration, for a net savings of 463 milliseconds.

In order to support an Execute-In-Place con-

Boot Stage	Non-XIP time	XIP time
Copy kernel to RAM	85	12
Decompress kernel	453	0
Kernel initialization	819	882
Total kernel boot time	1357	894

Note: Times are in milliseconds. Results are for PowerPC 405 LP at 266 MHz

Table 3: Comparison of Non-XIP vs. XIP bootup times

figuration, the kernel must be compiled and linked so that the code is ready to be executed from a fixed memory location. There are examples of XIP configurations for ARM, MIPS and SH platforms in the CELinux source tree, available at: `http://tree.celinuxforum.org/`

4.1 XIP Design Tradeoffs

There are tradeoffs involved in the use of XIP. First, it is common for access times to flash memory to be greater than access times to RAM. Thus, a kernel executing from Flash usually runs a bit slower than a kernel executing from RAM. Table 4 shows some of the results from running the `lmbench` benchmark on system, with the kernel executing in a standard non-XIP configuration versus an XIP configuration.

Operation	Non-XIP	XIP
stat() syscall	22.4	25.6
fork a process	4718	7106
context switching for 16 processes and 64k data size	932	1109
pipe communication	248	548

Note: Times are in microseconds. Results are for lmbench benchmark run on OMAP 1510 (ARM9 at 168 MHz) processor

Table 4: Comparison of Non-XIP and XIP performance

Some of the operations in the benchmark took significantly longer with the kernel run in the XIP configuration. Most individual operations took about 20% to 30% longer. This performance penalty is suffered permanently while the kernel is running, and thus is a serious drawback to the use of XIP for reducing bootup time.

A second tradeoff with kernel XIP is between the sizes of various types of memory in the system. In the XIP configuration the kernel must be stored uncompressed, so the amount of Flash required for the kernel increases, and is usually about doubled, versus a compressed kernel image used with a non-XIP configuration. However, the amount of RAM required for the kernel is decreased, since the kernel code segment is never copied to RAM. Therefore, kernel XIP is also of interest for reducing the runtime RAM footprint for Linux in embedded systems.

There is additional research under way to investigate ways of reducing the performance impact of using XIP. One promising technique appears to be the use of "partial-XIP," where a highly active subset of the kernel is loaded into RAM, but the majority of the kernel is executed in place from Flash.

5 Delay Calibration Avoidance

One time-consuming operation inside the kernel is the process of calibrating the value used for delay loops. One of the first routines in the kernel, `calibrate_delay()`, executes a series of delays in order to determine the correct value for a variable called `loops_per_jiffy`, which is then subsequently used to execute short delays in the kernel.

The cost of performing this calibration is, interestingly, independent of processor speed. Rather, it is dependent on the number of iter-

ations required to perform the calibration, and the length of each iteration. Each iteration requires 1 jiffy, which is the length of time defined by the HZ variable.

In 2.4 versions of the Linux kernel, most platforms defined HZ as 100, which makes the length of a jiffy 10 milliseconds. A typical number of iterations for the calibration operation is 20 to 25, making the total time required for this operation about 250 milliseconds.

In 2.6 versions of the Linux kernel, a few platforms (notably i386) have changed HZ to 1000, making the length of a jiffy 1 millisecond. On those platforms, the typical cost of this calibration operation has decreased to about 25 milliseconds. Thus, the benefit of eliminating this operation on most standard desktop systems has been reduced. However, for many embedded systems, HZ is still defined as 100, which makes bypassing the calibration useful.

It is easy to eliminate the calibration operation. You can directly edit the code in `init/main.c:calibrate_delay()` to hardcode a value for `loops_per_jiffy`, and avoid the calibration entirely. Alternatively, there is a patch available at `http://tree.celinuxforum.org/pubwiki/moin.cgi/PresetLPJ`

This patch allows you to use a kernel configuration option to specify a value for `loops_per_jiffy` at kernel compile time. Alternatively, the patch also allows you to use a kernel command line argument to specify a preset value for `loops_per_jiffy` at kernel boot time.

6 Avoiding Probing During Bootup

Another technique for reducing bootup time is to avoid probing during bootup. As a general technique, this can consist of identifying hardware which is known not to be present on one's machine, and making sure the kernel is compiled without the drivers for that hardware.

In the specific case of IDE, the kernel supports options at the command line to allow the user to avoid performing probing for specific interfaces and devices. To do this, you can use the IDE and harddrive `noprobe` options at the kernel command line. Please see the file `Documentation/ide.txt` in the kernel source tree for details on the syntax of using these options.

On the test machine, IDE `noprobe` options were used to reduce the amount of probing during startup. The test machine had only a hard drive on hda (ide0 interface, first device) and a CD-ROM drive on hdc (ide1 interface, first device).

In one test, `noprobe` options were specified to suppress probing of non-used interfaces and devices. Specifically, the following arguments were added to the kernel command line:

```
hdb=none hdd=none ide2=noprobe
```

The kernel was booted and the result was that the function `ide_delay_50ms()` was called only 68 times, and `delay_tsc()` was called only 3453 times. During a regular kernel boot without these options specified, the function `ide_delay_50ms()` is called 102 times, and `delay_tsc()` is called 5153 times. Each call to `delay_tsc()` takes about 1 millisecond, so the total time savings from using these options was 1700 milliseconds.

These IDE `noprobe` options have been available at least since the 2.4 kernel series, and are an easy way to reduce bootup time, without even having to recompile the kernel.

7 Reducing Probing Delays

As was noted on the test machine, IDE initialization takes a significant percentage of the total bootup time. Almost all of this time is spent busywaiting in the routine `ide_delay_50ms()`.

It is trivial to modify the value of the timeout used in this routine. As an experiment, this code (located in the file `drivers/ide/ide.c`) was modified to only delay 5 milliseconds instead of 50 milliseconds.

The results of this change were interesting. When a kernel with this change was run on the test machine, the total time for the `ide_init()` routine dropped from 3327 milliseconds to 339 milliseconds. The total time spent in all invocations of `ide_delay_50ms()` was reduced from 5471 milliseconds to 552 milliseconds. The overall bootup time was reduced accordingly, by about 5 seconds.

The ide devices were successfully detected, and the devices operated without problem on the test machine. However, this configuration was not tested exhaustively.

Reducing the duration of the delay in the `ide_delay_50ms()` routine provides a substantial reduction in the overall bootup time for the kernel on a typical desktop system. It also has potential use in embedded systems where PCI-based IDE drives are used.

However, there are several issues with this modification that need to be resolved. This change may not support legacy hardware which requires long delays for proper probing and initializing. The kernel code needs to be analyzed to determine if any callers of this routine really need the 50 milliseconds of delay that they are requesting. Also, it should be determined whether this call is used only in initialization context or if it is used during regular runtime use of IDE devices also.

Also, it may be that 5 milliseconds does not represent the lowest possible value for this delay. It is possible that this value will need to be tuned to match the hardware for a particular machine. This type of tuning may be acceptable in the embedded space, where the hardware configuration of a product may be fixed. But it may be too risky to use in desktop configurations of Linux, where the hardware is not known ahead of time.

More experimentation, testing and validation are required before this technique should be used.

IMPORTANT NOTE: You should probably not experiment with this modification on production hardware unless you have evaluated the risks.

8 Using the “quiet” Option

One non-obvious method to reduce overhead during booting is to use the `quiet` option on the kernel command line. This option changes the loglevel to 4, which suppresses the output of regular (non-emergency) printk messages. Even though the messages are not printed to the system console, they are still placed in the kernel printk buffer, and can be retrieved after bootup using the `dmesg` command.

When embedded systems boot with a serial console, the speed of printing the characters to the console is constrained by the speed of the serial output. Also, depending on the driver, some VGA console operations (such as scrolling the screen) may be performed in software. For slow processors, this may take a significant amount of time. In either case, the cost of performing output of printk messages during bootup may be high. But it is easily eliminated using the `quiet` command line option.

Table 5 shows the difference in bootup time of using the `quiet` option and not, for two different systems (one with a serial console and one with a VGA console).

9 RTC Read Synchronization

One routine that potentially takes a long time during kernel startup is `get_cmos_time()`. This routine is used to read the value of the external real-time clock (RTC) when the kernel boots. Currently, this routine delays until the edge of the next second rollover, in order to ensure that the time value in the kernel is accurate with respect to the RTC.

However, this operation can take up to one full second to complete, and thus introduces up to 1 second of variability in the total bootup time. For systems where the target bootup time is under 1 second, this variability is unacceptable.

The synchronization in this routine is easy to remove. It can be eliminated by removing the first two loops in the function `get_cmos_time()`, which is located in `include/asm-i386/mach-default/mach_time.h` for the i386 architecture. Similar routines are present in the kernel source tree for other architectures.

When the synchronization is removed, the routine completes very quickly.

One tradeoff in making this modification is that the time stored by the Linux kernel is no longer completely synchronized (to the boundary of a second) with the time in the machine's realtime clock hardware. Some systems save the system time back out to the hardware clock on system shutdown. After numerous bootups and shutdowns, this lack of synchronization will cause the realtime clock value to drift from the correct time value.

Since the amount of un-synchronization is up to a second per boot cycle, this drift can be significant. However, for some embedded applications, this drift is unimportant. Also, in some situations the system time may be synchronized with an external source anyway, so the drift, if any, is corrected under normal circumstances soon after booting.

10 User space Work

There are a number of techniques currently available or under development for user space bootup time reductions. These techniques are (mostly) outside the scope of kernel development, but may provide additional benefits for reducing overall bootup time for Linux systems.

Some of these techniques are mentioned briefly in this section.

10.1 Application XIP

One technique for improving application startup speed is application XIP, which is similar to the kernel XIP discussed in this paper. To support application XIP the kernel must be compiled with a file system where files can be stored linearly (where the blocks for a file are stored contiguously) and uncompressed. One file system which supports this is CRAMFS, with the LINEAR option turned on. This is a read-only file system.

With application XIP, when a program is executed, the kernel program loader maps the text segments for applications directly from the flash memory of the file system. This saves the time required to load these segments into system RAM.

Platform	Speed	console type	w/o `quiet` option	with `quiet` option	difference
SH-4 SH7751R	240 MHz	VGA	637	461	176
OMAP 1510 (ARM 9)	168 MHz	serial	551	280	271

Note: Times are in milliseconds

Table 5: Bootup time with and without the `quiet` option

10.2 RC Script improvements

Also, there are a number of projects which strive to decrease total bootup time of a system by parallelizing the execution of the system run-command scripts ("RC scripts"). There is a list of resources for some of these projects at the following web site:

`http://tree.celinuxforum.org/pubwiki/moin.cgi/BootupTimeWorkingGroup`

Also, there has been some research conducted in reducing the overhead of running RC scripts. This consists of modifying the multi-function program `busybox` to reduce the number and cost of forks during RC script processing, and to optimize the usage of functions builtin to the `busybox` program. Initial testing has shown a reduction from about 8 seconds to 5 seconds for a particular set of Debian RC scripts on an OMAP 1510 (ARM 9) processor, running at 168 MHz.

11 Results

By use of the some of the techniques mentioned in this paper, as well as additional techniques, Sony was able to boot a 2.4.20-based Linux system, from power on to user space display of a greeting image and sound playback, in 1.2 seconds. The time from power on to the end of kernel initialization (first user space instruction) in this configuration was about 110 milliseconds. The processor was a TI OMAP 1510 processor, with an ARM9-based core, running at 168 MHz.

Some of the techniques used for reducing the bootup time of embedded systems can also be used for desktop or server systems. Often, it is possible, with rather simple and small modifications, to decrease the bootup time of the Linux kernel to only a few seconds. In the desktop configuration of Linux presented here, techniques from this paper were used to reduced the total bootup time from around 7 seconds to around 1 second. This was with no loss of functionality that the author could detect (with limited testing).

12 Further Research

As stated in the beginning of the paper, numerous techniques can be employed to reduce the overall bootup time of Linux systems. Further work continues or is needed in a number of areas.

12.1 Concurrent Driver Init

One area of additional research that seems promising is to structure driver initializations in the kernel so that they can proceed in parallel. For some items, like IDE initialization, there are large delays as buses and devices are probed and initialized. The time spent in such busywaits could potentially be used to perform other startup tasks, concurrently with the ini-

tializations waiting for hardware events to occur or time out.

The big problem to be addressed with concurrent initialization is to identify what kernel startup activities can be allowed to occur in parallel. The kernel init sequence is already a carefully ordered sequence of events to make sure that critical startup dependencies are observed. Any system of concurrent driver initialization will have to provide a mechanism to guarantee sequencing of initialization tasks which have order dependencies.

12.2 Partial XIP

Another possible area of further investigation, which has already been mentioned, is "partial XIP," whereby the kernel is executed *mostly* in-place. Prototype code already exists which demonstrates the mechanisms necessary to move a subset of an XIP-configured kernel into RAM, for faster code execution. The key to making partial kernel XIP useful will be to ensure correct identification (either statically or dynamically) of the sections of kernel code that need to be moved to RAM. Also, experimentation and testing need to be performed to determine the appropriate tradeoff between the size of the RAM-based portion of the kernel, and the effect on bootup time and system runtime performance.

12.3 Pre-linking and Lazy Linking

Finally, research is needed into reducing the time required to fixup links between programs and their shared libraries.

Two systems that have been proposed and experimented with are pre-linking and lazy linking. Pre-linking involves fixing the location in virtual memory of the shared libraries for a system, and performing fixups on the programs of the system ahead of time. Lazy linking consists of only performing fixups on demand as library routines are called by a running program.

Additional research is needed with both of these techniques to determine if they can provide benefit for current Linux systems.

13 Credits

This paper is the result of work performed by the Bootup Time Working Group of the CE Linux forum (of which the author is Chair). I would like to thank developers at some of CELF's member companies, including Hitachi, Intel, Mitsubishi, MontaVista, Panasonic, and Sony, who contributed information or code used in writing this paper.

Linux on NUMA Systems

Martin J. Bligh
mbligh@aracnet.com
Matt Dobson
colpatch@us.ibm.com
Darren Hart
dvhltc@us.ibm.com
Gerrit Huizenga
gh@us.ibm.com

Abstract

NUMA is becoming more widespread in the marketplace, used on many systems, small or large, particularly with the advent of AMD Opteron systems. This paper will cover a summary of the current state of NUMA, and future developments, encompassing the VM subsystem, scheduler, topology (CPU, memory, I/O layouts including complex non-uniform layouts), userspace interface APIs, and network and disk I/O locality. It will take a broad-based approach, focusing on the challenges of creating subsystems that work for all machines (including AMD64, PPC64, IA-32, IA-64, etc.), rather than just one architecture.

1 What is a NUMA machine?

NUMA stands for non-uniform memory architecture. Typically this means that not all memory is the same "distance" from each CPU in the system, but also applies to other features such as I/O buses. The word "distance" in this context is generally used to refer to both latency and bandwidth. Typically, NUMA machines can access any resource in the system, just at different speeds.

NUMA systems are sometimes measured with a simple "NUMA factor" ratio of N:1—meaning that the latency for a cache miss memory read from remote memory is *N* times the latency for that from local memory (for NUMA machines, $N > 1$). Whilst such a simple descriptor is attractive, it can also be highly misleading, as it describes latency only, not bandwidth, on an uncontended bus (which is not particularly relevant or interesting), and takes no account of inter-node caches.

The term *node* is normally used to describe a grouping of resources—e.g., CPUs, memory, and I/O. On some systems, a node may contain only some types of resources (e.g., only memory, or only CPUs, or only I/O); on others it may contain all of them. The interconnect between nodes may take many different forms, but can be expected to be higher latency than the connection within a node, and typically lower bandwidth.

Programming for NUMA machines generally implies focusing on *locality*—the use of resources close to the device in question, and trying to reduce traffic between nodes; this type of programming generally results in better application throughput. On some machines with high-speed cross-node interconnects, bet-

ter performance may be derived under certain workloads by "striping" accesses across multiple nodes, rather than just using local resources, in order to increase bandwidth. Whilst it is easy to demonstrate a benchmark that shows improvement via this method, it is difficult to be sure that the concept is generally benefical (i.e., with the machine under full load).

2 Why use a NUMA architecture to build a machine?

The intuitive approach to building a large machine, with many processors and banks of memory, would be simply to scale up the typical 2–4 processor machine with all resources attached to a shared system bus. However, restrictions of electronics and physics dictate that accesses slow as the length of the bus grows, and the bus is shared amongst more devices.

Rather than accept this global slowdown for a larger machine, designers have chosen to instead give fast access to a limited set of local resources, and reserve the slower access times for remote resources.

Historically, NUMA architectures have only been used for larger machines (more than 4 CPUs), but the advantages of NUMA have been brought into the commodity marketplace with the advent of AMD's x86-64, which has one CPU per node, and local memory for each processor. Linux supports NUMA machines of every size from 2 CPUs upwards (e.g., SGI have machines with 512 processors).

It might help to envision the machine as a group of standard SMP machines, connected by a very fast interconnect somewhat like a network connection, except that the transfers over that bus are transparent to the operating system. Indeed, some earlier systems were built exactly like that; the older Sequent NUMA-Q hardware uses a standard 450NX 4 processor chipset, with an SCI interconnect plugged into the system bus of each node to unify them, and pass traffic between them. The complex part of the implementation is to ensure cache-coherency across the interconnect, and such machines are often referred to as *CC-NUMA* (cache coherent NUMA). As accesses over the interconnect are transparent, it is possible to program such machines as if they were standard SMP machines (though the performance will be poor). Indeed, this is exactly how the NUMA-Q machines were first bootstrapped.

Often, we are asked why people do not use clusters of smaller machines, instead of a large NUMA machine, as clusters are cheaper, simpler, and have a better price:performance ratio. Unfortunately, it makes the programming of applications much harder; all of the inter-communication and load balancing now has to be more explicit. Some large applications (e.g., database servers) do not split up across multiple cluster nodes easily—in those situations, people often use NUMA machines. In addition, the interconnect for NUMA boxes is normally very low latency, and very high bandwidth, yielding excellent performance. The management of a single NUMA machine is also simpler than that of a whole cluster with multiple copies of the OS.

We could either have the operating system make decisions about how to deal with the architecture of the machine on behalf of the user processes, or give the userspace application an API to specify how such decisions are to be made. It might seem, at first, that the userspace application is in a better position to make such decisions, but this has two major disadvantages:

1. Every application must be changed to support NUMA machines, and may need to

be revised when a new hardware platform is released.

2. Applications are not in a good position to make global holistic decisions about machine resources, coordinate themselves with other applications, and balance decisions between them.

Thus decisions on process, memory and I/O placement are normally best left to the operating system, perhaps with some hints from userspace about which applications group together, or will use particular resources heavily. Details of hardware layout are put in one place, in the operating system, and tuning and modification of the necessary algorithms are done once in that central location, instead of in every application. In some circumstances, the application or system administrator will want to override these decisions with explicit APIs, but this should be the exception, rather than the norm.

3 Linux NUMA Memory Support

In order to manage memory, Linux requires a page descriptor structure (`struct page`) for each physical page of memory present in the system. This consumes approximately 1% of the memory managed (assuming 4K page size), and the structures are grouped into an array called `mem_map`. For NUMA machines, there is a separate array for each node, called `lmem_map`. The `mem_map` and `lmem_map` arrays are simple contiguous data structures accessed in a linear fashion by their offset from the beginning of the node. This means that the memory controlled by them is assumed to be physically contiguous.

NUMA memory support is enabled by `CONFIG_DISCONTIGMEM` and `CONFIG_NUMA`. A node descriptor called a `struct pgdata_t` is created for each node. Currently we do not support discontiguous memory within a node (though large gaps in the physical address space are acceptable between nodes). Thus we must still create page descriptor structures for "holes" in memory within a node (and then mark them invalid), which will waste memory (potentially a problem for large holes).

Dave McCracken has picked up Daniel Phillips' earlier work on a better data structure for holding the page descriptors, called `CONFIG_NONLINEAR`. This will allow the mapping of discontigous memory ranges inside each node, and greatly simplify the existing code for discontiguous memory on non-NUMA machines.

`CONFIG_NONLINEAR` solves the problem by creating an artificial layer of linear addresses. It does this by dividing the physical address space into fixed size sections (akin to very large pages), then allocating an array to allow translations from linear physical address to true physical address. This added level of indirection allows memory with widely differing true physical addresses to appear adjacent to the page allocator and to be in the same zone, with a single struct page array to describe them. It also provides support for memory hotplug by allowing new physical memory to be added to an existing zone and struct page array.

Linux normally allocates memory for a process on the local node, i.e., the node that the process is currently running on. `alloc_pages` will call `alloc_pages_node` for the current processor's node, which will pass the relevant zonelist (`pgdat->node_zonelists`) to the core allocator (`__alloc_pages`). The zonelists are built by `build_zonelists`, and are set up to allocate memory in a round-robin fashion, starting from the local node (this creates a roughly even distribution of memory

pressure).

In the interest of reducing cross-node traffic, and reducing memory access latency for frequently accessed data and text, it is desirable to replicate any such memory that is read-only to each node, and use the local copy on any accesses, rather than a remote copy. The obvious candidates for such replication are the kernel text itself, and the text of shared libraries such as libc. Of course, this faster access comes at the price of increased memory usage, but this is rarely a problem on large NUMA machines. Whilst it might be technically possible to replicate read/write mappings, this is complex, of dubious utility, and is unlikely to be implemented.

Kernel text is assumed by the kernel itself to appear at a fixed virtual address, and to change this would be problematic. Hence the easiest way to replicate it is to change the virtual to physical mappings for each node to point at a different address. On IA-64, this is easy, since the CPU provides hardware assistance in the form of a pinned TLB entry.

On other architectures this proves more difficult, and would depend on the structure of the pagetables. On IA-32 with PAE enabled, as long as the user-kernel split is aligned on a PMD boundary, we can have a separate kernel PMD for each node, and point the vmalloc area (which uses small page mappings) back to a globally shared set of PTE pages. The PMD entries for the `ZONE_NORMAL` areas normally never change, so this is not an issue, though there is an issue with `ioremap_nocache` that can change them (GART trips over this) and speculative execution means that we will have to deal with that (this can be a slow-path that updates all copies of the PMDs though).

Dave Hansen has created a patch to replicate read only pagecache data, by adding a per-node data structure to each node of the pagecache radix tree. As soon as any mapping is opened for write, the replication is collapsed, making it safe. The patch gives a 5%–40% increase in performance, depending on the workload.

In the 2.6 Linux kernel, we have a per-node LRU for page management and a per-node LRU lock, in place of the global structures and locks of 2.4. Not only does this reduce contention through finer grained locking, it also means we do not have to search other nodes' page lists to free up pages on one node which is under memory pressure. Moreover, we get much better locality, as only the local kswapd process is accessing that node's pages. Before splitting the LRU into per-node lists, we were spending 50% of the system time during a kernel compile just spinning waiting for `pagemap_lru_lock` (which was the biggest global VM lock at the time). Contention for the `pagemap_lru_lock` is now so small it is not measurable.

4 Sched Domains—a Topology-aware Scheduler

The previous Linux scheduler, the O(1) scheduler, provided some needed improvements to the 2.4 scheduler, but shows its age as more complex system topologies become more and more common. With technologies such as NUMA, Symmetric Multi-Threading (SMT), and variations and combinations of these, the need for a more flexible mechanism to model system topology is evident.

4.1 Overview

In answer to this concern, the mainline 2.6 tree (linux-2.6.7-rc1 at the time of this writing) contains an updated scheduler with support for generic CPU topologies with a data structure, `struct sched_domain`, that models the architecture and defines scheduling policies.

Simply speaking, sched domains group CPUs together in a hierarchy that mimics that of the physical hardware. Since CPUs at the bottom of the hierarchy are most closely related (in terms of memory access), the new scheduler performs load balancing most often at the lower domains, with decreasing frequency at each higher level.

Consider the case of a machine with two SMT CPUs. Each CPU contains a pair of virtual CPU siblings which share a cache and the core processor. The machine itself has two physical CPUs which share main memory. In such a situation, treating each of the four effective CPUs the same would not result in the best possible performance. With only two tasks, for example, the scheduler should place one on CPU0 and one on CPU2, and not on the two virtual CPUs of the same physical CPU. When running several tasks it seems natural to try to place newly ready tasks on the CPU they last ran on (hoping to take advantage of cache warmth). However, virtual CPU siblings share a cache; a task that was running on CPU0, then blocked, and became ready when CPU0 was running another task and CPU1 was idle, would ideally be placed on CPU1. Sched domains provide the structures needed to realize these sorts of policies. With sched domains, each physical CPU represents a domain containing the pair of virtual siblings, each represented in a `sched_group` structure. These two domains both point to a parent domain which contains all four effective processors in two `sched_group` structures, each containing a pair of virtual siblings. Figure 1 illustrates this hierarchy.

Next consider a two-node NUMA machine with two processors per node. In this example there are no virtual sibling CPUs, and therefore no shared caches. When a task becomes ready and the processor it last ran on is busy, the scheduler needs to consider waiting until that CPU is available to take advantage of cache warmth. If the only available CPU is on another node, the scheduler must carefully weigh the costs of migrating that task to another node, where access to its memory will be slower. The lowest level sched domains in a machine like this will contain the two processors of each node. These two CPU level domains each point to a parent domain which contains the two nodes. Figure 2 illustrates this hierarchy.

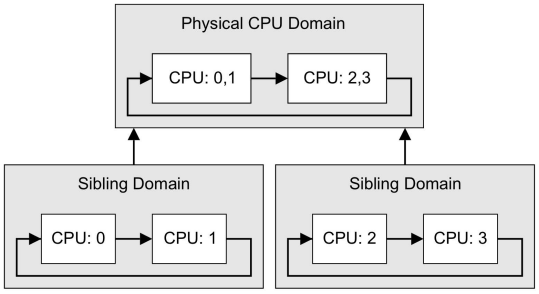


Figure 1: SMT Domains

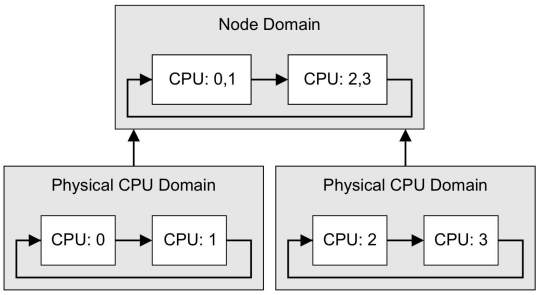


Figure 2: NUMA Domains

The next logical step is to consider an SMT NUMA machine. By combining the previous two examples, the resulting sched domain hierarchy has three levels, sibling domains, physical CPU domains, and the node domain. Figure 3 illustrates this hierarchy.

The unique AMD Opteron architecture warrants mentioning here as it creates a NUMA system on a single physical board. In this case, however, each NUMA node contains only one

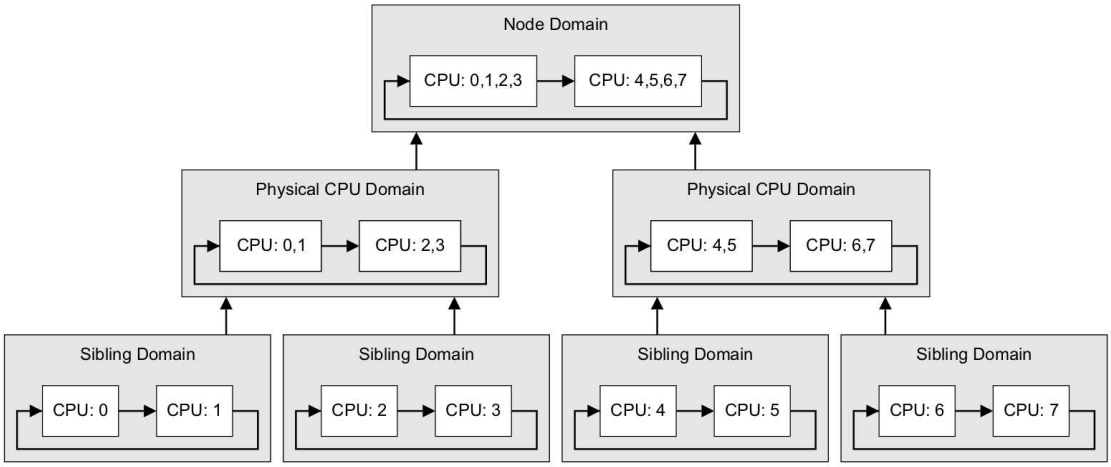


Figure 3: SMT NUMA Domains

physical CPU. Without careful consideration of this property, a typical NUMA sched domains hierarchy would perform badly, trying to load balance single CPU nodes often (an obvious waste of cycles) and between node domains only rarely (also bad since these actually represent the physical CPUs).

4.2 Sched Domains Implementation

4.2.1 Structure

The `sched_domain` structure stores policy parameters and flags and, along with the `sched_group` structure, is the primary building block in the domain hierarchy. Figure 4 describes these structures. The `sched_domain` structure is constructed into an upwardly traversable tree via the parent pointer, the top level domain setting parent to NULL. The groups list is a circular list of of `sched_group` structures which essentially define the CPUs in each child domain and the relative power of that group of CPUs (two physical CPUs are more powerful than one SMT CPU). The span member is simply a bit vector with a 1 for every CPU encompassed by that domain and is always the union of the bit vector stored in each element of the groups list. The remaining fields define the scheduling policy to be followed while dealing with that domain, see Section 4.2.2.

While the hierarchy may seem simple, the details of its construction and resulting tree structures are not. For performance reasons, the domain hierarchy is built on a per-CPU basis, meaning each CPU has a unique instance of each domain in the path from the base domain to the highest level domain. These duplicate structures do share the `sched_group` structures however. The resulting tree is difficult to diagram, but resembles Figure 5 for the machine with two SMT CPUs discussed earlier.

In accordance with common practice, each architecture may specify the construction of the sched domains hierarchy and the parameters and flags defining the various policies. At the time of this writing, only i386 and ppc64 defined custom construction routines. Both architectures provide for SMT processors and NUMA configurations. Without an architecture-specific routine, the kernel uses the default implementations in `sched.c`, which do take NUMA into account.

```
struct sched_domain {
        /* These fields must be setup */
        struct sched_domain *parent;            /* top domain must be null terminated */
        struct sched_group *groups;             /* the balancing groups of the domain */
        cpumask_t span;                         /* span of all CPUs in this domain */
        unsigned long min_interval;             /* Minimum balance interval ms */
        unsigned long max_interval;             /* Maximum balance interval ms */
        unsigned int busy_factor;               /* less balancing by factor if busy */
        unsigned int imbalance_pct;             /* No balance until over watermark */
        unsigned long long cache_hot_time;      /* Task considered cache hot (ns) */
        unsigned int cache_nice_tries;          /* Leave cache hot tasks for # tries */
        unsigned int per_cpu_gain;              /* CPU % gained by adding domain cpus */
        int flags;                              /* See SD_* */

        /* Runtime fields. */
        unsigned long last_balance;             /* init to jiffies. units in jiffies */
        unsigned int balance_interval;          /* initialise to 1. units in ms. */
        unsigned int nr_balance_failed;         /* initialise to 0 */
};

struct sched_group {
        struct sched_group *next;               /* Must be a circular list */
        cpumask_t cpumask;
        unsigned long cpu_power;
};
```

Figure 4: Sched Domains Structures

4.2.2 Policy

The new scheduler attempts to keep the system load as balanced as possible by running rebalance code when tasks change state or make specific system calls, we will call this *event balancing*, and at specified intervals measured in jiffies, called *active balancing*. Tasks must do something for event balancing to take place, while active balancing occurs independent of any task.

Event balance policy is defined in each `sched_domain` structure by setting a combination of the #defines of figure 6 in the flags member.

To define the policy outlined for the dual SMT processor machine in Section 4.1, the lowest level domains would set `SD_BALANCE_NEWIDLE` and `SD_WAKE_IDLE` (as there is no cache penalty for running on a different sibling within the same physical CPU), `SD_SHARE_CPUPOWER` to indicate to the scheduler that this is an SMT processor (the scheduler will give full physical CPU access to a high priority task by idling the virtual sibling CPU), and a few common flags `SD_BALANCE_EXEC`, `SD_BALANCE_CLONE`, and `SD_WAKE_AFFINE`. The next level domain represents the physical CPUs and will not set `SD_WAKE_IDLE` since cache warmth is a concern when balancing across physical CPUs, nor `SD_SHARE_CPUPOWER`. This domain adds the `SD_WAKE_BALANCE` flag to compensate for the removal of `SD_WAKE_IDLE`. As discussed earlier, an SMT NUMA system will have these two domains and another node-level domain. This domain removes the `SD_BALANCE_NEWIDLE` and `SD_WAKE_AFFINE` flags, resulting in far fewer balancing across nodes than within nodes. When one of these events occurs, the scheduler search up the domain hierarchy and performs the load balancing at the highest level domain with the corresponding flag set.

Active balancing is fairly straightforward and aids in preventing CPU-hungry tasks from hogging a processor, since these tasks may only

```
#define SD_BALANCE_NEWIDLE    1   /* Balance when about to become idle */
#define SD_BALANCE_EXEC       2   /* Balance on exec */
#define SD_BALANCE_CLONE      4   /* Balance on clone */
#define SD_WAKE_IDLE          8   /* Wake to idle CPU on task wakeup */
#define SD_WAKE_AFFINE       16   /* Wake task to waking CPU */
#define SD_WAKE_BALANCE      32   /* Perform balancing at task wakeup */
#define SD_SHARE_CPUPOWER    64   /* Domain members share cpu power */
```

Figure 6: Sched Domains Policies

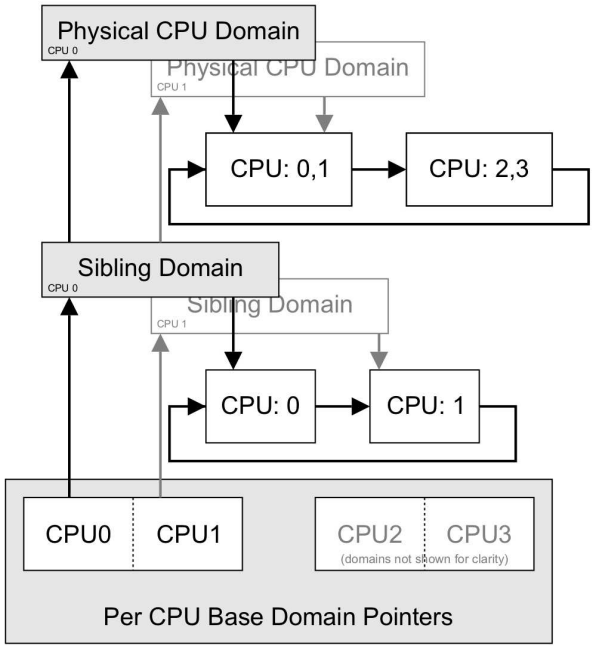


Figure 5: Per CPU Domains

rarely trigger event balancing. At each rebalance tick, the scheduler starts at the lowest level domain and works its way up, checking the `balance_interval` and `last_balance` fields to determine if that domain should be balanced. If the domain is already busy, the `balance_interval` is adjusted using the `busy_factor` field. Other fields define how out of balance a node must be before rebalancing can occur, as well as some sane limits on cache hot time and min and max balancing intervals. As with the flags for event balancing, the active balancing parameters are defined to perform less balancing at higher domains in the hierarchy.

4.3 Conclusions and Future Work

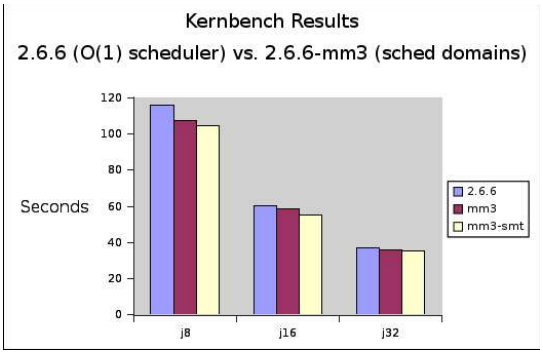


Figure 7: Kernbench Results

To compare the O(1) scheduler of mainline with the sched domains implementation in the mm tree, we ran kernbench (with the `-j` option to make set to 8, 16, and 32) on a 16 CPU SMT machine (32 virtual CPUs) on linux-2.6.6 and linux-2.6.6-mm3 (the latest tree with sched domains at the time of the benchmark) with and without `CONFIG_SCHED_SMT` enabled. The results are displayed in Figure 7. The O(1) scheduler evenly distributed compile tasks across virtual CPUs, forcing tasks to share cache and computational units between virtual sibling CPUs. The sched domains implementation with `CONFIG_SCHED_SMT` enabled balanced the load accross physical CPUs, making far better use of CPU resources when running fewer tasks than CPUs (as in the j8 case) since each compile task would have exclusive access to the physical CPU. Surprisingly, sched domains (which would seem to have more overhead than the mainline scheduler) even showed improvement for the j32 case, where it doesn't

benefit from balancing across physical CPUs before virtual CPUs as there are more tasks than virtual CPUs. Considering the sched domains implementation has not been heavily tested or tweaked for performance, some fine tuning is sure to further improve performance.

The sched domains structures replace the expanding set of `#ifdefs` of the O(1) scheduler, which should improve readability and maintainability. Unfortunately, the per CPU nature of the domain construction results in a non-intuitive structure that is difficult to work with. For example, it is natural to discuss the policy defined at "the" top level domain; unfortunately there are `NR_CPUS` top level domains and, since they are self-adjusting, each one could conceivably have a different set of flags and parameters. Depending on which CPU the scheduler was running on, it could behave radically differently. As an extension of this research, an effort to analyze the impact of a unified sched domains hierarchy is needed, one which only creates one instance of each domain.

Sched domains provides a needed structural change to the way the Linux scheduler views modern architectures, and provides the parameters needed to create complex scheduling policies that cater to the strengths and weaknesses of these systems. Currently only i386 and ppc64 machines benefit from arch specific construction routines; others must now step forward and fill in the construction and parameter setting routines for their architecture of choice. There is still plenty of fine tuning and performance tweaking to be done.

5 NUMA API

5.1 Introduction

One of the biggest impediments to the acceptance of a NUMA API for Linux was a lack of understanding of what its potential uses and users would be. There are two schools of thought when it comes to writing NUMA code. One says that the OS should take care of all the NUMA details, hide the NUMAness of the underlying hardware in the kernel and allow userspace applications to pretend that it's a regular SMP machine. Linux does this by having a process scheduler and a VMM that make intelligent decisions based on the hardware topology presented by arch-specific code. The other way to handle NUMA programming is to provide as much detail as possible about the system to userspace and allow applications to exploit the hardware to the fullest by giving scheduling hints, memory placement directives, etc., and the NUMA API for Linux handles this. Many applications, particularly larger applications with many concurrent threads of execution, cannot fully utilize a NUMA machine with the default scheduler and VM behavior. Take, for example, a database application that uses a large region of shared memory and many threads. This application may have a startup thread that initializes the environment, sets up the shared memory region, and forks off the worker threads. The default behavior of Linux's VM for NUMA is to bring pages into memory on the node that faulted them in. This behavior for our hypothetical app would mean that many pages would get faulted in by the startup thread on the node it is executing on, not necessarily on the node containing the processes that will actually use these pages. Also, the forked worker threads would get spread around by the scheduler to be balanced across all the nodes and their CPUs, but with no guarantees as to which

threads would be associated with which nodes. The NUMA API and scheduler affinity syscalls allow this application to specify that its threads be pinned to particular CPUs and that its memory be placed on particular nodes. The application knows which threads will be working with which regions of memory, and is better equipped than the kernel to make those decisions.

The Linux NUMA API allows applications to give regions of their own virtual memory space specific allocation behaviors, called policies. Currently there are four supported policies: PREFERRED, BIND, INTERLEAVE, and DEFAULT. The DEFAULT policy is the simplest, and tells the VMM to do what it would normally do (ie: pre-NUMA API) for pages in the policied region, and fault them in from the local node. This policy applies to all regions, but is overridden if an application requests a different policy. The PREFERRED policy allows an application to specify one node that all pages in the policied region should come from. However, if the specified node has no available pages, the PREFERRED policy allows allocation to fall back to any other node in the system. The BIND policy allows applications to pass in a nodemask, a bitmap of nodes, that the VM is required to use when faulting in pages from a region. The fourth policy type, INTERLEAVE, again requires applications to pass in a nodemask, but with the INTERLEAVE policy, the nodemask is used to ensure pages are faulted in in a round-robin fashion from the nodes in the nodemask. As with the PREFERRED policy, the INTERLEAVE policy allows page allocation to fall back to other nodes if necessary. In addition to allowing a process to policy a specific region of its VM space, the NUMA API also allows a process to policy its entire VM space with a process-wide policy, which is set with a different syscall: `set_mempolicy()`. Note that process-wide policies are not persistent over swapping, however per-VMA policies are. Please also note that none of the policies will migrate existing (already allocated) pages to match the binding.

The actual implementation of the in-kernel policies uses a `struct mempolicy` that is hung off the `struct vm_area_struct`. This choice involves some tradeoffs. The first is that, previous to the NUMA API, the per-VMA structure was exactly 32 bytes on 32-bit architectures, meaning that multiple `vm_area_structs` would fit conveniently in a single cacheline. The structure is now a little larger, but this allowed us to achieve a per-VMA granularity to policied regions. This is important in that it is flexible enough to bind a single page, a whole library, or a whole process' memory. This choice did lead to a second obstacle, however, which was for shared memory regions. For shared memory regions, we really want the policy to be shared amongst all processes sharing the memory, but VMAs are not shared across separate tasks. The solution that was implemented to work around this was to create a red-black tree of "shared policy nodes" for shared memory regions. Due to this, calls were added to the `vm_ops` structure which allow the kernel to check if a shared region has any policies and to easily retrieve these shared policies.

5.2 Syscall Entry Points

1. sys_mbind(unsigned long start, unsigned long len, unsigned long mode, unsigned long *nmask, unsigned long maxnode, unsigned flags);

   Bind the region of memory `[start, start+len)` according to `mode` and `flags` on the nodes enumerated in `nmask` and having a maximum possible node number of `maxnode`.

2. sys_set_mempolicy(int mode, unsigned

long *nmask, unsigned long maxnode);

Bind the entire address space of the current process according to `mode` on the nodes enumerated in `nmask` and having a maximum possible node number of `maxnode`.

3. sys_get_mempolicy(int *policy, unsigned long *nmask, unsigned long maxnode, unsigned long addr, unsigned long flags);

   Return the current binding's mode in `policy` and node enumeration in `nmask` based on the `maxnode`, `addr`, and `flags` passed in.

In addition to the raw syscalls discussed above, there is a user-level library called "libnuma" that attempts to present a more cohesive interface to the NUMA API, topology, and scheduler affinity functionality. This, however, is documented elsewhere.

5.3 At `mbind()` Time

After argument validation, the passed-in list of nodes is checked to make sure they are all online. If the node list is ok, a new memory policy structure is allocated and populated with the binding details. Next, the given address range is checked to make sure the vma's for the region are present and correct. If the region is ok, we proceed to actually install the new policy into all the vma's in that range. For most types of virtual memory regions, this involves simply pointing the `vma->vm_policy` to the newly allocated memory policy structure. For shared memory, hugetlbfs, and tmpfs, however, it's not quite this simple. In the case of a memory policy for a shared segment, a red-black tree root node is created, if it doesn't already exist, to represent the shared memory segment and is populated with "shared policy nodes." This allows a user to bind a single shared memory segment with multiple different bindings.

5.4 At Page Fault Time

There are now several new and different flavors of `alloc_pages()` style functions. Previous to the NUMA API, there existed `alloc_page()`, `alloc_pages()` and `alloc_pages_node()`. Without going into too much detail, `alloc_page()` and `alloc_pages()` both called `alloc_pages_node()` with the current node id as an argument. `alloc_pages_node()` allocated 2^{order} pages from a specific node, and was the only caller to the *real* page allocator, `__alloc_pages()`.

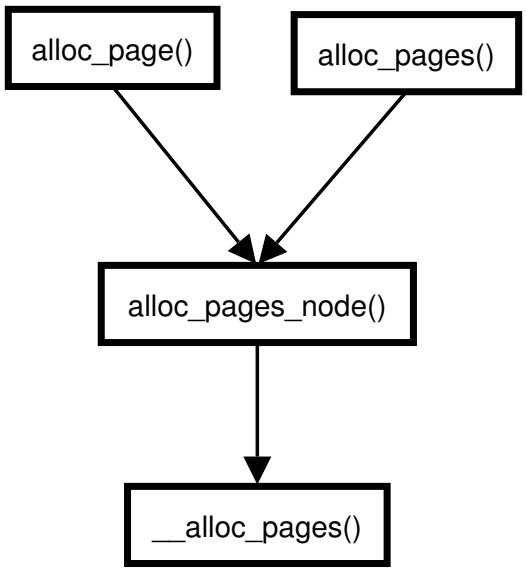


Figure 8: old `alloc_pages`

With the introduction of the NUMA API, non-NUMA kernels still retain the old `alloc_page*()` routines, but the NUMA allocators have changed. `alloc_pages_node()` and `__alloc_pages()`, the core routines remain untouched, but all calls to `alloc_page()`/`alloc_pages()` now end up calling `alloc_pages_current()`, a new function.

There has also been the addition of two new page allocation functions: `alloc_page_vma()` and `alloc_page_interleave()`. `alloc_pages_current()` checks that the system is not currently `in_interrupt()`, and if it isn't, uses the current process's process policy for allocation. If the system is currently in interrupt context, `alloc_pages_current()` falls back to the old default allocation scheme. `alloc_page_interleave()` allocates pages from regions that are bound with an interleave policy, and is broken out separately because there are some statistics kept for interleaved regions. `alloc_page_vma()` is a new allocator that allocates only single pages based on a per-vma policy. The `alloc_page_vma()` function is the only one of the new allocator functions that must be called explicity, so you will notice that some calls to `alloc_pages()` have been replaced by calls to `alloc_page_vma()` throughout the kernel, as necessary.

5.5 Problems/Future Work

There is no checking that the nodes requested are online at page fault time, so interactions with hotpluggable CPUs/memory will be tricky. There is an asymmetry between how you bind a memory region and a whole process's memory: One call takes a flags argument, and one doesn't. Also the `maxnode` argument is a bit strange, the get/set_affinity calls take a number of bytes to be read/written instead of a maximum CPU number. The `alloc_page_interleave()` function could be dropped if we were willing to forgo the statistics that are kept for interleaved regions. Again, a lack of symmetry exists because other types of policies aren't tracked in any way.

6 Legal statement

References

[LWN] LWN Editor, "Scheduling Domains," `http://lwn.net/Articles/80911/`

[MM2] Linux 2.6.6-rc2/mm2 source, `http://www.kernel.org`

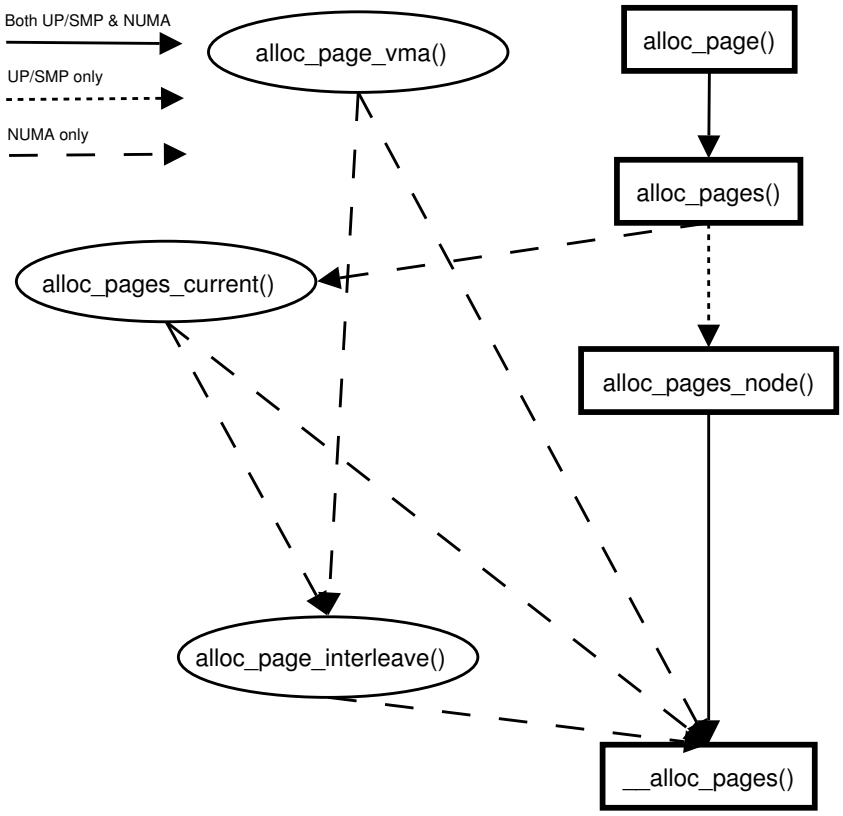


Figure 9: new `alloc_pages`

Improving Kernel Performance by Unmapping the Page Cache

James Bottomley
SteelEye Technology, Inc.
James.Bottomley@SteelEye.com

Abstract

The current DMA API is written on the founding assumption that the coherency is being done between the device and kernel virtual addresses. We have a different API for coherency between the kernel and userspace. The upshot is that every Process I/O must be flushed twice: Once to make the user coherent with the kernel and once to make the kernel coherent with the device. Additionally, having to map all pages for I/O places considerable resource pressure on x86 (where any highmem page must be separately mapped).

We present a different paradigm: Assume that by and large, read/write data is only required by a single entity (the major consumers of large multiply shared mappings are libraries, which are read only) and optimise the I/O path for this case. This means that any other shared consumers of the data (including the kernel) must separately map it themselves. The DMA API would be changed to perform coherence to the preferred address space (which could be the kernel). This is a slight paradigm shift, because now devices that need to peek at the data may have to map it first. Further, to free up more space for this mapping, we would break the assumption that any page in ZONE_NORMAL is automatically mapped into kernel space.

The benefits are that I/O goes straight from the device into the user space (for processors that have virtually indexed caches) and the kernel has quite a large unmapped area for use in kmapping highmem pages (for x86).

1 Introduction

In the Linux kernel[1] there are two addressing spaces: memory physical which is the location in the actual memory subsystem and CPU virtual, which is an address the CPU's Memory Management Unit (MMU) translates to a memory physical address internally. The Linux kernel operates completely in CPU virtual space, keeping separate virtual spaces for the kernel and each of the current user processes. However, the kernel also has to manage the mappings between physical and virtual spaces, and to do that it keeps track of where the physical pages of memory currently are.

In the Linux kernel, memory is split into zones in memory physical space:

- `ZONE_DMA`: A historical region where ISA DMAable memory is allocated from. On x86 this is all memory under 16MB.
- `ZONE_NORMAL`: This is where normally allocated kernel memory goes. Where

[1] This is not quite true, there are kernels for processors without memory management units, but these are very specialised and won't be considered further

this zone ends depends on the architecture. However, all memory in this zone is mapped in kernel space (visible to the kernel).

- `ZONE_HIGHMEM`: This is where the rest of the memory goes. Its characteristic is that it is not mapped in kernel space (thus the kernel cannot access it without first mapping it).

1.1 The x86 and Highmem

The main reason for the existence of `ZONE_HIGHMEM` is a peculiar quirk on the x86 processor which makes it rather expensive to have different page table mappings between the kernel and user space. The root of the problem is that the x86 can only keep one set of physical to virtual mappings on-hand at once. Since the kernel and the processes occupy different virtual mappings, the TLB context would have to be switched not only when the processor changes current user tasks, but also when the current user task calls on the kernel to perform an operation on its behalf. The time taken to change mappings, called the TLB flushing penalty, contributes to a degradation in process performance and has been measured at around 30%[1]. To avoid this penalty, the Kernel and user spaces share a partitioned virtual address space so that the kernel is actually mapped into user space (although protected from user access) and vice versa.

The upshot of this is that the x86 userspace is divided 3GB/1GB with the virtual address range `0x00000000-0xbfffffff` being available for the user process and `0xc0000000-0xffffffff` being reserved for the kernel.

The problem, for the kernel, is that it now only has 1GB of virtual address to play with *including* all memory mapped I/O regions. The result being that `ZONE_NORMAL` actually ends at around 850kb on most x86 boxes. Since the kernel must also manage the mappings for every user process (and these mappings must be memory resident), the larger the physical memory of the kernel becomes, the less of `ZONE_NORMAL` becomes available to the kernel. On a 64GB x86 box, the usable memory becomes minuscule and has lead to the proposal[2] to use a 4G/4G split and just accept the TLB flushing penalty.

1.2 Non-x86 and Virtual Indexing

Most other architectures are rather better implemented and are able to cope easily with separate virtual spaces for the user and the kernel without imposing a performance penalty transitioning from one virtual address space to another. However, there are other problems the kernel's penchant for keeping all memory mapped causes, notably with Virtual Indexing.

Virtual Indexing[3] (VI) means that the CPU cache keeps its data indexed by virtual address (rather than by physical address like the x86 does). The problem this causes is that if multiple virtual address spaces have the same physical address mapped, but at different virtual addresses then the cache may contain duplicate entries, called aliases. Managing these aliases becomes impossible if there are multiple ones that become dirty.

Most VI architectures find a solution to the multiple cache line problem by having a "congruence modulus" meaning that if two virtual addresses are equal modulo this congruence (usually a value around 4MB) then the cache will detect the aliasing and keep only a single copy of the data that will be seen by all the virtual addresses.

The problems arise because, although architectures go to great lengths to make sure all user mappings are congruent, because the ker-

nel memory is always mapped, it is highly unlikely that any given kernel page would be congruent to a user page.

1.3 The solution: Unmapping ZONE_NORMAL

It has already been pointed out[4] that x86 could recover some of its precious ZONE_NORMAL space simply by moving page table entries into unmapped highmem space. However, the penalty of having to map and unmap the page table entries to modify them turned out to be unacceptable.

The solution, though, remains valid. There are many pages of data currently in ZONE_NORMAL that the kernel doesn't ordinarily use. If these could be unmapped and their virtual address space given up then the x86 kernel wouldn't be facing quite such a memory crunch.

For VI architectures, the problems stem from having unallocated kernel memory already mapped. If we could keep the majority of kernel memory unmapped, and map it only when we really need to use it, then we would stand a very good chance of being able to map the memory congruently even in kernel space.

The solution this paper will explore is that of keeping the majority of kernel memory unmapped, mapping it only when it is used.

2 A closer look at Virtual Indexing

As well as the aliasing problem, VI architectures also have issues with I/O coherency on DMA. The essence of the problem stems from the fact that in order to make a device access to physical memory coherent, any cache lines that the processor is holding need to be flushed/invalidates as part of the DMA transaction. In order to do DMA, a device simply presents a physical address to the system with a request to read or write. However, if the processor indexes the caches virtually, it will have no idea whether it is caching this physical address or not. Therefore, in order to give the processor an idea of where in the cache the data might be, the DMA engines on VI architectures also present a virtual index (called the "coherence index") along with the physical address.

2.1 Coherence Indices and DMA

The Coherence Index is computed by the processor on a per page basis, and is used to identify the line in the cache belonging to the physical address the DMA is using.

One will notice that this means the coherence index must be computed on *every* DMA transaction for a *particular* address space (although, if all the addresses are congruent, one may simply pick any one). Since, at the time the dma mapping is done, the only virtual address the kernel knows about is the kernel virtual address, it means that DMA is always done coherently with the kernel.

In turn, since the kernel address is pretty much not congruent with any user address, before the DMA is signalled as being completed to the user process, the kernel mapping and the user mappings must likewise be made coherent (using the `flush_dcache_page()` function). However, since the majority of DMA transactions occur on *user* data in which the kernel has no interest, the extra flush is simply an unnecessary performance penalty.

This performance penalty would be eliminated if either we knew that the designated kernel address was congruent to all the user addresses or we didn't bother to map the DMA region into kernel space and simply computed the coherence index from a given user process. The latter would be preferable from a performance point of view since it eliminates an unneces-

sary map and unmap.

2.2 Other Issues with Non-Congruence

On the parisc architecture, there is an architectural requirement that we don't simultaneously enable multiple read and write translations of a non-congruent address. We can either enable a single write translation or multiple read (but no write) translations. With the current manner of kernel operation, this is almost impossible to satisfy without going to enormous lengths in our page translation and fault routines to work around the issues.

Previously, we were able to get away with ignoring this restriction because the machine would only detect it if we allowed multiple aliases to become dirty (something Linux never does). However, in the next generation systems, this condition will be detected when it occurs. Thus, addressing it has become critical to providing a bootable kernel on these new machines.

Thus, as well as being a simple performance enhancement, removing non-congruence becomes vital to keeping the kernel booting on next generation machines.

2.3 VIPT vs VIVT

This topic is covered comprehensively in [3]. However, there is a problem in VIPT caches, namely that if we are reusing the virtual address in kernel space, we must flush the processor's cache for that page on this re-use otherwise it may fall victim to stale cache references that were left over from a prior use.

Flushing a VIPT cache is easier said than done, since in order to flush, a valid translation must exist for the virtual address in order for the flush to be effective. This causes particular problems for pages that were mapped to a user space process, since the address translations are destroyed *before* the page is finally freed.

3 Kernel Virtual Space

Although the kernel is nominally mapped in the same way the user process is (and can theoretically be fragmented in physical space), in fact it is usually offset mapped. This means there is a simple mathematical relation between the physical and virtual addresses:

$$virtual = physical + \texttt{\_\_PAGE\_OFFSET}$$

where `__PAGE_OFFSET` is an architecture defined quantity. This type of mapping makes it very easy to calculate virtual addresses from physical ones and vice versa without having to go to all the bother (and CPU time) of having to look them up in the kernel page tables.

3.1 Moving away from Offset Mapping

There's another wrinkle on some architectures in that if an interruption occurs, the CPU turns off virtual addressing to begin processing it. This means that the kernel needs to save the various registers and turn virtual addressing back on, all in physical space. If it's no longer a simple matter of subtracting `__PAGE_OFFSET` to get the kernel stack for the process, then extra time will be consumed in the critical path doing potentially cache cold page table lookups.

3.2 Keeping track of Mapped pages

In general, when mapping a page we will either require that it goes in the first available slot (for x86), or that it goes at the first available slot congruent with a given address (for VI architectures). All we really require is a simple mechanism for finding the first free page

virtual address given some specific constraints. However, since the constraints are architecture specific, the specifics of this tracking are also implemented in architectures (see section 5.2 for details on parisc).

3.3 Determining Physical address from Virtual and Vice-Versa

In the Linux kernel, the simple macros `__pa()` and `__va()` are used to do physical to virtual translation. Since we are now filling the mappings in randomly, this is no longer a simple offset calculation.

The kernel does have help for finding the virtual address of a given page. There is an optional `virtual` entry which is turned on and populated with the page's current virtual address when the architecture defines `WANT_PAGE_VIRTUAL`. The `__va()` macro can be programmed simply to do this lookup.

To find the physical address, the best method is probably to look the page up in the kernel page table mappings. This is obviously less efficient than a simple subtraction.

4 Implementing the unmapping of `ZONE_NORMAL`

It is not surprising, given that the entire kernel is designed to operate with `ZONE_NORMAL` mapped it is surprising that unmapping it turns out to be fairly easy. The primary reason for this is the existence of highmem. Since pages in `ZONE_HIGHMEM` are always unmapped and since they are usually assigned to user processes, the kernel must proceed on the assumption that it potentially has to map into its address space any page from a user process that it wishes to touch.

4.1 Booting

The kernel has an entire bootmem API whose sole job is to cope with memory allocations while the system is booting and before paging has been initialised to the point where normal memory allocations may proceed. On parisc, we simply get the available page ranges from the firmware, map them all and turn them over lock stock and barrel to bootmem.

Then, when we're ready to begin paging, we simply release all the unallocated bootmem pages for the kernel to use from its `mem_map`[2] array of pages.

We can implement the unmapping idea simply by covering all our page ranges with an offset map for bootmem, but then unmapping all the unreserved pages that bootmem releases to the `mem_map` array.

This leaves us with the kernel text and data sections contiguously offset mapped, and all other boot time

4.2 Pages Coming From User Space

The standard mechanisms for mapping potential highmem pages from user space for the kernel to see are `kmap`, `kunmap`, `kmap_atomic`, and `kmap_atomic_to_page`. Simply hijacking them and divorcing their implementation from `CONFIG_HIGHMEM` is sufficient to solve all user to kernel problems that arise because of the unmapping of `ZONE_NORMAL`.

4.3 In Kernel Problems: Memory Allocation

Since now every free page in the system will be unmapped, they will have to be mapped

[2] This global array would be a set of per-zone arrays on NUMA

before the *kernel* can use them (pages allocated for use in user space have no need to be mapped additionally in kernel space at allocation time). The engine for doing this is a single point in `__alloc_pages()` which is the central routine for allocating every page in the system. In the single successful page return, the page is mapped for the kernel to use it if `__GFP_HIGH` is not set—this simple test is sufficient to ensure that kernel pages only are mapped here.

The unmapping is done in two separate routines: `__free_pages_ok()` for freeing bulk pages (accumulations of contiguous pages) and `free_hot_cold_page()` for freeing single pages. Here, since we don't know the gfp mask the page was allocated with, we simply check to see if the page is currently mapped, and unmap it if it is before freeing it. There is another side benefit to this: the routine that transfers all the unreserved bootmem to the `mem_map` array does this via `__free_pages()`. Thus, we additionally achieve the unmapping of all the free pages in the system after booting with virtually no additional effort.

4.4 Other Benefits: Variable size pages

Although it wasn't the design of this structure to provide variable size pages, one of the benefits of this approach is now that the pages that are mapped as they are allocated. Since pages in the kernel are allocated with a specified order (the power of two of the number of contiguous pages), it becomes possible to cover them with a TLB entry that is larger than the usual page size (as long as the architecture supports this). Thus, we can take the `order` argument to `__alloc_pages()` and work out the smallest number of TLB entries that we need to allocate to cover it.

Implementation of variable size pages is actually transparent to the system; as far as Linux is concerned, the page table entries it deal with describe 4k pages. However, we add additional flags to the pte to tell the software TLB routine that actually we'd like to use a larger size TLB to access this region.

As a further optimisation, in the architecture specific routines that free the boot mem, we can remap the kernel text and data sections with the smallest number of TLB entries that will entirely cover each of them.

5 Achieving The VI architecture Goal: Fully Congruent Aliasing

The system possesses every attribute it now needs to implement this. We no-longer map any user pages into kernel space unless the kernel actually needs to touch them. Thus, the pages will have congruent user addresses allocated to them in user space *before* we try to map them in kernel space. Thus, all we have to do is track up the free address list in increments of the congruence modulus until we find an empty place to map the page congruently.

5.1 Wrinkles in the I/O Subsystem

The I/O subsystem is designed to operate without mapping pages into the kernel *at all*. This becomes problematic for VI architectures because we have to know the user virtual address to compute the coherence index for the I/O. If the page is unmapped in kernel space, we can no longer make it coherent with the kernel mapping and, unfortunately, the information in the BIO is insufficient to tell us the user virtual address.

The proposal for solving this is to add an architecture defined set of elements to `struct bio_vec` and an architecture specific function for populating this (possibly empty) set of elements as the biovec is created. In parisc,

we need to add an extra unsigned long for the coherence index, which we compute from a pointer to the mm and the user virtual address. The architecture defined components are pulled into `struct scatterlist` by yet another callout when the request is mapped for DMA.

5.2 Tracking the Mappings in `ZONE_DMA`

Since the tracking requirements vary depending on architectures: x86 will merely wish to find the first free pte to place a page into; however VI architectures will need to find the first free pte satisfying the congruence requirements (which vary by architecture), the actual mechanism for finding a free pte for the mapping needs to be architecture specific.

On parisc, all of this can be done in `kmap_kernel()` which merely uses rmap[5] to determine if the page is mapped in user space and find the congruent address if it is. We use a simple hash table based bitmap with one bucket representing the set of available congruent pages. Thus, finding a page congruent to any given virtual address is the simple computation of finding the first set bit in the congruence bucket. To find an arbitrary page, we keep a global bucket counter, allocating a page from that bucket and then incrementing the counter[3].

6 Implementation Details on PA-RISC

Since the whole thrust of this project was to improve the kernel on PA-RISC (and bring it back into architectural compliance), it is appropriate to investigate some of the other problems that turned up during the implementation.

[3]This can all be done locklessly with atomic increments, since it doesn't really matter if we get two allocations from the same bucket because of race conditions

6.1 Equivalent Mapping

The PA architecture has a software TLB meaning that in Virtual mode, if the CPU accesses an address that isn't in the CPU's TLB cache, it will take a TLB fault so the software routine can locate the TLB entry (by walking the page tables) and insert it into the CPU's TLB. Obviously, this type of interruption must be handled purely by referencing physical addresses. In fact, the PA CPU is designed to have fast and slow paths for faults and interruptions. The fast paths (since they cannot take another interruption, i.e. not a TLB miss fault) must all operate on physical addresses. To assist with this, the PA CPU even turns off virtual addressing when it takes an interruption.

When the CPU turns off virtual address translation, it is said to be operating in absolute mode. All address accesses in this mode are physical. However, all accesses in this mode also go through the CPU cache (which means that for this particular mode the cache is actually Physically Indexed). Unfortunately, this can also set up unwanted aliasing between the physical address and its virtual translation. The fix for this is to obey the architectural definition for "equivalent mapping." Equivalent mapping is defined as virtual and physical addresses being equal; however, we benefit from the obvious loophole in that the physical and virtual addresses don't have to be exactly equal, merely equal modulo the congruent modulus.

All of this means that when a page is allocated for use by the kernel, we must determine if it will ever be used in absolute mode, and make it equivalently mapped if it will be. At the time of writing, this was simply implemented by making all kernel allocated pages equivalent. However, really all that needs to be equivalently mapped is

1. the page tables (pgd, pmd and pte),

2. the task structure and
3. the kernel stacks.

6.2 Physical to Virtual address Translation

In the interruption slow path, where we save all the registers and transition to virtual mode, there is a point where execution must be switched (and hence pointers moved from physical to virtual). Currently, with offset mapping, this is simply done by and addition of `__PAGE_OFFSET`. However, in the new scheme we cannot do this, nor can we call the address translation functions when in absolute mode. Therefore, we had to reorganise the interruption paths in the PA code so that both the physical and virtual address was available. Currently parisc uses a control register (`%cr30`) to store the virtual address of the `struct thread_info`. We altered all paths to change `%cr30` to contain the physical address of `struct thread_info` and also added a physical address pointer to the `struct task_struct` to the thread info. This is sufficient to perform all the necessary register saves in absolute addressing mode.

6.3 Flushing on Page Freeing

as was documented in section 2.3, we need to find a way of flushing a user virtual address *after* its translation is gone. Actually, this turns out to be quite easy on PARISC. We already have an area of memory (called the tmpalias space) that we use to copy to priming the user cache (it is simply a 4MB memory area we dynamically program to map to the page). Therefore, as long as we know the user virtual address, we can simply flush the page through the tmpalias space. In order to confound any attempted kernel use of this page, we reserve a separate 4MB virtual area that produces a page fault if referenced, and point the page's `virtual` address into this when it is *removed* from process mappings (so that any kernel attempt to use the page produces an immediate fault). Then, when the page is freed, if its `virtual` pointer is within this range, we convert it to a tmpalias address and flush it using the tmpalias mechanism.

7 Results and Conclusion

The best result is that on a parisc machine, the total amount of memory the operational kernel keeps mapped is around 10MB (although this alters depending on conditions).

The current implementation makes all pages congruent or equivalent, but the allocation routine contains `BUG_ON()` asserts to detect if we run out of equivalent addresses. So far, under fairly heavy stress, none of these has tripped.

Although the primary reason for the unmapping was to move parisc back within its architectural requirements, it also produces a knock on effect of speeding up I/O by eliminating the cache flushing from kernel to user space. At the time of writing, the effects of this were still unmeasured, but expected to be around 6% or so.

As a final side effect, the flush on free necessity releases the parisc from a very stringent "flush the entire cache on process death or exec" requirement that was producing horrible latencies in the parisc fork/exec. With this code in place, we see a vast (50%) improvement in the fork/exec figures.

References

[1] Andrea Arcangeli *3:1 4:4 100HZ 1000HZ comparison with the HINT benchmark* 7 April 2004 `http://www.kernel.org/pub/`

linux/kernel/people/andrea/misc/31-44-100-1000/31-44-100-1000.html

[2] Ingo Molnar *[announce, patch] 4G/4G split on x86, 64 GB RAM (and more) support* 8 July 2003 http://marc.theaimsgroup.com/?t=105770467300001

[3] James E.J. Bottomley *Understanding Caching* Linux Journal January 2004, Issue 117 p58

[4] Ingo Molnar *[patch] simpler 'highpte' design* 18 February 2002 http://marc.theaimsgroup.com/?l=linux-kernel&m=101406121032371

[5] Rik van Riel *Re: Rmap code?* 22 August 2001 http://marc.theaimsgroup.com/?l=linux-mm&m=99849912207578

Linux Virtualization on IBM POWER5 Systems

Dave Boutcher
IBM
boutcher@us.ibm.com

Dave Engebretsen
IBM
engebret@us.ibm.com

Abstract

In 2004 IBM® is releasing new systems based on the POWER5™ processor. There is new support in both the hardware and firmware for virtualization of multiple operating systems on a single platform. This includes the ability to have multiple operating systems share a processor. Additionally, a hypervisor firmware layer supports virtualization of I/O devices such as SCSI, LAN, and console, allowing limited physical resources in a system to be shared.

At its extreme, these new systems allow 10 Linux images per physical processor to run concurrently, contending for and sharing the system's physical resources. All changes to support these new functions are in the 2.4 and 2.6 Linux kernels.

This paper discusses the virtualization capabilities of the processor and firmware, as well as the changes made to the PPC64 kernel to take advantage of them.

1 Introduction

IBM's new POWER5** processor is being used in both IBM iSeries® and pSeries® systems capable of running any combination of Linux, AIX®, and OS/400® in logical partitions. The hardware and firmware, including a *hypervisor* [AAN00], in these systems provide the ability to create "virtual" system images with virtual hardware. The virtualization technique used on POWER™ hardware is known as paravirtualization, where the operating system is modified in select areas to make calls into the hypervisor. PPC64 Linux has been enhanced to make use of these virtualization interfaces. Note that the same PPC64 Linux kernel binary works on both virtualized systems and previous "bare metal" pSeries systems that did not offer a hypervisor.

All changes related to virtualization have been made in the kernel, and almost exclusively in the PPC64 portion of the code. One challenge has been keeping as much code common as possible between POWER5 portions of the code and other portions, such as those supporting the Apple G5.

Like previous generations of POWER processors such as the RS64 and POWER4™ families, POWER5 includes hardware enablement for logical partitioning. This includes features such as a hypervisor state which is more privileged than supervisor state. This higher privilege state is used to restrict access to system resources, such as the hardware page table, to hypervisor only access. All current systems based on POWER5 run in a hypervised environment, even if only one partition is active on the system.

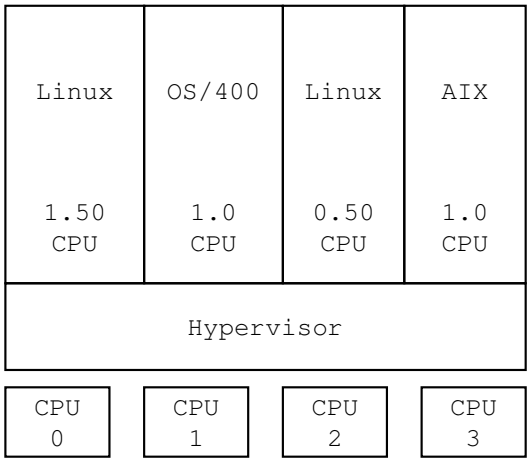


Figure 1: POWER5 Partitioned System

2 Processor Virtualization

2.1 Virtual Processors

When running in a partition, the operating system is allocated virtual processors (VP's), where each VP can be configured in either shared or dedicated mode of operation. In shared mode, as little as 10%, or 10 *processing units*, of a physical processor can be allocated to a partition and the hypervisor layer timeslices between the partitions. In dedicated mode, 100% of the processor is given to the partition such that its capacity is never multiplexed with another partition.

It is possible to create more virtual processors in the partition than there are physical processors on the system. For example, a partition allocated 100 processing units (the equivalent of 1 processor) of capacity could be configured to have 10 virtual processors, where each VP has 10% of a physical processor's time. While not generally valuable, this extreme configuration can be used to help test SMP configurations on small systems.

On POWER5 systems with multiple logical partitions, an important requirement is to be able to move processors (either shared or dedicated) from one logical partition to another. In the case of dedicated processors, this truly means moving a CPU from one logical partition to another. In the case of shared processors, it means adjusting the number of processors used by Linux on the fly.

This "hotplug CPU" capability is far more interesting in this environment than in the case that the covers are going to be removed from a real system and a CPU physically added. The goal of virtualization on these systems is to dynamically create and adjust operating system images as required. Much work has been done, particularly by Rusty Russell, to get the architecture independent changes into the mainline kernel to support hotplug CPU.

Hypervisor interfaces exist that help the operating system optimize its use of the physical processor resources. The following sections describe some of these mechanisms.

2.2 Virtual Processor Area

Each virtual processor in the partition can create a *virtual processor area* (VPA), which is a small (one page) data structure shared between the hypervisor and the operating system. Its primary use is to communicate information between the two software layers. Examples of the information that can be communicated in the VPA include whether the OS is in the idle loop, if floating point and performance counter register state must be saved by the hypervisor between operating system dispatches, and whether the VP is running in the partition's operating system.

2.3 Spinlocks

The hypervisor provides an interface that helps minimize wasted cycles in the operating system when a lock is held. Rather than simply spin on the held lock in the OS, a new hypervi-

sor call, `h_confer`, has been provided. This interface is used to confer any remaining virtual processor cycles from the lock requester to the lock holder.

The PPC64 spinlocks were changed to identify the logical processor number of the lock holder, examine that processor's VPA *yield count* field to determine if it is not running in the OS (even values indicate the VP is running in the OS), and to make the `h_confer` call to the hypervisor to give any cycles remaining in the virtual processor's timeslice to the lock holder. Obviously, this more expensive leg of spinlock processing is only taken if the spinlock cannot be immediately acquired. In cases where the lock is available, no additional pathlength is incurred.

2.4 Idle

When the operating system no longer has active tasks to run and enters its idle loop, the `h_cede` interface is used to indicate to the hypervisor that the processor is available for other work. The operating system simply sets the VPA *idle* bit and calls `h_cede`. Under this call, the hypervisor is free to allocate the processor resources to another partition, or even to another virtual processor within the same partition. The processor is returned to the operating system if an external, decrementer (timer), or interprocessor interrupt occurs. As an alternative to sending an IPI, the ceded processor can be awoken by another processor calling the *h_prod* interface, which has slightly less overhead in this environment.

Making use of the cede interface is especially important on systems where partitions configured to run *uncapped* exist. In uncapped mode, any physical processor cycles not used by other partitions can be allocated by the hypervisor to a non-idle partition, even if that partition has already consumed its defined quantity of processor units. For example, a partition that is defined as uncapped, 2 virtual processors, and 20 processing units could consume 2 full processors (200 processing units), if all other partitions are idle.

2.5 SMT

The POWER5 processor provides symmetric multithreading (SMT) capabilities that allow two threads of execution to simultaneously execute on one physical processor. This results in twice as many processor contexts being presented to the operating system as there are physical processors. Like other processor threading mechanisms found in POWER RS64 and Intel® processors, the goal of SMT is to enable higher processor utilization.

At Linux boot, each processor thread is discovered in the open firmware device tree and a logical processor is created for Linux. A command line option, `smt-enabled = [on, off, dynamic]`, has been added to allow the Linux partition to config SMT in one of three states. The *on* and *off* modes indicate that the processor always runs with SMT either on or off. The dynamic mode allows the operating system and firmware to dynamically configure the processor to switch between threaded (SMT) and a single threaded (ST) mode where one of the processor threads is dormant. The hardware implementation is such that running in ST mode can provide additional performance when only a single task is executing.

Linux can cause the processor to switch between SMT and ST modes via the `h_cede` hypervisor call interface. When entering its idle loop, Linux sets the VPA *idle* state bit, and after a selectable delay, calls `h_cede`. Under this interface, the hypervisor layer determines if only one thread is idle, and if so, switches the processor into ST mode. If both threads are

idle (as indicated by the VPA *idle* bit), then the hypervisor keeps the processor in SMT mode and returns to the operating system.

The processor switches back to SMT mode if an external or decrementer interrupt is presented, or if another processor calls the `h_prod` interface against the dormant thread.

3 Memory Virtualization

Memory is virtualized only to the extent that all partitions on the system are presented a contiguous range of logical addresses that start at zero. Linux sees these logical addresses as its real storage. The actual real memory is allocated by the hypervisor from any available space throughout the system, managing the storage in *logical memory blocks* (LMB's). Each LMB is presented to the partition via a memory node in the open firmware device tree. When Linux creates a mapping in the hardware page table for effective addresses, it makes a call to the hypervisor (`h_enter`) indicating the effective and partition logical address. The hypervisor translates the logical address to the corresponding real address and inserts the mapping into the hardware page table.

One additional layer of memory virtualization managed by the hypervisor is a *real mode offset* (RMO) region. This is a 128 or 256 MB region of memory covering the first portion of the logical address space within a partition. It can be accessed by Linux when address relocation is off, for example after an exception occurs. When a partition is running relocation off and accesses addresses within the RMO region, a simple offset is added by the hardware to generate the actual storage access. In this manner, each partition has what it considers logical address zero.

4 I/O Virtualization

Once CPU and memory have been virtualized, a key requirement is to provide virtualized I/O. The goal of the POWER5 systems is to have, for example, 10 Linux images running on a small system with a single CPU, 1GB of memory, and a single SCSI adapter and Ethernet adapter.

The approach taken to virtualize I/O is a cooperative implementation between the hypervisor and the operating system images. One operating system image always "owns" physical adapters and manages all I/O to those adapters (DMA, interrupts, etc.)

The hypervisor and Open Firmware then provide "virtual" adapters to any operating systems that require them. Creation of virtual adapters is done by the system administrator as part of logically partitioning the system. A key concept is that these virtual adapters do not interact in any way with the physical adapters. The virtual adapters interact with other operating systems in other logical partitions, which may choose to make use of physical adapters.

Virtual adapters are presented to the operating system in the Open Firmware device tree just as physical adapters are. They have very similar attributes as physical adapters, including DMA windows and interrupts.

The adapters currently supported by the hypervisor are virtual SCSI adapters, virtual Ethernet adapters, and virtual TTY adapters.

4.1 Virtual Bus

Virtual adapters, of course, exist on a virtual bus. The bus has slots into which virtual adapters are configured. The number of slots available on the virtual bus is configured by the system administrator. The goal is to make

the behavior of virtual adapters consistent with physical adapters. The virtual bus is *not* presented as a PCI bus, but rather as its own bus type.

4.2 Virtual LAN

Virtual LAN adapters are conceptually the simplest kind of virtual adapter. The hypervisor implements a switch, which supports 802.1Q semantics for having multiple VLANs share a physical switch. Adapters can be marked as 802.1Q aware, in which case the hypervisor expects the operating system to handle the 802.1Q VLAN headers, or 802.1Q unaware, in which case the hypervisor connects the adapter to a single VLAN. Multiple virtual Ethernet adapters can be created for a given partition.

Virtual Ethernet adapters have an additional attribute called "Trunk Adapter." An adapter marked as a Trunk Adapter will be delivered all frames that don't match any MAC address on the virtual Ethernet. This is similar, but not identical, to promiscuous mode on a real adapter.

For a logical partition to have network connectivity to the outside world, the partition owning a "real" network adapter generally has both the real Ethernet adapter and a virtual Ethernet adapter marked as a Trunk adapter. That partition then performs either routing or bridging between the real adapter and the virtual adapter. The Linux bridge-utils package works well to bridge the two kinds of networks.

Note that there is no architected link between the real and virtual adapters, it is the responsibility of some operating system to route traffic between them.

The implementation of the virtual Ethernet adapters involves a number of hypervisor interfaces. Some of the more significant interfaces are `h_register_logical_lan` to establish the initial link between a device driver and a virtual Ethernet device, `h_send_logical_lan` to send a frame, and `h_add_logical_lan_buffer` to tell the hypervisor about a data buffer into which a received frame is to be placed. The hypervisor interfaces then support either polled or interrupt driven notification of new frames arriving.

For additional information on the virtual Ethernet implementation, the code is the documentation (`drivers/net/ibmveth.c`).

4.3 Virtual SCSI

Unlike virtual Ethernet adapters, virtual SCSI adapters come in two flavors. A "client" virtual SCSI adapter behaves just as a regular SCSI host bus adapter and is implemented within the SCSI framework of the Linux kernel. The SCSI mid-layer issues standard SCSI commands such as Inquiry to determine devices connected to the adapter, and issues regular SCSI operations to those devices.

A "server" virtual SCSI adapter, generally in a different partition than the client, receives all the SCSI commands from the client and is responsible for handling them. The hypervisor is not involved in what the server does with the commands. There is no requirement for the server to link a virtual SCSI adapter to any kind of real adapter. The server can process and return SCSI responses in any fashion it likes. If it happens to issue I/O operations to a real adapter as part of satisfying those requests, that is an implementation detail of the operating system containing the server adapter.

The hypervisor provides two very primitive interpartition communication mechanisms on which the virtual SCSI implementation is built. There is a queue of 16 byte messages referred to as a "Command/Response Queue" (CRQ). Each partition provides the hypervisor with a

page of memory where its receive queue resides, and a partition wishing to send a message to its partner's queue issues an `h_send_crq` hypervisor call. When a message is received on the queue, an interrupt is (optionally) generated in the receiving partition.

The second hypervisor mechanism is a facility for issuing DMA operations between partitions. The `h_copy_rdma` call is used to DMA a block of memory from the memory space of one logical partition to the memory space of another.

The virtual SCSI interpartition protocol is implemented using the ANSI "SCSI RDMA Protocol" (SRP) (available at `http://www.t10.org`). When the client wishes to issue a SCSI operation, it builds an SRP frame, and sends the address of the frame in a 16 byte CRQ message. The server DMA's the SRP frame from the client, and processes it. The SRP frame may itself contain DMA addresses required for data transfer (read or write buffers, for example) which may require additional interpartition DMA operations. When the operation is complete, the server DMA's the SRP response back to the same location as the SRP command came from and sends a 16 byte CRQ message back indicating that the SCSI command has completed.

The current Linux virtual SCSI server decodes incoming SCSI commands and issues block layer commands (`generic_make_request`). This allows the SCSI server to share any block device (e.g., `/dev/sdb6` or `/dev/loop0`) with client partitions as a virtual SCSI device.

Note that consideration was given to using protocols such as iSCSI for device sharing between partitions. The virtual SCSI SRP design above, however, is a much simpler design that does not rely on riding above an existing IP stack. Additionally, the ability to use DMA operations between partitions fits much better into the SRP model than an iSCSI model.

The Linux virtual SCSI client (`drivers/scsi/ibmvscsi/ibmvscsi.c`) is close, at the time of writing, to being accepted into the Linux mainline. The Linux virtual SCSI server is sufficiently unlike existing SCSI drivers that it will require much more mailing list "discussion."

4.4 Virtual TTY

In addition to virtual Ethernet and SCSI adapters, the hypervisor supports virtual serial (TTY) adapters. As with SCSI adapter, these can be configured as "client" adapters, and "server" adapters and connected between partitions. The first virtual TTY adapter is used as the system console, and is treated specially by the hypervisor. It is automatically connected to the partition console on the Hardware Management Console.

To date, multiple concurrent "consoles" have not been implemented, but they could be. Similarly, this interface could be used for kernel debugging as with any serial port, but such an implementation has not been done.

5 Dynamic Resource Movement

As mentioned for processors, the logical partition environment lends itself to moving resources (processors, memory, I/O) between partitions. In a perfect world, such movement should be done dynamically while the operating system is running. Dynamic movement of processors is currently being implemented, and dynamic movement of I/O devices (including dynamically adding and removing virtual I/O devices) is included in the kernel mainline.

The one area for future work in Linux is the dynamic movement of memory into and out of an

active partition. This function is already supported on other POWER5 operating systems, so there is an opportunity for Linux to catch up.

6 Multiple Operating Systems

A key feature of the POWER5 systems is the ability to run different operating systems in different logical partitions on the same physical system. The operating systems currently supported on the POWER5 hardware are AIX, OS/400, and Linux.

While running multiple operating systems, all of the functions for interpartion interaction described above must work between operating systems. For example, idle cycles from an AIX partition can be given to Linux. A processor can be moved from OS/400 to Linux while both operating systems are active.

For I/O, multiple operating systems must be able to communicate over the virtual Ethernet, and SCSI devices must be sharable from (say) an AIX virtual SCSI server to a Linux virtual SCSI client.

These requirements, along with the architected hypervisor interfaces, limit the ability to change implementations just to fit a Linux kernel internal behavior.

7 Conclusions

While many of the basic virtualization technologies described in this paper existed in the Linux implementation provided on POWER RS64 and POWER4 iSeries systems [Bou01], they have been significantly enhanced for POWER5 to better use the firmware provided interfaces.

The introduction of POWER5-based systems converged all of the virtualization interfaces provided by firmware on legacy iSeries and pSeries systems to a model in line with the legacy pSeries partitioned system architecture. As a result much of the PPC64 Linux virtualization code was updated to use these new virtualization interface definitions.

8 Acknowledgments

The authors would like to thank the entire Linux/PPC64 team for the work that went into the POWER5 virtualization effort. In particular Anton Blanchard, Paul Mackerras, Rusty Rusell, Hollis Blanchard, Santiago Leon, Ryan Arnold, Will Schmidt, Colin Devilbiss, Kyle Lucke, Mike Corrigan, Jeff Scheel, and David Larson.

9 Legal Statement

References

[AAN00] Bill Armstrong, Troy Armstrong, Naresh Nayar, Ron Peterson, Tom Sand, and Jeff Scheel. *Logical Partitioning*, `http://www-1.ibm.com/servers/eserver/iseries/beyondtech/lpar.htm`.

[Bou01] David Boutcher, *The iSeries Linux Kernel* 2001 Linux Symposium, (July 2001).

The State of ACPI in the Linux Kernel

A. Leonard Brown
Intel
len.brown@intel.com

Abstract

ACPI puts Linux in control of configuration and power management. It abstracts the platform BIOS and hardware so Linux and the platform can interoperate while evolving independently.

This paper starts with some background on the ACPI specification, followed by the state of ACPI deployment on Linux.

It describes the implementation architecture of ACPI on Linux, followed by details on the configuration and power management features.

It closes with a summary of ACPI bugzilla activity, and a list of what is next for ACPI in Linux.

1 ACPI Specification Background

> "ACPI (Advanced Configuration and Power Interface) is an open industry specification co-developed by Hewlett-Packard, Intel, Microsoft, Phoenix, and Toshiba.
>
> ACPI establishes industry-standard interfaces for OS-directed configuration and power management on laptops, desktops, and servers.
>
> ACPI evolves the existing collection of power management BIOS code, Advanced Power Management (APM) application programming interfaces (APIs, PNPBIOS APIs, Multiprocessor Specification (MPS) tables and so on into a well-defined power management and configuration interface specification."[1]

ACPI 1.0 was published in 1996. 2.0 added 64-bit support in 2000. ACPI 3.0 is expected in summer 2004.

2 Linux ACPI Deployment

Linux supports ACPI on three architectures: `ia64`, `i386`, and `x86_64`.

2.1 `ia64` Linux/ACPI support

Most `ia64` platforms require ACPI support, as they do not have the legacy configuration methods seen on `i386`. All the Linux distributions that support `ia64` include ACPI support, whether they're based on Linux-2.4 or Linux-2.6.

2.2 `i386` Linux/ACPI support

Not all Linux-2.4 distributions enabled ACPI by default on `i386`. Often they used just enough table parsing to enable Hyper-Threading (HT), ala `acpi=ht` below, and relied on MPS and PIRQ routers to configure the

[1] `http://www.acpi.info`

```
setup_arch()
  dmi_scan_machine()
    Scan DMI blacklist
    BIOS Date vs Jan 1, 2001
  acpi_boot_init()
    acpi_table_init()
      locate and checksum all ACPI tables
      print table headers to console
    acpi_blacklisted()
      ACPI table headers vs. blacklist
    parse(BOOT) /* Simple Boot Flags */
    parse(FADT) /* PM timer address */
    parse(MADT) /* LAPIC, IOAPIC */
    parse(HPET) /* HiPrecision Timer */
    parse(MCFG) /* PCI Express base */
```

Figure 1: Early ACPI init on `i386`

machine. Some included ACPI support by default, but required the user to add `acpi=on` to the cmdline to enable it.

So far, the major Linux 2.6 distributions all support ACPI enabled by default on `i386`.

Several methods are used to make it more practical to deploy ACPI onto `i386` installed base. Figure 1 shows the early ACPI startup on the `i386` and where these methods hook in.

1. Most modern system BIOS support DMI, which exports the date of the BIOS. Linux DMI scan in `i386` disables ACPI on platforms with a BIOS older than January 1, 2001. There is nothing magic about this date, except it allowed developers to focus on recent platforms without getting distracted debugging issues on very old platforms that:

   (a) had been running Linux w/o ACPI support for years.

   (b) had virtually no chance of a BIOS update from the OEM.

   Boot parameter `acpi=force` is available to enable ACPI on platforms older than the cutoff date.

2. DMI also exports the hardware manufacturer, baseboard name, BIOS version, etc. that you can observe with `dmidecode`.[2] `dmi_scan.c` has a general purpose blacklist that keys off this information, and invokes various platform-specific workarounds. `acpi=off` is the most severe—disabling all ACPI support, even the simple table parsing needed to enable Hyper-Threading (HT). `acpi=ht` does the same, excepts parses enough tables to enable HT. `pci=noacpi` disables ACPI for PCI enumeration and interrupt configuration. And `acpi=noirq` disables ACPI just for interrupt configuration.

3. The ACPI tables also contain header information, which you see near the top of the kernel messages. ACPI maintains a blacklist based on the table headers. But this blacklist is somewhat primitive. When an entry matches the system, it either prints warnings or invokes `acpi=off`.

All three of these methods share the problem that if they are successful, they tend to hide root-cause issues in Linux that should be fixed. For this reason, adding to the blacklists is discouraged in the upstream kernel. Their main value is to allow Linux distributors to quickly react to deployment issues when they need to support deviant platforms.

2.3 `x86_64` Linux/ACPI support

All `x86_64` platforms I've seen include ACPI support. The major `x86_64` Linux distributions, whether Linux-2.4 or Linux-2.6 based, all support ACPI.

[2] `http://www.nongnu.org/dmidecode`

3 Implementation Overview

The ACPI specification describes platform registers, ACPI tables, and operation of the ACPI BIOS. Figure 2 shows these ACPI components logically as a layer above the platform specific hardware and firmware.

The ACPI kernel support centers around the ACPICA (ACPI Component Architecture[3]) core. ACPICA includes the AML[4] interpreter that implements ACPI's hardware abstraction. ACPICA also implements other OS-agnostic parts of the ACPI specification. The ACPICA code does not implement any policy, that is the realm of the Linux-specific code. A single file, `osl.c`, glues ACPICA to the Linux-specific functions it requires.

The box in Figure 2 labeled "Linux/ACPI" represents the Linux-specific ACPI code, including boot-time configuration.

Optional "ACPI drivers," such as Button, Battery, Processor, etc. are (optionally loadable) modules that implement policy related to those specific features and devices.

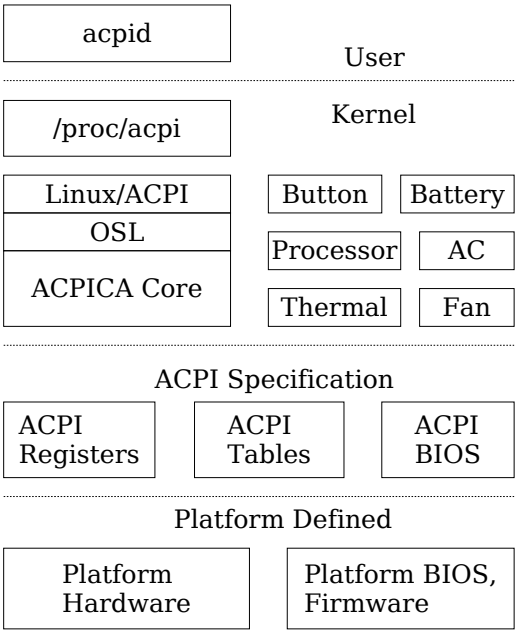


Figure 2: Implementation Architecture

3.1 Events

ACPI registers for a "System Control Interrupt" (SCI) and all ACPI events come through that interrupt.

The kernel interrupt handler de-multiplexes the possible events using ACPI constructs. In some cases, it then delivers events to a user-space application such as `acpid` via `/proc/acpi/events`.

[3] `http://www.intel.com/technology/iapc/acpi`

[4] AML, ACPI Machine Language.

4 ACPI Configuration

Interrupt configuration on `i386` dominated the ACPI bug fixing activity over the last year.

The algorithm to configure interrupts on an `i386` system with an IOAPIC is shown in Figure 3. ACPI mandates that all PIC mode IRQs be identity mapped to IOAPIC pins. Exceptions are specified in MADT [5] interrupt source override entries.

Over-rides are often used, for example, to specify that the 8254 timer on IRQ0 in PIC mode does not use pin0 on the IOAPIC, but uses pin2. Over-rides also often move the ACPI SCI to a different pin in IOAPIC mode than it had in PIC mode, or change its polarity or trigger from the default.

[5] MADT, Multiple APIC Description Table.

```
setup_arch()
 acpi_boot_init()
  parse(MADT);
   parse(LAPIC); /* processors */
   parse(IOAPIC)
    parse(INT_SRC_OVERRIDE);
    add_identity_legacy_mappings();
    /* mp_irqs[] initialized */

init()
 smp_boot_cpus()
  setup_IO_APIC()
   enable_IO_APIC();
   setup_IO_APIC_irqs(); /* mp_irqs[] */
 do_initcalls()
  acpi_init()
"ACPI: Subsystem revision 20040326"
   acpi_initialize_subsystem();
   /* AML interpreter */
   acpi_load_tables(); /* DSDT */
   acpi_enable_subsystem();
   /* HW into ACPI mode */
"ACPI: Interpreter enabled"
   acpi_bus_init_irq();
    AML(_PIC, PIC | IOAPIC | IOSAPIC);

  acpi_pci_link_init()
   for(every PCI Link in DSDT)
    acpi_pci_link_add(Link)
     AML(_PRS, Link);
     AML(_CRS, Link);
"... Link [LNKA] (IRQs 9 10 *11)"

  pci_acpi_init()
"PCI: Using ACPI for IRQ routing"
  acpi_irq_penalty_init();
  for (PCI devices)
   acpi_pci_irq_enable(device)
    acpi_pci_irq_lookup()
     find _PRT entry
      if (Link) {
       acpi_pci_link_get_irq()
        acpi_pci_link_allocate()
         examine possible & current IRQs
          AML(_SRS, Link)
      } else {
        use hard-coded IRQ in _PRT entry
      }
    acpi_register_gsi()
     mp_register_gsi()
      io_apic_set_pci_routing()
"PCI: PCI interrupt 00:06.0[A] ->
  GSI 26 (level, low) -> IRQ 26"
```

Figure 3: Interrupt Initialization

So after identifying that the system will be in IOAPIC mode, the 1st step is to record all the Interrupt Source Overrides in `mp_irqs[]`. The second step is to add the legacy identity mappings where pins and IRQs have not been consumed by the over-rides.

Step three is to digest `mp_irqs[]` in `setup_IO_APIC_irqs()`, just like it would be if the system were running in legacy MPS mode.

But that is just the start of interrupt configuration in ACPI mode. The system still needs to enable the mappings for PCI devices, which are stored in the DSDT[6] _PRT[7] entries. Further, the _PRT can contain both static entries, analogous to MPS table entries, or it can contain dynamic _PRT entries that use PCI Interrupt Link Devices.

So Linux enables the AML interpreter and informs the ACPI BIOS that it plans to run the system in IOAPIC mode.

Next the PCI Interrupt Link Devices are parsed. These “links” are abstract versions of what used to be called PIRQ-routers, though they are more general. `acpi_pci_link_init()` searches the DSDT for Link Devices and queries each about the IRQs it can be set to (_PRS)[8] and the IRQ that it is already set to (_CRS)[9]

A penalty table is used to help decide how to program the PCI Interrupt Link Devices. Weights are statically compiled into the table to avoid programming the links to well known legacy IRQs. `acpi_irq_penalty_init()` updates the table to add penalties to the IRQs where the Links have possible set-

[6] DSDT, Differentiated Services Description Table, written in AML

[7] _PRT, PCI Routing Table

[8] PRS, Possible Resource Settings.

[9] CRS, Current Resource Settings.

tings. The idea is to minimize IRQ sharing, while not conflicting with legacy IRQ use. While it works reasonably well in practice, this heuristic is inherently flawed because it assumes the legacy IRQs rather than asking the DSDT what legacy IRQs are actually in use.[10]

The PCI sub-system calls `acpi_pci_irq_enable()` for every device. ACPI looks up the device in the _PRT by device-id and if it a simple static entry, programs the IOAPIC. If it is a dynamic entry, `acpi_pci_link_allocate()` chooses an IRQ for the link and programs the link via AML (_SRS).[11] Then the associated IOAPIC entry is programmed.

Later, the drivers initialize and call `request_irq(IRQ)` with the IRQ the PCI sub-system told it to request.

One issue we have with this scheme is that it can't automatically recover when the heuristic balancing act fails. For example when the parallel port grabs IRQ7 and a PCI Interrupt Links gets programmed to the same IRQ, then `request_irq(IRQ)` correctly fails to put ISA and PCI interrupts on the same pin. But the system doesn't realize that one of the contenders could actually be re-programmed to a different IRQ.

The fix for this issue will be to delete the heuristic weights from the IRQ penalty table. Instead the kernel should scan the DSDT to enumerate exactly what legacy devices reserve exactly what IRQs.[12]

[10] In PIC mode, the default is to keep the BIOS provided current IRQ setting, unless cmdline `acpi_irq_balance` is used. Balancing is always enabled in IOAPIC mode.

[11] SRS, Set Resource Setting

[12] bugzilla 2733

4.1 Issues With PCI Interrupt Link Devices

Most of the issues have been with PCI Interrupt Link Devices, an ACPI mechanism primarily used to replace the chip-set-specific Legacy PIRQ code.

- The status (_STA) returned by a PCI Interrupt Link Device does not matter. Some systems mark the ones we should use as enabled, some do not.
- The status set by Linux on a link is important on some chip sets. If we do not explicitly disable some unused links, they result in tying together IRQs and can cause spurious interrupts.
- The current setting returned by a link (_CRS) can not always be trusted. Some systems return invalid settings always. Linux must assume that when it sets a link, the setting was successful.
- Some systems return a current setting that is outside the list of possible settings. Per above, this must be ignored and a new setting selected from the possible-list.

4.2 Issues With ACPI SCI Configuration

Another area that was ironed out this year was the ACPI SCI (System Control Interrupt). Originally, the SCI was always configured as level/low, but SCI failures didn't stop until we implemented the algorithm in Figure 4. During debugging, the kernel gained the cmdline option that applies to either PIC or IOAPIC mode: `acpi_sci={level,edge,high,low}` but production systems seem to be working properly and this has seen use recently only to work around prototype BIOS bugs.

```
if (PIC mode) {
  set ELCR to level trigger();
} else { /* IOAPIC mode */
  if (Interrupt Source Override) {
    Use IRQ specified in override
    if(trigger edge or level)
      use edge or level
    else (compatible trigger)
      use level

    if (polarity high or low)
      use high or low
    else
      use low
  } else { /* no Override */
    use level-trigger
    use low-polarity
  }
}
```

Figure 4: SCI configuration algorithm

4.3 Unresolved: Local APIC Timer Issue

The most troublesome configuration issue today is that many systems with no IO-APIC will hang during boot unless their LOCAL-APIC has been disabled, eg. by booting `nolapic`. While this issue has gone away on several systems with BIOS upgrades, entire product lines from high-volume OEMS appear to be subject to this failure. The current workaround to disable the LAPIC timer for the duration of the SMI-CMD update that enables ACPI mode. [13]

4.4 Wanted: Generic Linux Driver Manager

The ACPI DSDT enumerates motherboard devices via PNP identifiers. This method is used to load the ACPI specific devices today, eg. battery, button, fan, thermal etc. as well as `8550_acpi`. PCI devices are enumerated via PCI-ids from PCI config space. Legacy devices probe out using hard-coded address values.

But a device driver should not have to know or care how it is enumerated by its parent bus. An 8250 driver should worry about the 8250 and not if it is being discovered by legacy means, ACPI enumeration, or PCI.

[13] `http://bugzilla.kernel.org 1269`

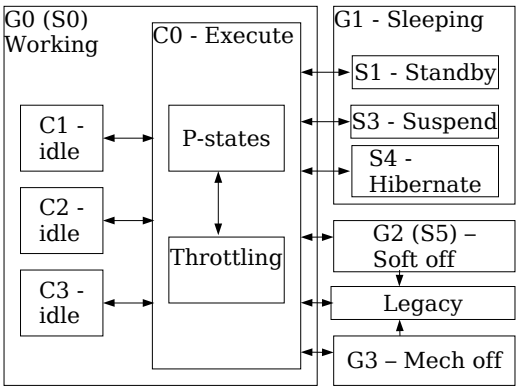


Figure 5: ACPI Global, CPU, and Sleep states.

One fix would be to be to abstract the PCI-ids, PNP-ids, and perhaps even some hard-coded values into a generic device manager directory that maps them to device drivers.

This would simply add a veneer to the PCI device configuration, simplifying a very small number of drivers that can be configured by PCI or ACPI. However, it would also fix the real issue that the configuration information in the ACPI DSDT for most motherboard devices is currently not parsed and not communicated to any Linux drivers.

The Device driver manager would also be able to tell the power management sub-system which methods are used to power-manage the device. Eg. PCI or ACPI.

5 ACPI Power Management

The Global System States defined by ACPI are illustrated in Figure 5. G0 is the working state, G1 is sleeping, G2 is soft-off and G3 is mechanical off. The “Legacy” state illustrates where the system is not in ACPI mode.

5.1 P-states

In the context of G0 – Global Working State, and C0 – CPU Executing State, P-states (Performance states) are available to reduce power of the running processor. P-states simultaneously modulate both the MHz and the voltage. As power varies by voltage squared, P-states are extremely effective at saving power.

While P-states are extremely important, the `cpufreq` sub-system handles P-states on a number of different platforms, and the topic is best addressed in that larger context.

5.2 Throttling

In the context of the G0-Working, C0-Executing state, Throttling states are defined to modulate the frequency of the running processor.

Power varies (almost) directly with MHz, so when the MHz is cut if half, so is the power. Unfortunately, so is the performance.

Linux currently uses Throttling only in response to thermal events where the processor is too hot. However, in the future, Linux could add throttling when the processor is already in the lowest P-state to save additional power.

Note that most processors also include a backup Thermal Monitor throttling mechanism in hardware, set with higher temperature thresholds than ACPI throttling. Most processors also have in hardware an thermal emergency shutdown mechanism.

5.3 C-states

In the context of G0 Working system state, C-state (CPU-state) C0 is used to refer to the executing state. Higher number C-states are entered to save successively more power when the processor is idle. No instructions are executed when in C1, C2, or C3.

ACPI replaces the default idle loop so it can enter C1, C2 or C3. The deeper the C-state, the more power savings, but the higher the latency to enter/exit the C-state. You can observe the C-states supported by the system and the success at using them in `/proc/acpi/processor/CPU0/power`

C1 is included in every processor and has negligible latency. C1 is implemented with the HALT or MONITOR/MWAIT instructions. Any interrupt will automatically wake the processor from C1.

C2 has higher latency (though always under 100 usec) and higher power savings than C1. It is entered through writes to ACPI registers and exits automatically with any interrupt.

C3 has higher latency (though always under 1000 usec) and higher power savings than C2. It is entered through writes to ACPI registers and exits automatically with any interrupt or bus master activity. The processor does not snoop its cache when in C3, which is why bus-master (DMA) activity will wake it up. Linux sees several implementation issues with C3 today:

1. C3 is enabled even if the latency is up to 1000 usec. This compares with the Linux 2.6 clock tick rate of 1000Hz = 1ms = 1000usec. So when a clock tick causes C3 to exit, it may take all the way to the next clock tick to execute the next kernel instruction. So the benefit of C3 is lost because the system effectively pays C3 latency and gets negligible C3 residency to save power.

2. Some devices do not tolerate the DMA latency introduced by C3. Their device buffers underrun or overflow. This is cur-

rently an issue with the ipw2100 WLAN NIC.

3. Some platforms can lie about C3 latency and transparently put the system into a higher latency C4 when we ask for C3—particularly when running on batteries.

4. Many processors halt their local APIC timer (a.k.a. TSC – Timer Stamp Counter) when in C3. You can observe this by watching LOC fall behind IRQ0 in /proc/interrupts.

5. USB makes it virtually impossible to enter C3 because of constant bus master activity. The workaround at the moment is to unplug your USB devices when idle. Longer term, it will take enhancements to the USB sub-system to address this issue. Ie. USB software needs to recognize when devices are present but idle, and reduce the frequency of bus master activity.

Linux decides which C-state to enter on idle based on a promotion/demotion algorithm. The current algorithm measures the residency in the current C-state. If it meets a threshold the processor is promoted to the deeper C-state on re-entrance into idle. If it was too short, then the processor is demoted to a lower-numbered C-state.

Unfortunately, the demotion rules are overly simplistic, as Linux tracks only its previous success at being idle, and doesn't yet account for the load on the system.

Support for deeper C-states via the _CST method is currently in prototype. Hopefully this method will also give the OS more accurate data than the FADT about the latency associated with C3. If it does not, then we may need to consider discarding the table-provided latencies and measuring the actual latency at boot time.

5.4 Sleep States

ACPI names sleeps states S0 – S5. S0 is the non-sleep state, synonymous with G0. S1 is standby, it halts the processor and turns off the display. Of course turning off the display on an idle system saves the same amount of power without taking the system off line, so S1 isn't worth much. S2 is deprecated. S3 is suspend to RAM. S4 is hibernate to disk. S5 is soft-power off, AKA G2.

Sleep states are unreliable enough on Linux today that they're best considered "experimental." Suspend/Resume suffers from (at least) two systematic problems:

- `_init()` and `_initdata()` on items that may be referenced after boot, say, during resume, is a bad idea.

- PCI configuration space is not uniformly saved and restored either for devices or for PCI bridges. This can be observed by using `lspci` before and after a suspend/resume cycle. Sometimes `setpci` can be used to repair this damage from user-space.

5.5 Device States

Not shown on the diagram, ACPI defines power saving states for devices: D0 – D3. D0 is on, D3 is off, D1 and D2 are intermediate. Higher device states have

1. more power savings,

2. less device context saved by hardware,

3. more device driver state restoring,

4. higher restore latency.

ACPI defines semantics for each device state in each device class. In practice, D1 and D2 are often optional - as many devices support only on and off either because they are low-latency, or because they are simple.

Linux-2.6 includes an updated device driver model to accommodate power management.[14] This model is highly compatible with PCI and ACPI. However, this vision is not yet fully realized. To do so, Linux needs a global power policy manager.

5.6 Wanted: Generic Linux Run-time Power Policy Manager

PCI device drivers today call `pci_set_power_state()` to enter D-states. This uses the power management capabilities in the PCI power management specification.

The ACPI DSDT supplies methods for ACPI enumerated devices to access ACPI D-states. However, no driver calls into ACPI to enter D-states today.[15]

Drivers shouldn't have to care if they are power managed by PCI or by ACPI. Drivers should be able to up-call to a generic run-time power policy manager. That manager should know about calling the PCI layer or the ACPI layer as appropriate.

The power manager should also put those requests in the context of user-specified power policy. Eg. Does the user want maximum performance, or maximum battery life? Currently there is no method to specify the detailed policy, and the kernel wouldn't know how to handle it anyway.

In a related point, it appears that devices currently only suspend upon system suspend. This is probably not the path to industry leading battery life.

Device drivers should recognize when their device has gone idle. They should invoke a suspend up-call to a power manager layer which will decide if it really is a good idea to grant that request now, and if so, how. In this case by calling the PCI or ACPI layer as appropriate.

6 ACPI as seen by bugzilla

Over the last year the ACPI developers have made heavy use of bugzilla[16] to help prioritize and track 460 bugs. 300 bugs are closed or resolved, 160 are open. [17]

We cc: `acpi-bugzilla@lists.sourceforge.net` on these bugs, and we encourage the community to add that alias to ACPI-specific bugs in other bugzillas so that the team can help out wherever the problems are found.

We haven't really used the bugzilla priority field. Instead we've split the bugs into categories and have addressed the configuration issues first. This explains why most of the interrupt bugs are resolved, and most of the suspend/resume bugs are unresolved.

We've seen an incoming bug rate of 10-bugs/week for many months, but the new reports favor the power management features over configuration, so we're hopeful that the torrent of configuration issues is behind us.

[14] Patrick Mochel, Linux Kernel Power Management, OLS 2003.

[15] Actually, the ACPI hot-plug driver invokes D-states, but that is the only exception.

[16] `http://bugzilla.kernel.org/`

[17] The resolved state indicates that a patch is available for testing, but that it is not yet checked into the kernel.org kernel.

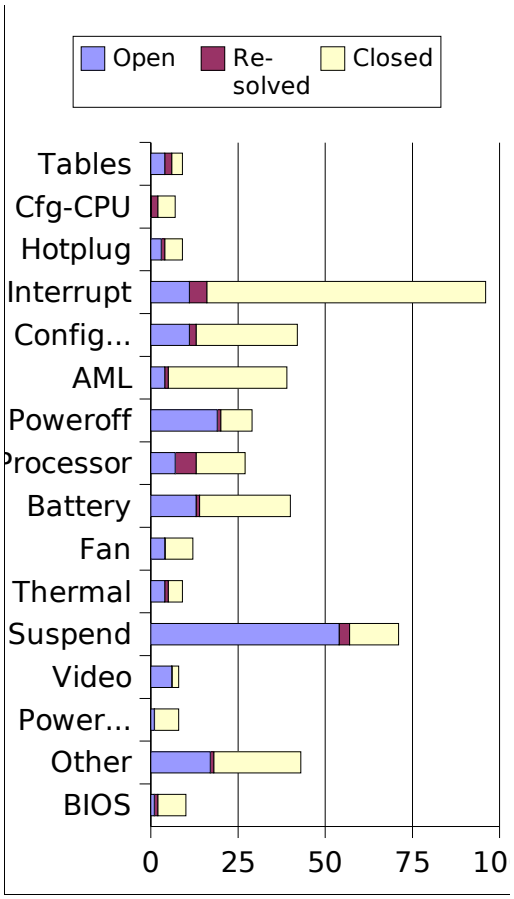


Figure 6: ACPI bug profile

7 Future Work

7.1 Linux 2.4

Going forward, I expect to back-port only critical configuration related fixes to Linux-2.4. For the latest power management code, users need to migrate to Linux-2.6.

7.2 Linux 2.6

Linux-2.6 is a "stable" release, so it is not appropriate to integrate significant new features. However, the power management side of ACPI is widely used in 2.6 and there will be plenty of bug-fixes necessary. The most visible will probably be anything that makes Suspend/Resume work on more platforms.

7.3 Linux 2.7

These feature gaps will not be addressed in Linux 2.6, and so are candidates for Linux 2.7:

- Device enumeration is not abstracted in a generic device driver manager that can shield drivers from knowing if they're enumerated by ACPI, PCI, or other.
- Motherboard devices enumerated by ACPI in the DSDT are ignored, and probed instead via legacy methods. This can lead to resource conflicts.
- Device power states are not abstracted in a generic device power manager that can shield drivers from knowing whether to call ACPI or PCI to handle D-states.
- There is no power policy manager to translate the user-requested power policy into kernel policy.
- No devices invoke ACPI methods to enter D-states.
- Devices do not detect that they are idle and request of a power manager whether they should enter power saving device states.
- There is no MP/SMT coordination of P-states. Today, P-states are disabled on SMP systems. Coordination needs to account for multiple threads and multiple cores per package.
- Coordinate P-states and T-states. Throttling should be used only after the system is put in the lowest P-state.
- Idle states above C1 are disabled on SMP.
- Enable Suspend in PAE mode. [18]

[18] PAE, Physical Address Extended—MMU mode to handle > 4GB RAM—optional on `i386`, always used on `x86_64`.

- Enable Suspend on SMP.
- Tick timer modulation for idle power savings.
- Video control extensions. Video is a large power consumer. The ACPI spec Video extensions are currently in prototype.
- Docking Station support is completely absent from Linux.
- ACPI 3.0 features. TBD after the specification is published.

7.4 ACPI 3.0

Although ACPI 3.0 has not yet been published, two ACPI 3.0 tidbits are already in Linux.

- PCI Express table scanning. This is the basic PCI Express support, there will be more coming. Those in the PCI SIG can read all about it in the PCI Express Firmware Specification.
- Several clarifications to the ACPI 2.0b spec resulted directly from open source development, [19] and the text of ACPI 3.0 has been updated accordingly. For example, some subtleties of SCI interrupt configuration and device enumeration.

When the ACPI 3.0 specification is published there will instantly be a multiple additions to the ACPI/Linux feature to-do list.

7.5 Tougher Issues

- Battery Life on Linux is not yet competitive. This single metric is the sum of all the power savings features in the platform, and if any of them are not working properly, it comes out on this bottom line.
- Laptop Hot Keys are used to control things such as video brightness, etc. ACPI does not specify Hot Keys. But when they work in APM mode and don't work in ACPI mode, ACPI gets blamed. There are 4 ways to implement hot keys:
  1. SMI[20] handler, the BIOS handles interrupts from the keys, and controls the device directly. This acts like "hardware" control as the OS doesn't know it is happening. But on many systems this SMI method is disabled as soon as the system transitions into ACPI mode. Thus the complaint "the button works in APM mode, but doesn't work in ACPI mode."

     But ACPI doesn't specify how hot keys work, so in ACPI mode one of the other methods listed here needs to handle the keys.
  2. Keyboard Extension driver, such as `i8k`. Here the keys return scan codes like any other keys on the keyboard, and the keyboard driver needs to understand those scan code. This is independent of ACPI, and generally OEM specific.
  3. OEM-specific ACPI hot key driver. Some OEMs enumerate the hot keys as OEM-specific devices in the ACPI tables. While the device is described in AML, such devices are not described in the ACPI spec so we can't build generic ACPI support for them. The OEM must supply the appropriate hot-key driver since only they know how it is supposed to work.
  4. Platform-specific "ACPI" driver. Today Linux includes Toshiba and

[19] FreeBSD deserves kudos in addition to Linux

[20] SMI, System Management Interrupt; invisible to the OS, handled by the BIOS, generally considered evil.

Asus platform specific extension drivers to ACPI. They do not use portable ACPI compliant methods to recognize and talk to the hot keys, but generally use the methods above.

The correct solution to the the Hot Key issue on Linux will require direct support from the OEMs, either by supplying documentation, or code to the community.

8 Summary

This past year has seen great strides in the configuration aspects of ACPI. Multiple Linux distributors now enable ACPI on multiple architectures.

This sets the foundation for the next era of ACPI on Linux where we can evolve the more advanced ACPI features to meet the expectations of the community.

9 Resources

The ACPI specification is published at `http://www.acpi.info`.

The home page for the Linux ACPI development community is here: `http://acpi.sourceforge.net/` It contains numerous useful pointers, including one to the `acpi-devel` mailing list.

The latest ACPI code can be found against various recent releases in the BitKeeper repositories: `http://linux-acpi.bkbits.net/`

Plain patches are available on `kernel.org`.[21] Note that Andrew Morton currently includes the latest ACPI test tree in the `-mm` patch, so you can test the latest ACPI code combined with other recent updates there.[22]

10 Acknowledgments

Many thanks to the following people whose direct contributions have significantly improved the quality of the ACPI code in the last year: Jesse Barnes, John Belmonte, Dominik Brodowski, Bruno Ducrot, Bjorn Helgaas, Nitin, Kamble, Andi Kleen, Karol Kozimor, Pavel Machek, Andrew Morton, Jun Nakajima, Venkatesh Pallipadi, Nate Lawson, David Shaohua Li, Suresh Siddha, Jes Sorensen, Andrew de Quincey, Arjan van de Ven, Matt Wilcox, and Luming Yu. Thanks also to all the bug submitters, and the enthusiasts on `acpi-devel`.

Special thanks to Intel's Mobile Platforms Group, which created ACPICA, particularly Bob Moore and Andy Grover.

[21] `http://ftp.kernel.org/pub/linux/kernel/people/lenb/acpi/patches/`

[22] `http://ftp.kernel.org/pub/linux/kernel/people/akpm/patches/`

Scaling Linux® to the Extreme

From 64 to 512 Processors

Ray Bryant raybry@sgi.com
John Hawkes hawkes@sgi.com
Jack Steiner steiner@sgi.com
Jesse Barnes jbarnes@sgi.com
Jeremy Higdon jeremy@sgi.com

Silicon Graphics, Inc.

Abstract

In January 2003, SGI announced the SGI® Altix® 3000 family of servers. As announced, the SGI Altix 3000 system supported up to 64 Intel® Itanium® 2 processors and 512 GB of main memory in a single Linux® image. Altix now supports up to 256 processors in a single Linux system, and we have a few early-adopter customers who are running 512 processors in a single Linux system; others are running with as much as 4 terabytes of memory. This paper continues the work reported on in our 2003 OLS paper by describing the changes necessary to get Linux to efficiently run high-performance computing workloads on such large systems.

Introduction

At OLS 2003 [1], we discussed changes to Linux that allowed us to make Linux scale to 64 processors for our high-performance computing (HPC) workloads. Since then, we have continued our scalability work, and we now support up to 256 processors in a single Linux image, and we have a few early-adopter customers who are running 512 processors in a single-system image; other customers are running with as much as 4 terabytes of memory.

As can be imagined, the type of changes necessary to get a single Linux system to scale on a 512 processor system or to support 4 terabytes of memory are of a different nature than those necessary to get Linux to scale up to a 64 processor system, and the majority of this paper will describe such changes.

While much of this work has been done in the context of a Linux 2.4 kernel, Altix is now a supported platform in the Linux 2.6 series (`www.kernel.org` versions of Linux 2.6 boot and run well on many small to moderate sized Altix systems), and our plan is to port many of these changes to Linux 2.6 and propose them as enhancements to the community kernel. While some of these changes will be unique to the Linux kernel for Altix, many of the changes we propose will also improve performance on smaller SMP and NUMA systems, so should be of general interest to the Linux scalability community.

In the rest of this paper, we will first provide a brief review of the SGI Altix 3000 hardware. Next we will describe why we believe that very large single-system image, shared-memory machine can be more effective tools for HPC than similar sized non-shared memory clusters. We will then discuss changes that we made to Linux for Altix in order to make

that system a more effective system for HPC on systems with as many as 512 processors. A second large topic of discussion will be the changes to support high-performance I/O on Altix and some of the hardware underpinnings for that support. We believe that the latter set of problems are general in the sense that they apply to any large scale NUMA system and the solutions we have adopted should be of general interest for this reason.

Even though this paper is focused on the changes that we have made to Linux to effectively support very large Altix platforms, it should be remembered that the total number of such changes is small in relation to the overall size of the Linux kernel and its supporting software. SGI is committed to supporting the Linux community and continues to support Linux for Altix as a member of the Linux family of kernels, and in general to support binary compatibility between Linux for Altix and Linux on other Itanium Processor Family platforms.

In many cases, the scaling changes described in this paper have already been submitted to the community for consideration for inclusion in Linux 2.6. In other cases, the changes are under evaluation to determine if they need to be added to Linux 2.6, or whether they are fixes for problems in Linux 2.4.21 (the current product base for Linux for Altix) that are no longer present in Linux 2.6.

Finally, this paper contains forward-looking statements regarding SGI® technologies and third-party technologies that are subject to risks and uncertainties. The reader is cautioned not to rely unduly on these forward-looking statements, which are not a guarantee of future or current performance, nor are they a guarantee that features described herein will or will not be available in future SGI products.

The SGI Altix Hardware

This section is condensed from [1]; the reader should refer to that paper for additional details.

An Altix system consists of a configurable number of rack-mounted units, each of which SGI refers to as a *brick*. The most common type of brick is the C-brick (or compute brick). A fully configured C-brick consists of two separate dual-processor Intel Itanium 2 systems, each of which is a bus-connected multiprocessor or *node*.

In addition to the two processors on the bus, there is also a SHUB chip on each bus. The SHUB is a proprietary ASIC that (1) acts as a memory controller for the local memory, (2) provides the interface to the interconnection network, (3) manages the global cache coherency protocol, and (4) some other functions as discussed in [1].

Memory accesses in an Altix system are either local (i.e., the reference is to memory in the same node as the processor) or remote. The SHUB detects whether a reference is local, in which case it directs the request to the memory on the node, or remote, in which case it forwards the request across the interconnection network to the SHUB chip where the memory reference will be serviced.

Local memory references have lower latency; the Altix system is thus a NUMA (non-uniform memory access) system. The ratio of remote to local memory access times on an Altix system varies from 1.9 to 3.5, depending on the size of the system and the relative locations of the processor and memory module involved in the transfer.

The cache-coherency policy in the Altix system can be divided into two levels: *local* and *global*. The local cache-coherency protocol is defined by the processors on the local

bus and is used to maintain cache-coherency between the Itanium processors on the bus. The global cache-coherency protocol is implemented by the SHUB chip. The global protocol is directory-based and is a refinement of the protocol originally developed for DASH [2].

The Altix system interconnection network uses routing bricks to provide connectivity in system sizes larger than 16 processors. In systems with 128 or more processors a second layer of routing bricks is used to forward requests among subgroups of 32 processors each. The routing topology is a fat-tree topology with additional "express" links being inserted to improve performance.

Why Big SSI?

In this section we discuss the rationale for building such a large single-system image (SSI) box as an Altix system with 512 CPU's and (potentially) several TB of main memory:

(1) Shared memory systems are more flexible and easier to manage than a cluster. One can simulate message passing on shared memory, but not the other way around. Software for cluster management and system maintenance exists, but can be expensive or complex to use.

(2) Shared memory style programming is generally simpler and more easily understood than message passing. Debugging of code is often simpler on a SSI system than on a cluster.

(3) It is generally easier to port or write codes from scratch using the shared memory paradigm. Additionally it is often possible to simply ignore large sections of the code (e.g. those devoted to data input and output) and only parallelize the part that matters.

(4) A shared memory system supports easier load balancing within a computation. The mapping of grid points to a node determines the computational load on the node. Some grid points may be located near more rapidly changing parts of computation, resulting in higher computational load. Balancing this over time requires moving grid points from node to node in a cluster, where in a shared memory system such re-balancing is typically simpler.

(5) Access to large global data sets is simplified. Often, the parallel computation depends on a large data set describing, for example, the precise dimensions and characteristics of the physical object that is being modeled. This data set can be too large to fit into the node memories available on a clustered machine, but it can readily be loaded into memory on a large shared memory machine.

(6) Not everything fits into the cluster model. While many production codes have been converted to message passing, the overall computation may still contain one or more phases that are better performed using a large shared memory system. Or, there may be a subset of users of the system who would prefer a shared memory paradigm to a message passing one. This can be a particularly important consideration in large data-center environments.

Kernel Changes

In this section we describe the most significant kernel problems we have encountered in running Linux on a 512 processor Altix system.

Cache line and TLB Conflicts

Cache line conflicts occur in every cache-coherent multiprocessor system, to one extent or another, and whether or not the conflict exhibits itself as a performance problem is dependent on the rate at which the conflict occurs and the time required by the hardware to resolve

the conflict. The latter time is typically proportional to the number of processors involved in the conflict. On Altix systems with 256 processors or more, we have encountered some cache line conflicts that can effectively halt forward progress of the machine. Typically, these conflicts involve global variables that are updated at each timer tick (or faster) by every processor in the system.

One example of this kind of problem is the default kernel profiler. When we first enabled the default kernel profiler on a 512 CPU system, the system would not boot. The reason was that once per timer tick, each processor in the system was trying to update the profiler bin corresponding to the CPU idle routine. A work around to this problem was to initialize `prof_cpu_mask` to `CPU_MASK_NONE` instead of the default. This disables profiling on all processors until the user sets the `prof_cpu_mask`.

Another example of this kind of problem was when we imported some timer code from Red Hat® AS 3.0. The timer code included a global variable that was used to account for differences between HZ (typically a power of 2) and the number of microseconds in a second (nominally 1,000,000). This global variable was updated by each processor on each timer tick. The result was that on Altix systems larger than about 384 processors, forward progress could not be made with this version of the code. To fix this problem, we made this global variable a per processor variable. The result was that the adjustment for the difference between HZ and microseconds is done on a per processor rather than on a global basis, and now the system will boot.

Still other cache line conflicts were remedied by identifying cases of false cache line sharing i.e., those cache lines that inadvertently contain a field that is frequently written by one CPU and another field (or fields) that are frequently read by other CPUs.

Another significant bottleneck is the ia64 `do_gettimeofday()` with its use of `cmpxchg`. That operation is expensive on most architectures, and concurrent `cmpxchg` operations on a common memory location scale worse than concurrent simple writes from multiple CPUs. On Altix, four concurrent user `gettimeofday()` system calls complete in almost an order of magnitude more time than a single `gettimeofday()`; eight are 20 times slower than one; and the scaling deteriorates nonlinearly to the point where 32 concurrent system calls is 100 times slower than one. At the present time, we are still exploring a way to improve this scaling problem in Linux 2.6 for Altix.

While moving data to per-processor storage is often a solution to the kind of scaling problems we have discussed here, it is not a panacea, particularly as the number of processors becomes large. Often, the system will want to inspect some data item in the per-processor storage of each processor in the system. For small numbers of processors this is not a problem. But when there are hundreds of processors involved, such loops can cause a TLB miss each time through the loop as well as a couple of cache-line misses, with the result that the loop may run quite slowly. (A TLB miss is caused because the per-processor storage areas are typically isolated from one another in the kernel's virtual address space.)

If such loops turn out to be bottlenecks, then what one must often do is to move the fields that such loops inspect out of the per-processor storage areas, and move them into a global static array with one entry per CPU.

An example of this kind of problem in Linux 2.6 for Altix is the current allocation scheme of the per-CPU run queue structures. Each

per-CPU structure on an Altix system requires a unique TLB to address it, and each structure begins at the same virtual offset in a page, which for a virtually indexed cache means that the same fields will collide at the same index. Thus, a CPU scheduler that wishes to do a quick peek at every other CPU's `nr_running` or `cpu_load` will not only suffer a TLB miss on every access, but will also likely suffer a cache miss because these same virtual offsets will collide in the cache. Cache coloring of these addresses would be one way to solve this problem; we are still exploring ways to fix this problem in Linux 2.6 for Altix.

Lock Conflicts

A cousin of cache line conflicts are the lock conflicts. Indeed, the root mechanism of the lock bottleneck is a cache line conflict. For a `spinlock_t` the conflict is the `cmpxchg` operation on the word that signifies whether or not the lock is owned. For a `rwlock_t` the conflict is the cmpxchg or fetch-and-add operation on the count of the number of readers or the bit signifying whether or not the lock is owned exclusively by a writer. For a `seqlock_t` the conflict is the increment of the sequence number.

For some lock conflicts, such as the `rcu_ctrlblk.mutex`, the remedy is to make the spinlock more fine-grained, e.g., by making it hierarchical or per-CPU. For other lock conflicts, the most effective remedy is to reduce the use of the lock.

The O(1) CPU scheduler replaced the global `runqueue_lock` with per-CPU run queue locks, and replaced the global run queue with per-CPU run queues. While this did substantially decrease the CPU scheduling bottleneck for CPU counts in the 8 to 32 range, additional effort has been necessary to remedy additional bottlenecks that appear with even large configurations.

For example, we discovered that at 256 processors and above, we encountered a live lock early in system boot because hundreds of idle CPUs are load-balancing and are racing in contention on one or a few busy CPUs. The contention is so severe that the busy CPU's scheduler cannot itself acquire its own run queue lock, and thus the system live locks.

A remedy we applied in our Altix 2.4-based kernel was to introduce a progressively longer back off between successive load-balancing attempts, if the load-balancing CPU continues to be unsuccessful in finding a task to pull-migrate. Perhaps all the busiest CPU's tasks are pinned to that CPU, or perhaps all the tasks are still cache-hot. Regardless of the reason, a load-balancing failure results in that CPU delaying the next load-balance attempt by another incremental increase in time. This algorithm effectively solved the live lock, as well as improved other high-contention conflicts on a busiest CPU's run queue lock (e.g., always finding pinned tasks that can never be migrated).

This load-balance back off algorithm did not get accepted into the early 2.6 kernels. The latest 2.6.7 CPU scheduler, as developed by Nick Piggin, incorporates a similar back off algorithm. However, this algorithm (at least as it appears in 2.6.7-rc2) continues to cause a boot-time live lock at 512 processors on Altix so we are continuing to investigate this matter.

Page Cache

Managing the page cache in Altix has been a challenging problem. The reason is that while a large Altix system may have a lot of memory, each node in the system only has a relatively small fraction of that memory available as local memory. For example, on a 512 CPU sys-

tem, if the entire system has 512 GB of memory, each node on the system has only 2 GB of local memory; less than 0.4% of the available memory on the system is local. When you consider that it is quite common on such systems to deal with files that are tens of GB in size, it is easy to understand how the page cache could consume all of the memory on several nodes in the system just doing normal, buffered-file I/O.

Stated another way, this is the challenge of a large NUMA system: all memory is addressable, but only a tiny fraction of that memory is local. Users of NUMA systems need to place their most frequently accessed data in local memory; this is crucial to obtain the maximum performance possible from the system. Typically this is done by allocating pages on a first-touch basis; that is, we attempt to allocate a page on the node where it is first referenced. If all of the local memory on a node is consumed by the page cache, then these local storage allocations will spill over to other (remote) nodes, the result being a potentially significant impact on program performance.

Similarly, it is important that the amount of free memory be balanced across idle nodes in the system. An imbalance could lead to some components of a parallel computation running slower than others because not all components of the computation were able to allocate their memory entirely out of local storage. Since the overall speed of parallel computation is determined by the execution of its slowest component, the performance of the entire application can be impacted by a non-local storage allocation on only a few nodes.

One might think that `bdflush` or `kupdated` (in a Linux 2.4 system) would be responsible for cleaning up unused page-cache pages. As the OLS reader knows, these daemons are responsible not for deallocating page-cache pages, but cleaning them. It is the swap daemon `kswapd` that is responsible for causing page-cache pages to be deallocated. However, in many situations we have encountered, even though multiple nodes of the system would be completely out of local memory, there would still be lots of free memory elsewhere in the system. As a result, `kswapd` will never start. Once the system gets into such a state, the local memory on those nodes can remain allocated entirely to page-cache pages for very long stretches of time since as far as the kernel is concerned there is no memory "pressure". To get around this problem, particularly for benchmarking studies, users have often resorted to programs that allocate and touch all of the memory on the system, thus causing `kswapd` to wake up and free unneeded buffer cache pages.

We have dealt with this problem in a number of ways, but the first approach was to change `page_cache_alloc()` so that instead of allocating the page on the local node, we spread allocations across all nodes in the system. To do this, we added a new GFP flag: `GFP_ROUND_ROBIN` and a new procedure: `alloc_pages_round_robin()`. `alloc_pages_round_robin()` maintains a counter in per-CPU storage; the counter is incremented on each call to `page_cache_alloc()`. The value of the counter, modulus the number of nodes in the system, is used to select the `zonelist` passed to `__alloc_pages()`. Like other NUMA implementations, in Linux for Altix there is a `zonelist` for each node, and the `zonelist`s are sorted in nearest neighbor order with the `zone` for the local node as the first entry of the `zonelist`. The result is that each time `page_cache_alloc()` is called, the returned page is allocated on the next node in sequence, or as close as possible to that node.

The rationale for allocating page-cache pages

in this way is that while pages are local resources, the page cache is a global resource, usable by all processes on the system. Thus, even if a process is bound to a particular node, in general it does not make sense to allocate page-cache pages just on that node, since some other process in the system may be reading that same file and hence sharing the pages. So instead of flooding the current node with the page-cache pages for files that processes on that node have opened, we "tax" every node in the system with a fraction of the page-cache pages. In this way, we try to conserve a scarce resource (local memory) by spreading page-cache allocations over all nodes in the system.

However, even this step was not enough to keep local storage usage balanced among nodes in the system. After reading a 10 GB file, for example, we found that the node where the reading process was running would have up to 40,000 pages more storage allocated than other nodes in the system. It turned out the reason for this was that buffer heads for the read operation were being allocated locally. To solve this problem in our Linux 2.4.21 kernel for Altix, we modified `kmem_cache_grow()` so that it would pass the `GFP_ROUND_ROBIN` flag to `kmem_getpages()` with the result that the slab caches on our systems are now also allocated out of round-robin storage. Of course, this is not a perfect solution, since there are situations where it makes perfect sense to allocate a slab cache entry locally; but this was an expedient solution appropriate for our product. For Linux 2.6 for Altix we would like to see the slab allocator be made NUMA aware. (Manfred Spraul has created some patches to do this and we are currently evaluating these changes.)

The previous two changes solved many of the cases where a local storage could be exhausted by allocation of page-cache pages. However, they still did not solve the problem of local allocations spilling off node, particularly in those cases where storage allocation was tight across the entire system. In such situations, the system would often start running the synchronous swapping code even though most (if not all) of the page-cache pages on the system were clean and unreferenced outside of the page-cache. With the very-large memory sizes typical of our larger Altix customers, entering the synchronous swapping code needs to be avoided if at all possible since this tends to freeze the system for 10s of seconds. Additionally, the round robin allocation fixes did not solve the problem of poor and unrepeatable performance on benchmarks due to the existence of significant amounts of page-cache storage left over from previous executions.

To solve these problems, we introduced a routine called `toss_buffer_cache_pages_node()` (referred to here as `toss()`, for brevity). In a related change, we made the active and inactive lists per node rather than global. `toss()` first scans the inactive list (on a particular node) looking for idle page-cache pages to release back to the free page pool. If not enough such pages are found on the inactive list, then the active list is also scanned. Finally, if `toss()` has not called `shrink_slab_caches()` recently, that routine is also invoked in order to more aggressively free unused slab-cache entries. `toss()` was patterned after the main loop of `shrink_caches()` except that it would never call `swap_out()` and if it encountered a page that didn't look to be easily free able, it would just skip that page and go on to the next page.

A call to `toss()` was added in `__alloc_pages()` in such a way that if allocation on the current node fails, then before trying to allocate from some other node (i. e. spilling to another node), the system will first see if it can free enough page-cache pages from the current node so that the current node alloca-

tion can succeed. In subsequent allocation passes, `toss()` is also called to free page-cache pages on nodes other than the current one. The result of this change is that clean page-cache pages are effectively treated as free memory by the page allocator.

At the same time that the `toss()` code was added, we added a new user command `bcfree` that could be used to free all idle page-cache pages. (On the `__alloc_pages()` path, `toss()` would only try to free 32 pages per node.) The `bcfree` command was intended to be used only for resetting the state of the page cache before running a benchmark, and in lieu of rebooting the system in order to get a clean system state. However, our customers found that this command could be used to reduce the size of the page cache and to avoid situations where large amounts of buffered-file I/O could force the system to begin swapping. Since `bcfree` kills the entire page-cache, however, this was regarded as a substandard solution that could also hurt read performance of cached data and we began looking for another way to solve this "BIGIO" problem.

Just to be specific, the BIGIO problem we were trying to solve was based on the behavior of our Linux 2.4.21 kernel for Altix. A customer reported that on a 256 GB Altix system, if 200 GB were allocated and 50 GB free, that if the user program then tried to write 100 GB of data out to disk, the system would start to swap, and then in many cases fill up the swap space. At that point our Out-of-memory (OOM) killer would wake up and kill the user program! (See the next section for discussion of our OOM killer changes.)

Initially we were able to work around this problem by increasing the amount of swap space on the system. Our experiments showed that with an amount of swap space equal to one-quarter the main memory size, the 256 GB example discussed above would continue to completion without the OOM killer being invoked. I/O performance during this phase was typically one-half of what the hardware could deliver, since two I/O operations often had to be completed: one to read the data in from the swap device, and one to write the data to the output file. Additionally, while the swap scan was active, the system was very sluggish. These problems led us to search for another solution.

Eventually what we developed is an aggressive method of trimming the page cache when it started to grow too big. This solution involved several steps:

(1) We first added a new page list, the `reclaim_list`. This increased the size of `struct page` by another 16 bytes. On our system, `struct page` is allocated on cache-aligned boundaries anyway, so this really did not cause an increase in storage, since the current `struct page` size was less than 112 bytes. Pages were added to the reclaim list when they were inserted into the page cache. The reclaim list is per node, with per node locking. Pages were removed from the reclaim list when they were no longer reclaimable; that is, they were removed from the reclaim list when they were marked as dirty due to buffer file-I/O or when they were mapped into an address space.

(2) We rewrote `toss()` to scan the reclaim list instead of the inactive and active lists. Herein we will refer to the new version of `toss()` as `toss_fast()`.

(3) We introduced a variant of `page_cache_alloc()` called `page_cache_alloc_limited()`. Associated with this new routine were two control variables settable via `sysctl()`: `page_cache_limit` and `page_cache_limit_threshold`.

(4) We modified the `generic_file_write()` path to call `page_cache_alloc_limited()` instead of `page_cache_alloc()`. `page_cache_alloc_limited()` examines the size of the page cache. If the total amount of free memory in the system is less than `page_cache_limit_threshold` and the size of the page cache is larger than `page_cache_limit`, then `page_cache_alloc_limited()` calls `page_cache_reduce()` to free enough page-cache pages on the system to bring the page cache size down below `page_cache_limit`. If this succeeds, then `page_cache_alloc_limited()` calls `page_cache_alloc` to allocate the page. If not, then we wakeup `bdflush` and the current thread is put to sleep for 30ms (a tunable parameter)

The rationale for the `reclaim_list` and `toss_fast()` was that when we needed to trim the page cache, practically all pages in the system would typically be on the inactive list. The existing `toss()` routine scanned the inactive list and thus was too slow to call from `generic_file_write`. Moreover, most of the pages on the inactive list were not reclaimable anyway. Most of the pages on the `reclaim_list` are reclaimable. As a result `toss_fast()` runs much faster and is more efficient at releasing idle page-cache pages than the old routine.

The rationale for the `page_cache_limit_threshold` in addition to the `page_cache_limit` is that if there is lots of free memory then there is no reason to trim the page cache. One might think that because we only trim the page cache on the file write path that this approach would still let the page cache to grow arbitrarily due to file reads. Unfortunately, this is not the case, since the Linux kernel in normal multiuser operation is constantly writing something to the disk. So, a page cache limit enforced at file write time is also an effective limit on the size of the page cache due to file reads.

Finally, the rationale for delaying the calling task when `page_cache_reduce()` fails is that we do not want the system to start swapping to make space for new buffered I/O pages, since that will reduce I/O bandwidth by as much as one-half anyway, as well as take a lot of CPU time to figure out which pages to swap out. So it is better to reduce the I/O bandwidth directly, by limiting the rate of requested I/O, instead of allowing that I/O to proceed at rate that causes the system to be overrun by page-cache pages.

Thus far, we have had good experience with this algorithm. File I/O rates are not substantially reduced from what the hardware can provide, the system does not start swapping, and the system remains responsive and usable during the period of time when the BIGIO is running.

Of course, this entire discussion is specific to Linux 2.4.21. For Linux 2.6, we have plans to evaluate whether this is a problem in the system at all. In particular, we want to see if an appropriate setting for `vm_swappiness` to zero can eliminate the “BIGIO causes swapping” problem. We also are interested in evaluating the recent set of VM patches that Nick Piggin [6] has assembled to see if they eliminate this problem for systems of the size of a large Altix.

VM and Memory Allocation Fixes

In addition to the page-cache changes described in the last section, we have made a number of smaller changes related to virtual memory and paging performance.

One set of such changes increased the parallelism of page-fault handling for anonymous

pages in multi-threaded applications. These applications allocate space using routines that eventually call `mmap()`; the result is that when the application touches the data area for the first time, it causes a minor page fault. These faults are serviced while holding the address space's `page_table_lock`. If the address space is large and there are a large number of threads executing in the address space, this spinlock can be an initialization-time bottleneck for the application. Examination of the `handle_mm_fault()` path for this case shows that the `page_table_lock` is acquired unconditionally but then released as soon as we have determined that this is a not-present fault for an anonymous page. So, we reordered the code checks in `handle_mm_fault()` to determine in advance whether or not this was the case we were in, and if so, to skip acquiring the lock altogether.

The second place the `page_table_lock` was used on this path was in `do_anonymous_page()`. Here, the lock was re-acquired to make sure that the process of allocating a page frame and filling in the pte is atomic. On Itanium, stores to page-table entries are normal stores (that is, the `set_pte` macro evaluates to a simple store). Thus, we can use `cmpxchg` to update the pte and make sure that only one thread allocates the page and fills in the pte. The compare and exchange effectively lets us lock on each individual pte. So, for Altix, we have been able to completely eliminate the `page_table_lock` from this particular page-fault path.

The performance improvement from this change is shown in Figure 1. Here we show the time required to initially touch 96 GB of data. As additional processors are added to the problem, the time required for both the baseline-Linux and Linux for Altix versions decrease until around 16 processors. At that point the `page_table_lock` starts to become a significant bottleneck. For the largest number of processors, even the time for the Linux for Altix case is starting to increase again. We believe that this is due to contention for the address space's `mmap` semaphore.

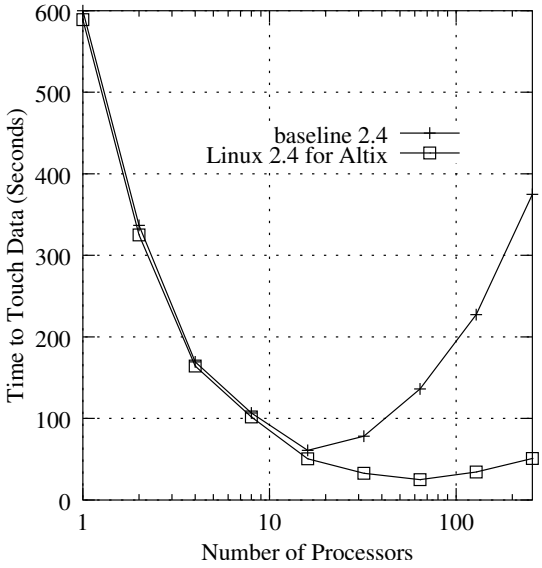


Figure 1: *Time to initially touch 96 GB of data.*

This is particularly important for HPC applications since OpenMP™[5], a common parallel programming model for FORTRAN, is implemented using a single address space, multiple-thread programming model. The optimization described here is one of the reasons that Altix has recently set new performance records for the SPEC® SPEComp® L2001 benchmark [7].

While the above measurements were taken using Linux 2.4.21 for Altix, a similar problem exists in Linux 2.6. For many other architectures, this same kind of change can be made; i386 is one of the exceptions to this statement. We are planning on porting our Linux 2.4.21 based changes to Linux 2.6 and submitting the changes to the Linux community for inclusion in Linux 2.6. This may require moving part of `do_anonymous_page()` to architecture

dependent code to allow for the fact that not all architectures can use the compare and exchange approach to eliminate the use of the `page_table_lock` in `do_anonymous_page()`. However, the performance improvement shown in Figure 1 is significant for Altix so we would we would like to explore some way of incorporating this code into the mainline kernel.

We have encountered similar scalability limitations for other kinds of page-fault behavior. Figure 2 shows the number of page faults per second of wall clock time measured for multiple processes running simultaneously and faulting in a 1 GB `/dev/zero` mapping. Unlike the previous case described here, in this case each process has its own private mapping. (Here the number of processes is equal to the number of CPUs.) The dramatic difference between the baseline 2.4 and 2.6 cases and Linux for Altix is due to elimination of a lock in the super block for `/dev/zero`.

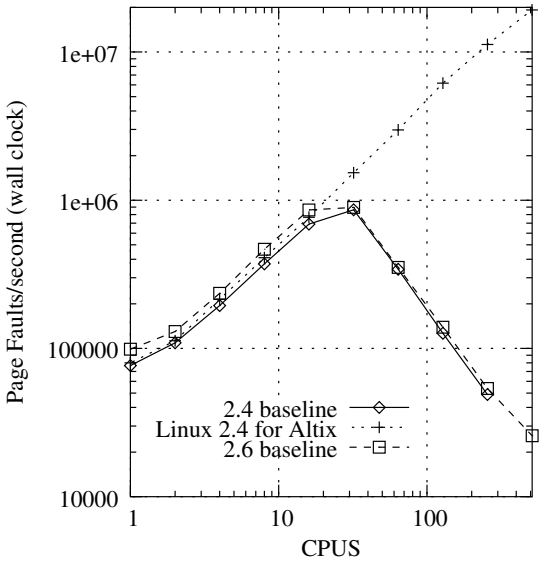


Figure 2: *Page Faults per Second of Wall Clock Time.*

The lock in the super block protects two counts: One count limits the maximum number of `/dev/zero` mappings to 2^{63}; the second count limits the number of pages assigned to a `/dev/zero` mapping to 2^{63}. Neither one of these counts is particularly useful for a `/dev/zero` mapping. We eliminated this lock and obtained a dramatic performance improvement for this micro-benchmark (at 512 CPUs the improvement was in excess of 800x). This optimization is important in decreasing startup time for large message-passing applications on the Altix system.

A related change is to distribute the count of pages in the page cache from a single global variable to a per node variable. Because every processor in the system needs to update the page-cache count when adding or removing pages from the page cache, contention for the cache line containing this global variable becomes significant. We changed this global count to a per-node count. When a page is inserted into (or removed from) the page cache, we update the page cache-count on the same node as the page itself. When we need the total number of pages in the page cache (for example if someone reads `/proc/meminfo`) we run a loop that sums the per node counts. However, since the latter operation is much less frequent than insertions and deletions from the page cache, this optimization is an overall performance improvement.

Another change we have made in the VM subsystem is in the out-of-memory (OOM) killer for Altix. In Linux 2.4.21, the OOM killer is called from the top of memory-free and swap-out call chain. `oom_kill()` is called from `try_to_free_pages_zone()` when calls to `shrink_caches()` at memory priority levels 6 through 0 have all failed. Inside `oom_kill()` a number of checks are performed, and if any of these checks succeed, the system is declared to not be out-of-memory. One of those checks is "if it has been more than 5 seconds since `oom_kill()` was last called, then we are not

OOM." On a large-memory Altix system, it can easily take much longer than that to complete the necessary calls to `shrink_caches()`. The result is that an Altix system never goes OOM in spite of the fact that swap space is full and there is no memory to be allocated.

It seemed to us that part of the problem here is the amount of time it can take for a swap full condition (readily detectable in `try_to_swap_out()` to bubble all the way up to the top level in `try_to_free_pages_zone()`, especially on a large memory machine. To solve this problem on Altix, we decided to drive the OOM killer directly off of detection of swap-space-full condition provided that the system also continues to try to swap out additional pages. A count of the number of successful swaps and unsuccessful swap attempts is maintained in `try_to_swap_out()`. If, in a 10 second interval, the number of successful swap outs is less than one percent of the number of attempted swap outs, and the total number of swap out attempts exceeds a specified threshold, then `try_to_swap_out()`) will directly wake the OOM killer thread (also new in our implementation). This thread will wait another 10 seconds, and if the out-of-swap condition persists, it will invoke `oom_kill()` to select a victim and kill it. The OOM killer thread will repeat this sleep and kill cycle until it appears that swap space is no longer full or the number of attempts to swap out new pages (since the thread went to sleep) falls below the threshold.

In our experience, this has made invocation of the OOM killer much more reliable than it was before, at least on Altix. Once again, this implementation was for Linux 2.4.21; we are in the process of evaluating this problem and the associated fix on Linux 2.6 at the present time.

Another fix we have made to the VM system in Linux 2.4.21 for Altix is in handling of HUGETLB pages. The existing implementation in Linux 2.4.21 allocates HUGETLB pages to an address space at `mmap()` time (see `hugetlb_prefault()`); it also zeroes the pages at this time. This processing is done by the thread that makes the `mmap()` call. In particular, this means that zeroing of the allocated HUGETLB pages is done by a single processor. On a machine with 4 TB of memory and with as much memory allocated to HUGETLB pages as possible, our measurements have shown that it can take as long as 5,000 seconds to allocate and zero all available HUGETLB pages. Worse yet, the thread that does this operation holds the address space's `mmap_sem` and the `page_table_lock` for the entire 5,000 seconds. Unfortunately, many commands that query system state (such as `ps` and `w`) also wish to acquire one of these locks. The result is that the system appears to be hung for the entire 5,000 seconds.

We solved this problem on Altix by changing the implementation of HUGETLB page allocation from *prefault* to *allocate on fault*. Many others have created similar patches; our patch was unique in that it also allowed zeroing of pages to occur in parallel if the HUGETLB page faults occurred on different processors. This was crucial to allow a large HUGETLB page region to be faulted into an address space in parallel, using as many processors as possible. For example, we have observed speedups of 25x using 16 processors to touch O(100 GB) of HUGETLB pages. (The speedup is super linear because if you use just one processor it has to zero many remote pages, whereas if you use more processors, at least some of the pages you are zeroing are local or on nearby nodes.) Assuming we can achieve the same kind of speedup on a 4 TB system, we would reduce the 5,000 second time stated above to 200 seconds.

Recently, we have worked with Kenneth Chen

to get a similar set of changes proposed for Linux 2.6 [3]. Once this set of changes is accepted into the mainline this particular problem will be solved for Linux 2.6. These changes are also necessary for Andi Kleen's NUMA placement algorithms [4] to apply to HUGETLB pages, since otherwise pages are placed at `hugetlb_prefault()` time.

A final set of changes is related to large kernel tables. As previously mentioned, on an Altix system with 512 processors, less than 0.4% of the available memory is local. Certain tables in the Linux kernel are sized to be on the order of one percent of available memory. (An example of this is the TCP/IP hash table.) Allocating a table of this size can use all of the local memory on a node, resulting in exactly the kind of storage-allocation imbalance we developed the page-cache changes to solve. To avoid this problem, we also implement round-robin allocation of these large tables. Our current technique uses `vm_alloc()` to do this. Unfortunately, this is not portable across all architectures, since certain architectures have limited amounts of space that can be allocated by `vm_alloc()`. Nonetheless, this is a change that we need to make; we are still exploring ways of making this change acceptable to the Linux community.

Once we have solved the initial allocation problem for these tables, there is still the problem of getting them appropriately sized for an Altix system. Clearly if there are 4 TB of main memory, it does not make much sense to allocate a TCP/IP hash table of 40 GB, particularly since the TCP/IP traffic into an Altix system does not increase with memory size the way one might expect it to scale with a traditional Linux server. We have seen cases where system performance is significantly hampered due to lookups in these overly large tables. At the moment, we are still exploring a solution acceptable to the community to solve this particular problem.

I/O Changes for Altix

One of the design goals for the Altix system is that it support standard PCI devices and their associated Linux drivers as much as possible. In this section we discuss the performance improvements built into the Altix hardware and supported through new driver interfaces in Linux that help us to meet this goal with excellent performance even on very large Altix systems.

According to the PCI specification, DMA writes and PIO read responses are strongly ordered. On large NUMA systems, however, DMA writes can take a long time to complete. Since most PIO reads do not imply completion of a previous DMA write, relaxing the ordering rules of DMA writes and PIO read responses can greatly improve system performance.

Another large system issue relates to initiating PIO writes from multiple CPUs. PIO writes from two different CPUs may arrive out of order at a device. The usual way to ensure ordering is through a combination of locking and a PIO read (see Documentation/io_ordering.txt). On large systems, however, doing this read can be very expensive, particularly if it must be ordered with respect to unrelated DMA writes.

Finally, the NUMA nature of large machines make some optimizations obvious and desirable. Many devices use so-called consistent system memory for retrieving commands and storing status information; allocating that memory close to its associated device makes sense.

Making non–dependent PIO reads fast

In its I/O chipsets, SGI chose to relax the ordering between DMAs and PIOs, instead adding

a barrier attribute to certain DMA writes (to consistent PCI allocations on Altix) and to interrupts. This works well with controllers that use DMA writes to indicate command completions (for example a SCSI controller with a response queue, where the response queue is allocated using `pci_alloc_consistent`, so that writes to the response queue have the barrier attribute). When we ported Linux to Altix, this behavior became a problem, because many Linux PCI drivers use PIO read responses to imply a status of a DMA write. For example, on an IDE controller, a bit status register read is performed to find out if a command is complete (command complete status implies that DMA writes of that command's data are completed). As a result, SGI had to implement a rather heavyweight mechanism to guarantee ordering of DMA writes and PIO reads. This mechanism involves doing an explicit flush of DMA write data after each PIO read.

For the cases in which strong ordering of PIO read responses and DMA writes are not necessary, a new API was needed so that drivers could communicate that a given PIO read response could used relaxed ordering with respect to prior DMA writes. The `read_relaxed` API [8] was added early in the 2.6 series for this purpose, and mirrors the normal read routines, which have variants for various sized reads.

The results below show how expensive a normal PIO read transaction can be, especially on a system doing a lot of I/O (and thus DMA).

Type of PIO	Time (ns)
normal PIO read	3875
relaxed PIO read	1299

Table 1: Normal vs. relaxed PIO reads on an idle system

It remains to be seen whether this API will also apply to the newly added RO bit in the PCI-X specification—the author is hopeful! Either way, it does give hardware vendors who want to support Linux some additional flexibility in their design.

Type of PIO	Time (ns)
normal PIO read	4889
relaxed PIO read	1646

Table 2: Normal vs. relaxed PIO reads on a busy system

Ordering posted writes efficiently

On many platforms, PIO writes from different CPUs will not necessarily arrive in order (i.e., they may be intermixed) even when locking is used. Since the platform has no way of knowing whether a given PIO read depends on preceding writes, it has to guarantee that all writes have completed before allowing a read transaction to complete. So performing a read prior to releasing a lock protecting a region doing writes is sufficient to guarantee that the writes arrive in the correct order.

However, performing PIO reads can be an expensive operation, especially if the device is on a distant node. SGI chipset designers foresaw this problem, however, and provided a way to ensure ordering by simply reading a register from the chipset on the local node. When the register indicates that all PIO writes are complete, it means they have arrived at the chipset attached to the device, and so are guaranteed to arrive at the device in the intended order. The SGI sn2 specific portion of the Linux ia64 port (sn2 is the architecture name for Altix in the Linux kernel source tree) provides a small function, `sn_mmiob()` (for memory–mapped I/O barrier, analogous to the `mb()` macro), to do just that. It can be used in place of reads that are intended to deal with posted writes and provides some benefit:

Type of flush	Time (ns)
regular PIO read	5940
relaxed PIO read	2619
`sn_mmiob()`	1610
(local chipset read alone)	399

Table 3: Normal vs. fast flushing of 5 PIO writes

Adding this API to Linux (i.e., making it non-sn2-specific) was discussed some time ago [9], and may need to be raised again, since it does appear to be useful on Altix, and is probably similarly useful on other platforms.

Local allocation of consistent DMA mappings

Consistent DMA mappings are used frequently by drivers to store command and status buffers. They are frequently read and written by the device that owns them, so making sure they can be accessed quickly is important. The table below shows the difference in the number of operations per second that can be achieved using local versus remote allocation of consistent DMA buffers. Local allocations were guaranteed by changing the `pci_alloc_consistent` function so that it calls `alloc_pages_node` using the node closest to the PCI device in question.

Type	I/Os per second
Local consistent buffer	46231
Remote consistent buffer	41295

Table 4: Local vs. remote DMA buffer allocation

Although this change is platform specific, it can be made generic if a `pci_to_node` or `pci_to_nodemask` routine is added to the Linux topology API.

Concluding Remarks

Today, our Linux 2.4.21 kernel for Altix provides a productive platform for our high-performance-computing users who desire to exploit the features of the SGI Altix 3000 hardware. To achieve this goal, we have made a number of changes to our Linux for Altix kernel. We are now in the process of either moving those changes forward to Linux 2.6 for Altix, or of evaluating the Linux 2.6 kernel on Altix in order to determine if these changes are indeed needed at all. Our goal is to develop a version of the Linux 2.6 kernel for Altix that not only supports our HPC customers equally well as our existing Linux 2.4.21 kernel, but also consists as much as possible of community supported code.

References

[1] Ray Bryant and John Hawkes, Linux Scalability for Large NUMA Systems, *Proceedings of the 2003 Ottawa Linux Symposium*, Ottawa, Ontario, Canada, (July 2003).

[2] Daniel Lenoski, James Laudon, Truman Joe, David Nakahira, Luis Stevens, Anoop Gupta, and John Hennesy, The DASH prototype: Logic overhead and performance, *IEEE Transactions on Parallel and Distributed Systems*, 4(1):41-61, January 1993.

[3] Kenneth Chen, "hugetlb demand paging patch part [0/3]," linux-kernel@vger.kernel.org, 2004-04-13 23:17:04, `http://marc.theaimsgroup.com/?l=linux-kernel&m=108189860419356&w=2`

[4] Andi Kleen, "Patch: NUMA API for Linux," linux-kernel@vger.kernel.org,

Tue, 6 Apr 2004 15:33:22 +0200, http://lwn.net/Articles/79100/

[5] http://www.openmp.org

[6] Nick Piggin, "MM patches," http://www.kerneltrap.org/~npiggin/nickvm-267r1m1.gz

[7] http://www.spec.org/omp/results/ompl2001.html

[8] http://linux.bkbits.net:8080/linux-2.5/cset%4040213ca0d3eIznHTPAR_kLCsMZI9VQ?nav=index.html|ChangeSet@-1d

[9] http://www.cs.helsinki.fi/linux/linux-kernel/2002-01/1540.html

Get More Device Drivers out of the Kernel!

Peter Chubb[*]
National ICT Australia
and
The University of New South Wales
peterc@gelato.unsw.edu.au

Abstract

Now that Linux has fast system calls, good (and getting better) threading, and cheap context switches, it's possible to write device drivers that live in user space for whole new classes of devices. Of course, some device drivers (Xfree, in particular) have always run in user space, with a little bit of kernel support. With a little bit more kernel support (a way to set up and tear down DMA safely, and a generalised way to be informed of and control interrupts) almost any PCI bus-mastering device could have a user-mode device driver.

I shall talk about the benefits and drawbacks of device drivers being in user space or kernel space, and show that performance concerns are not really an issue—in fact, on some platforms, our user-mode IDE driver out-performs the in-kernel one. I shall also present profiling and benchmark results that show where time is spent in in-kernel and user-space drivers, and describe the infrastructure I've added to the Linux kernel to allow portable, efficient user-space drivers to be written.

[*]This work was funded by HP, National ICT Australia, the ARC, and the University of NSW through the Gelato programme (http://www.gelato.unsw.edu.au)

1 Introduction

Normal device drivers in Linux run in the kernel's address space with kernel privilege. This is not the only place they can run—see Figure 1.

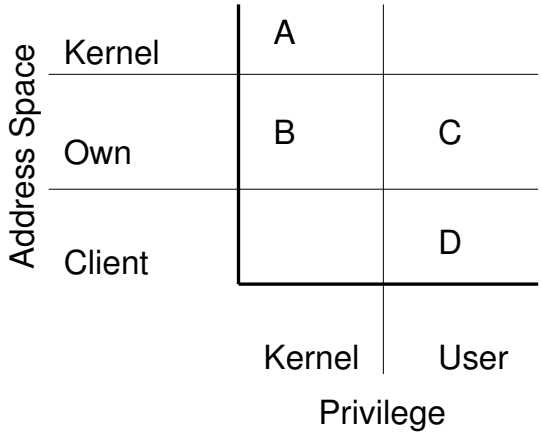


Figure 1: Where a Device Driver can Live

Point A is the normal Linux device driver, linked with the kernel, running in the kernel address space with kernel privilege.

Device drivers can also be linked directly with the applications that use them (Point B)—the so-called 'in-process' device drivers proposed by [Keedy, 1979]—or run in a separate process, and be talked to by an IPC mechanism (for example, an X server, point D). They can also run with kernel privilege, but with a separate kernel address space (Point

C) (as in the Nooks system described by [Swift et al., 2002]).

2 Motivation

Traditionally, device drivers have been developed as part of the kernel source. As such, they *have* to be written in the C language, and they have to conform to the (rapidly changing) interfaces and conventions used by kernel code. Even though drivers can be written as modules (obviating the need to reboot to try out a new version of the driver[1]), in-kernel driver code has access to all of kernel memory, and runs with privileges that give it access to all instructions (not just unprivileged ones) and to all I/O space. As such, bugs in drivers can easily cause kernel lockups or panics. And various studies (e.g., [Chou et al., 2001]) estimate that more than 85% of the bugs in an operating system are driver bugs.

Device drivers that run as user code, however, can use any language, can be developed using any IDE, and can use whatever internal threading, memory management, etc., techniques are most appropriate. When the infrastructure for supporting user-mode drivers is adequate, the processes implementing the driver can be killed and restarted almost with impunity as far as the rest of the operating system goes.

Drivers that run in the kernel have to be updated regularly to match in-kernel interface changes. Third party drivers are therefore usually shipped as source code (or with a compilable stub encapsulating the interface) that has to be compiled against the kernel the driver is to be installed into.

This means that everyone who wants to run a third-party driver also has to have a toolchain and kernel source on his or her system, or obtain a binary for their own kernel from a trusted third party.

Drivers for uncommon devices (or devices that the mainline kernel developers do not use regularly) tend to lag behind. For example, in the 2.6.6 kernel, there are 81 drivers known to be broken because they have not been updated to match the current APIs, and a number more that are still using APIs that have been deprecated.

User/kernel interfaces tend to change much more slowly than in-kernel ones; thus a user-mode driver has much more chance of not needing to be changed when the kernel changes. Moreover, user mode drivers can be distributed under licences other than the GPL, which may make them more attractive to some people[2].

User-mode drivers can be either closely or loosely coupled with the applications that use them. Two obvious examples are the X server (XFree86) which uses a socket to communicate with its clients and so has isolation from kernel and client address spaces and can be very complex; and the Myrinet drivers, which are usually linked into their clients to gain performance by eliminating context switch overhead on packet reception.

The Nooks work [Swift et al., 2002] showed that by isolating drivers from the kernel address space, the most common programming errors could be made recoverable. In Nooks, drivers are insulated from the rest of the kernel by running each in a separate address space, and replacing the driver $\leftrightarrow$ kernel interface with a new one that uses cross-domain procedure calls to replace any procedure calls in the ABI, and that creates shadow copies of any

[1] except that many drivers currently cannot be unloaded

[2] for example, the ongoing problems with the Nvidia graphics card driver could possibly be avoided.

shared variables in the protected address space of the driver.

This approach provides isolation, but also has problems: as the driver model changes, there is quite a lot of wrapper code that has to be changed to accommodate the changed APIs. Also, the value of any shared variable is frozen for the duration of a driver ABI call. The Nooks work is uniprocessor only; locking issues therefore have not yet been addressed.

Windriver [Jungo, 2003] allows development of user mode device drivers. It loads a proprietary device module `/dev/windrv6`; user code can interact with this device to setup and teardown DMA, catch interrupts, etc.

Even from user space, of course, it is possible to make your machine unusable. Device drivers have to be trusted to a certain extent to do what they are advertised to do; this means that they can program their devices, and possibly corrupt or spy on the data that they transfer between their devices and their clients. Moving a driver to user space does not change this. It does however make it less likely that a fault in a driver will affect anything other than its clients

3 Existing Support

Linux has good support for user-mode drivers that do not need DMA or interrupt handling—see, e.g., [Nakatani, 2002].

The `ioperm()` and `iopl()` system calls allow access to the first 65536 I/O ports; and, with a patch from Albert Calahan[3] one can map the appropriate parts of */proc/bus/pci/...* to gain access to memory-mapped registers. Or on some architectures it is safe to `mmap()` */dev/mem*.

[3] `http://lkml.org/lkml/2003/7/13/258`

It is usually best to use MMIO if it is available, because on many 64-bit platforms there are more than 65536 ports—the PCI specification says that there are 2^{32} ports available—(and on many architectures the ports are emulated by mapping memory anyway).

For particular devices—USB input devices, SCSI devices, devices that hang off the parallel port, and video drivers such as XFree86—there is explicit kernel support. By opening a file in */dev*, a user-mode driver can talk through the USB hub, SCSI controller, AGP controller, etc., to the device. In addition, the *input* handler allows input events to be queued back into the kernel, to allow normal event handling to proceed.

libpci allows access to the PCI configuration space, so that a driver can determine what interrupt, IO ports and memory locations are being used (and to determine whether the device is present or not).

Other recent changes—an improved scheduler, better and faster thread creation and synchronisation, a fully preemptive kernel, and faster system calls—mean that it is possible to write a driver that operates in user space that is almost as fast as an in-kernel driver.

4 Implementing the Missing Bits

The parts that are missing are:

1. the ability to claim a device from user space so that other drivers do not try to handle it;
2. The ability to deliver an interrupt from a device to user space,
3. The ability to set up and tear-down DMA between a device and some process's memory, and

4. the ability to loop a device driver's control and data interfaces into the appropriate part of the kernel (so that, for example, an IDE driver can appear as a standard block device), preferably without having to copy any payload data.

The work at UNSW covers only PCI devices, as that is the only bus available on all of the architectures we have access to (IA64, X86, MIPS, PPC, alpha and arm).

4.1 PCI interface

Each device should have only a single driver. Therefore one needs a way to associate a driver with a device, and to remove that association automatically when the driver exits. This has to be implemented in the kernel, as it is only the kernel that can be relied upon to clean up after a failed process. The simplest way to keep the association and to clean it up in Linux is to implement a new filesystem, using the PCI namespace. Open files are automatically closed when a process exits, so cleanup also happens automatically.

A new system call, `usr_pci_open(int bus, int slot, int fn)` returns a file descriptor. Internally, it calls `pci_enable_device()` and `pci_set_master()` to set up the PCI device after doing the standard filesystem boilerplate to set up a vnode and a `struct file`.

Attempts to open an already-opened PCI device will fail with `-EBUSY`.

When the file descriptor is finally closed, the PCI device is released, and any DMA mappings removed. All files are closed when a process dies, so if there is a bug in the driver that causes it to crash, the system recovers ready for the driver to be restarted.

4.2 DMA handling

On low-end systems, it's common for the PCI bus to be connected directly to the memory bus, so setting up a DMA transfer means merely pinning the appropriate bit of memory (so that the VM system can neither swap it out nor relocate it) and then converting virtual addresses to physical addresses.

There are, in general, two kinds of DMA, and this has to be reflected in the kernel interface:

1. Bi-directional DMA, for holding scatter-gather lists, etc., for communication with the device. Both the CPU and the device read and write to a shared memory area. Typically such memory is uncached, and on some architectures it has to be allocated from particular physical areas. This kind of mapping is called *PCI-consistent*; there is an internal kernel ABI function to allocate and deallocate appropriate memory.

2. Streaming DMA, where, once the device has either read or written the area, it has no further immediate use for it.

I implemented a new system call[4], `usr_pci_map()`, that does one of three things:

1. Allocates an area of memory suitable for a PCI-consistent mapping, and maps it into the current process's address space; or

2. Converts a region of the current process's virtual address space into a scatterlist in terms of virtual addresses (one entry per page), pins the memory, and converts the

[4] Although multiplexing system calls are in general deprecated in Linux, they are extremely useful while developing, because it is not necessary to change every architecture-dependent *entry.S* when adding new functionality

scatterlist into a list of addresses suitable for DMA (by calling `pci_map_sg()`, which sets up the IOMMU if appropriate), or

3. Undoes the mapping in point 2.

The file descriptor returned from `usr_pci_open()` is an argument to `usr_pci_map()`. Mappings are tracked as part of the private data for that open file descriptor, so that they can be undone if the device is closed (or the driver dies).

Underlying `usr_pci_map()` are the kernel routines `pci_map_sg()` and `pci_unmap_sg()`, and the kernel routine `pci_alloc_consistent()`.

Different PCI cards can address different amounts of DMA address space. In the kernel there is an interface to request that the dma addresses supplied are within the range addressable by the card. The current implementation assumes 32-bit addressing, but it would be possible to provide an interface to allow the real capabilities of the device to be communicated to the kernel.

4.2.1 The IOMMU

Many modern architectures have an IO memory management unit (see Figure 2), to convert from physical to I/O bus addresses—in much the same way that the processor's MMU converts virtual to physical addresses—allowing even thirty-two bit cards to do single-cycle DMA to anywhere in the sixty-four bit memory address space.

On such systems, after the memory has been pinned, the IOMMU has to be set up to translate from bus to physical addresses; and then after the DMA is complete, the translation can be removed from the IOMMU.

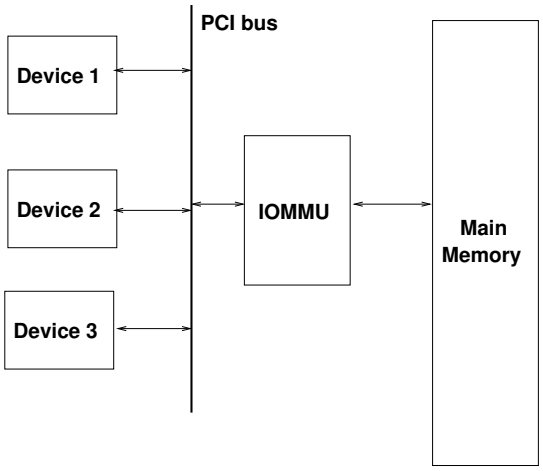


Figure 2: The IO MMU

The processor's MMU also protects one virtual address space from another. Currently shipping IOMMU hardware does not do this: all mappings are visible to all PCI devices, and moreover for some physical addresses on some architectures the IOMMU is bypassed.

For fully secure user-space drivers, one would want this capability to be turned off, and also to be able to associate a range of PCI bus addresses with a particular card, and disallow access by that card to other addresses. Only thus could one ensure that a card could perform DMA only into memory areas explicitly allocated to it.

4.3 Interrupt Handling

There are essentially two ways that interrupts can be passed to user level.

They can be mapped onto signals, and sent asynchronously, or a synchronous 'wait-for-signal' mechanism can be used.

A signal is a good intuitive match for what an interrupt *is*, but has other problems:

1. One is fairly restricted in what one can do in a signal handler, so a driver will usually

have to take extra context switches to respond to an interrupt (into and out of the signal handler, and then perhaps the interrupt handler thread wakes up)

2. Signals can be slow to deliver on busy systems, as they require the process table to be locked. It would be possible to short circuit this to some extent.

3. One needs an extra mechanism for registering interest in an interrupt, and for tearing down the registration when the driver dies.

For these reasons I decided to map interrupts onto file descriptors. */proc* already has a directory for each interrupt (containing a file that can be written to to adjust interrupt routing to processors); I added a new file to each such directory. Suitably privileged processes can open and read these files. The files have open-once semantics; attempts to open them while they are open return -1 with `EBUSY`.

When an interrupt occurs, the in-kernel interrupt handler masks just that interrupt in the interrupt controller, and then does an `up()` operation on a semaphore (well, actually, the implementation now uses a wait queue, but the effect is the same).

When a process reads from the file, then kernel enables the interrupt, then calls `down()` on a semaphore, which will block until an interrupt arrives.

The actual data transferred is immaterial, and in fact none ever is transferred; the `read()` operation is used merely as a synchronisation mechanism.

`poll()` is also implemented, so a user process is not forced into the 'wait for interrupt' model that we use.

Obviously, one cannot share interrupts between devices if there is a user process involved. The in-kernel driver merely passes the interrupt onto the user-mode process; as it knows nothing about the underlying hardware, it cannot tell if the interrupt is *really* for this driver or not. As such it always reports the interrupt as 'handled.'

This scheme works only for level-triggered interrupts. Fortunately, all PCI interrupts are level triggered.

If one really wants a signal when an interrupt happens, one can arrange for a `SIGIO` using `fcntl()`.

It may be possible, by more extensive rearrangement of the interrupt handling code, to delay the end-of-interrupt to the interrupt controller until the user process is ready to get an interrupt. As masking and unmasking interrupts is slow if it has to go off-chip, delaying the EOI should be significantly faster than the current code. However, interrupt delivery to userspace turns out not to be a bottleneck, so there's not a lot of point in this optimisation (profiles show less than 0.5% of the time is spent in the kernel interrupt handler and delivery even for heavy interrupt load—around 1000 cycles per interrupt).

5 Driver Structure

The user-mode drivers developed at UNSW are structured as a preamble, an interrupt thread, and a control thread (see Figure 3).

The preamble:

1. Uses *libpci.a* to find the device or devices it is meant to drive,

2. Calls `usr_pci_open()` to claim the device, and

3. Spawns the interrupt thread, then

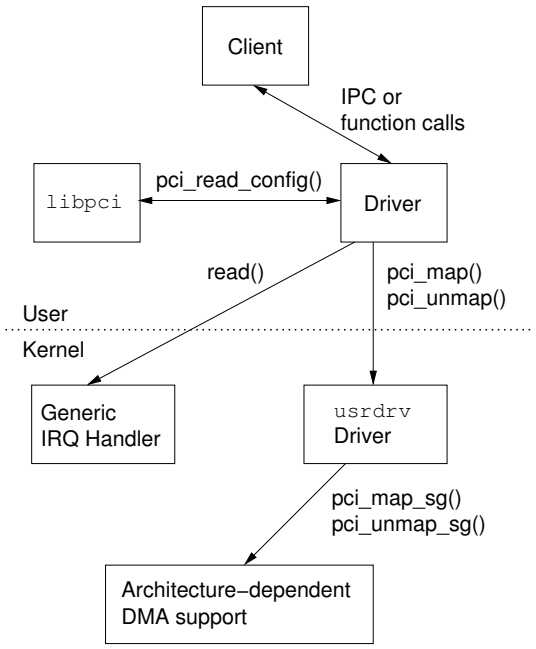


Figure 3: Architecture of a User-Mode Device Driver

4. Goes into a loop collecting client requests.

The interrupt thread:

1. Opens */proc/irq/***irq***/irq*
2. Loops calling `read()` on the resulting file descriptor and then calling the driver proper to handle the interrupt.
3. The driver handles the interrupt, calls out to the control thread(s) to say that work is completed or that there has been an error, queues any more work to the device, and then repeats from step 2.

For the lowest latency, the interrupt thread can be run as a real time thread. For our benchmarks, however, this was not done.

The control thread queues work to the driver then sleeps on a semaphore. When the driver, running in the interrupt thread, determines that a request is complete, it signals the semaphore so that the control thread can continue. (The semaphore is implemented as a pthreads mutex).

The driver relies on system calls and threading, so the fast system call support now available in Linux, and the NPTL are very important to get good performance. Each physical I/O involves at least three system calls, plus whatever is necessary for client communication: a `read()` on the interrupt FD, calls to set up and tear down DMA, and maybe a `futex()` operation to wake the client.

The system call overhead could be reduced by combining DMA setup and teardown into a single system call.

6 Looping the Drivers

An operating system has two functions with regard to devices: firstly to drive them, and secondly to abstract them, so that all devices of the same class have the same interface. While a standalone user-level driver is interesting in its own right (and could be used, for example, to test hardware, or could be linked into an application that doesn't like sharing the device with anyone), it is much more useful if the driver can be used like any other device.

For the network interface, that's easy: use the tun/tap interface and copy frames between the driver and */dev/net/tun*. Having to copy slows things down; others on the team here are planning to develop a zero-copy equivalent of tun/tap.

For the IDE device, there's no standard Linux way to have a user-level block device, so I implemented one. It is a filesystem that has pairs of directories: a master and a slave. When the filesystem is mounted, creating a file in the master directory creates a set of block device special files, one for each potential partition, in

the slave directory. The file in the master directory can then be used to communicate via a very simple protocol between a user level block device and the kernel's block layer. The block device special files in the slave directory can then be opened, closed, read, written or mounted, just as any other block device.

The main reason for using a mounted filesystem was to allow easy use of dynamic major numbers.

I didn't bother implementing ioctl; it was not necessary for our performance tests, and when the driver runs at user level, there are cleaner ways to communicate out-of-band data with the driver, anyway.

7 Results

Device drivers were coded up by [Leslie and Heiser, 2003] for a CMD680 IDE disc controller, and by another PhD student (Daniel Potts) for a DP83820 Gigabit ethernet controller. Daniel also designed and implemented the tuntap interface.

7.1 IDE driver

The disc driver was linked into a program that read 64 Megabytes of data from a Maxtor 80G disc into a buffer, using varying read sizes. Measurements were also made using Linux's in-kernel driver, and a program that read 64M of data from the same on-disc location using `O_DIRECT` and the same read sizes.

We also measured write performance, but the results are sufficiently similar that they are not reproduced here.

At the same time as the tests, a low-priority process attempted to increment a 64-bit counter as fast as possible. The number of increments was calibrated to processor time on an otherwise idle system; reading the counter before and after a test thus gives an indication of how much processor time is available to processes other than the test process.

The initial results were disappointing; the user-mode drivers spent far too much time in the kernel. This was tracked down to `kmalloc()`; so the `usr_pci_map()` function was changed to maintain a small cache of free mapping structures instead of calling `kmalloc()` and `kfree()` each time (we could have used the slab allocator, but it's easier to ensure that the same cache-hot descriptor is reused by coding a small cache ourselves). This resulted in the performance graphs in Figure 4.

The two drivers compared are the new CMD680 driver running in user space, and Linux's in-kernel SIS680 driver. As can be seen, there is very little to choose between them.

The graphs show average of ten runs; the standard deviations were calculated, but are negligible.

Each transfer request takes five system calls to do, in the current design. The client queues work to the driver, which then sets up DMA for the transfer (system call one), starts the transfer, then returns to the client, which then sleeps on a semaphore (system call two). The interrupt thread has been sleeping in `read()`, when the controller finishes its DMA, it cause an interrupt, which wakes the interrupt thread (half of system call three). The interrupt thread then tears down the DMA (system call four), and starts any queued and waiting activity, then signals the semaphore (system call five) and goes back to read the interrupt FD again (the other half of system call three).

When the transfer is above 128k, the IDE controller can no longer do a single DMA opera-

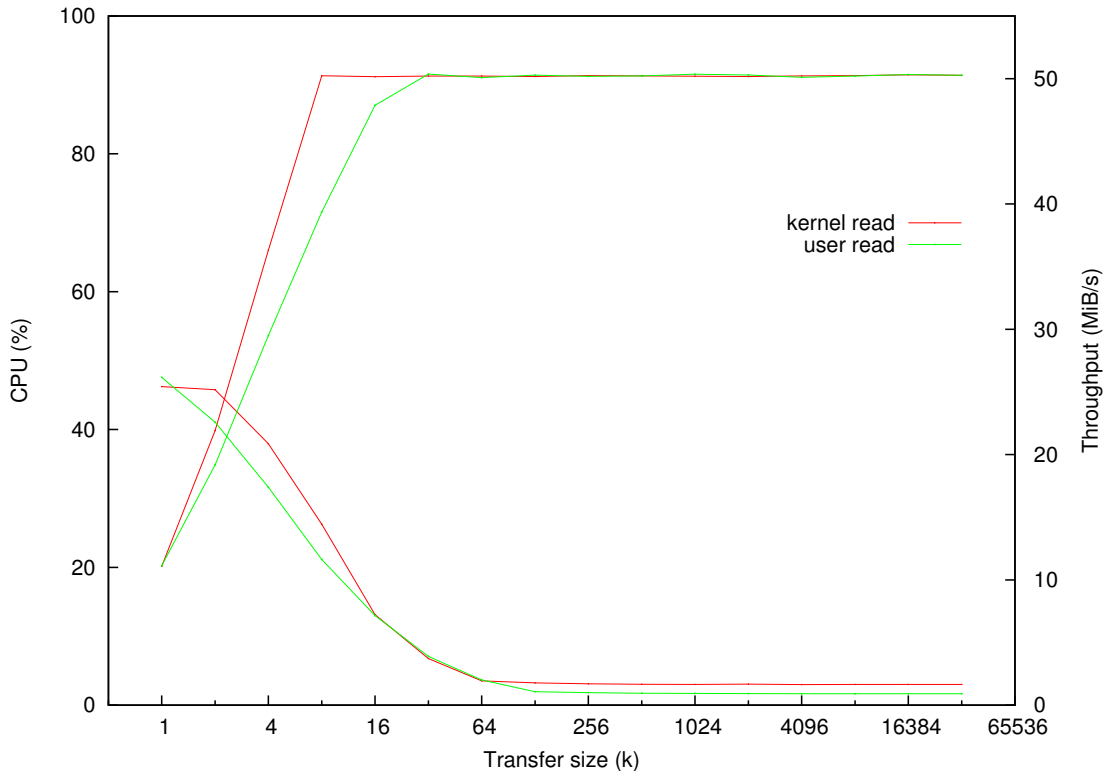


Figure 4: Throughput and CPU usage for the user-mode IDE driver on Itanium-2, reading from a disk

tion, so has to generate multiple transfers The Linux kernel splits DMA requests above 64k, thus increasing the overhead.

The time spent in this driver is divided as shown in Figure 5.

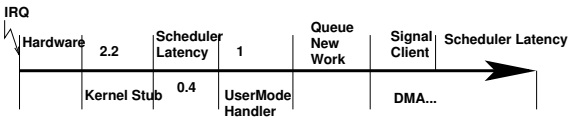


Figure 5: Timeline (in μseconds)

7.2 Gigabit Ethernet

The Gigabit driver results are more interesting. We tested these using [ipbench, 2004] with four clients, all with pause control turned off. We ran three tests:

1. Packet receive performance, where packets were dropped and counted at the layer immediately above the driver

2. Packet transmit performance, where packets were generated and fed to the driver, and

3. Ethernet-layer packet echoing, where the protocol layer swapped source and destination MAC-addresses, and fed received packets back into the driver.

We did not want to start comparing IP stacks, so none of these tests actually use higher level protocols.

We measured three different configurations: a standalone application linked with the driver, the driver looped back into */dev/net/tap* and the standard in-kernel driver, all with interrupt

holdoff set to 0, 1, or 2. (By default, the normal kernel driver sets the interrupt holdoff to 300 μseconds, which led to too many packets being dropped because of FIFO overflow) Not all tests were run in all configurations—for example the linux in-kernel packet generator is sufficiently different from ours that no fair comparison could be made.

For the tests that had the driver residing in or feeding into the kernel, we implemented a new protocol module to count and either echo or drop packets, depending on the benchmark.

In all cases, we used the amount of work achieved by a low priority process to measure time available for other work while the test was going on.

The throughput graphs in all cases are the same. The maximum possible speed on the wire is given for raw ethernet by $10^9 \times p/(p + 38)$ bits per second (the parameter 38 is the ethernet header size (14 octets), plus a 4 octet frame check sequence, plus a 7 octet preamble, plus a 1 octet start frame delimiter plus the minimum 12 octet interframe gap; p is the packet size in octets). For large packets the performance in all cases was the same as the theoretical maximum. For small packet sizes, the throughput is limited by the PCI bus; you'll notice that the slope of the throughput curve when echoing packets is around half the slope when discarding packets, because the driver has to do twice as many DMA operations per packet.

The user-mode driver ('Linux user' on the graph) outperforms the in-kernel driver ('Linux orig')—not in terms of throughput, where all the drivers perform identically, but in using *much* less processing time.

This result was so surprising that we repeated the tests using an EEpro1000, purportedly a card with a much better driver, but saw the same effect—in fact the achieved echo performance is worse than for the in-kernel ns83820 driver for some packet sizes.

The reason appears to be that our driver has a fixed number of receive buffers, which are reused when the client is finished with them—they are allocated only once. This is to provide congestion control at the lowest possible level—the card drops packets when the upper layers cannot keep up.

The Linux kernel drivers have an essentially unlimited supply of receive buffers. Overhead involved in allocating and setting up DMA for these buffers is excessive, and if the upper layers cannot keep up, congestion is detected and the packets dropped in the protocol layer—after significant work has been done in the driver.

One sees the same problem with the user mode driver feeding the tuntap interface, as there is no feedback to throttle the driver. Of course, here there is an extra copy for each packet, which also reduces performance.

7.3 Reliability and Failure Modes

In general the user-mode drivers are very reliable. Bugs in the drivers that would cause the kernel to crash (for example, a null pointer reference inside an interrupt handler) cause the driver to crash, but the kernel continues. The driver can then be fixed and restarted.

8 Future Work

The main foci of our work now lie in:

1. Reducing the need for context switches and system calls by merging system calls, and by trying new driver structures.

2. A zero-copy implementation of tun/tap.

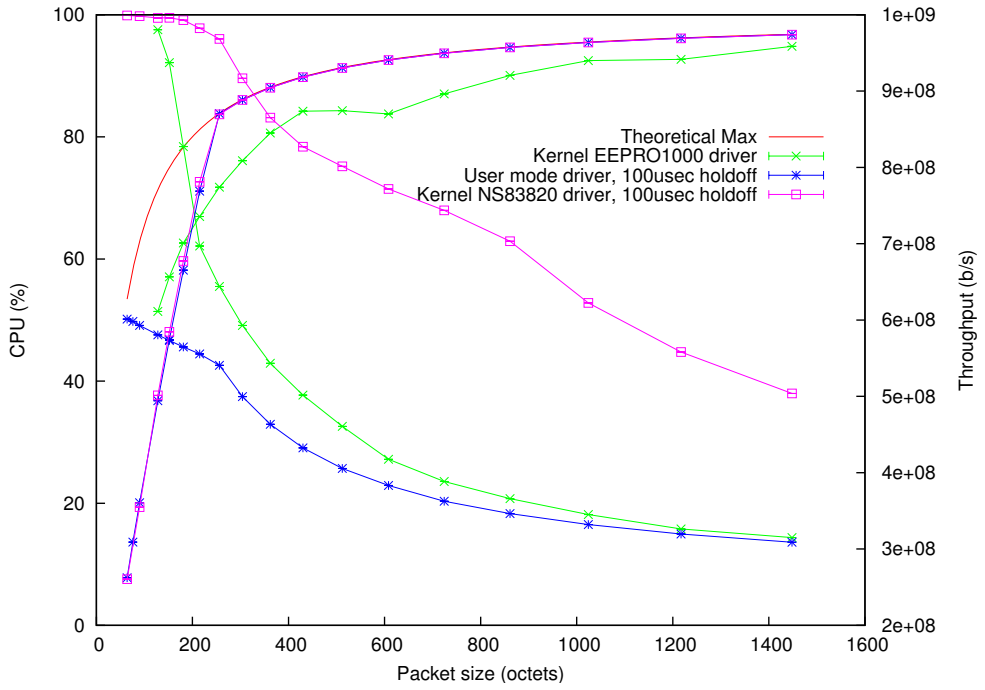


Figure 6: Receive Throughput and CPU usage for Gigabit Ethernet drivers on Itanium-2

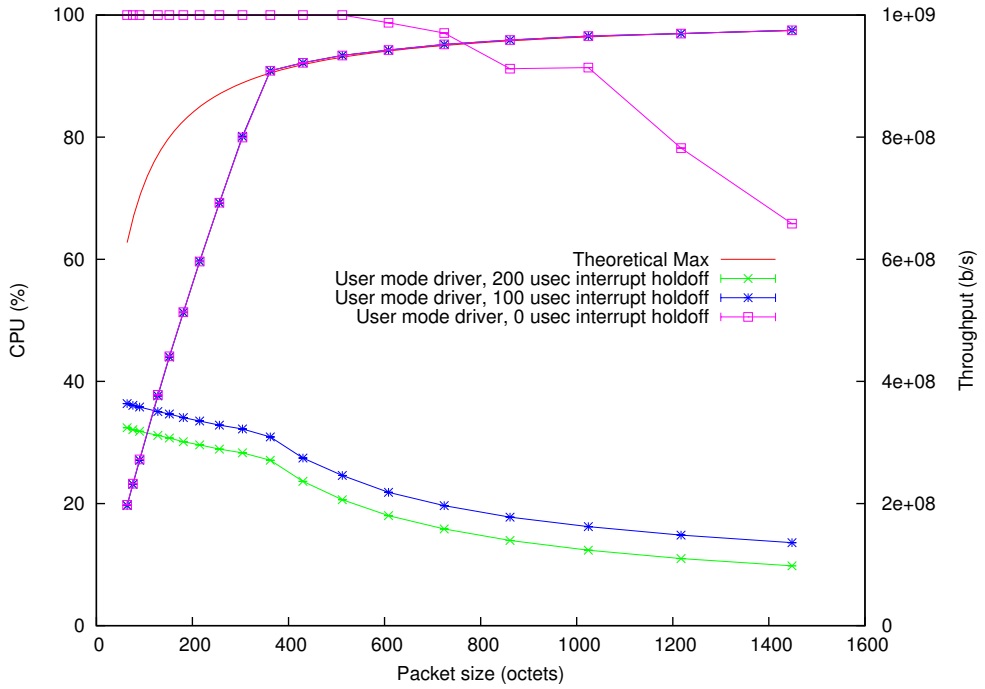


Figure 7: Transmit Throughput and CPU usage for Gigabit Ethernet drivers on Itanium-2

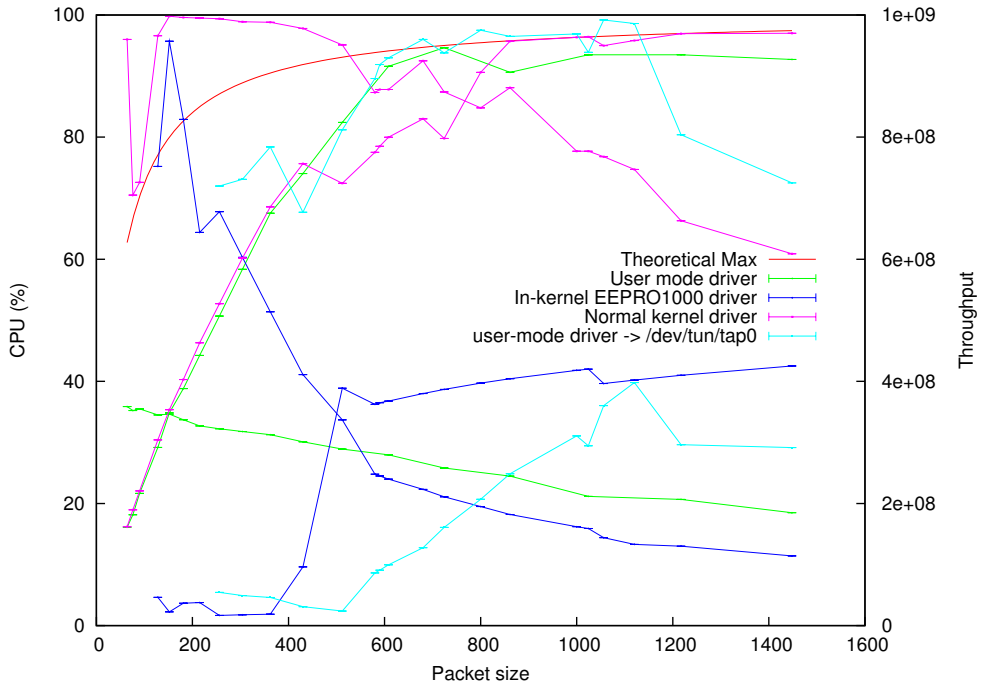


Figure 8: MAC-layer Echo Throughput and CPU usage for Gigabit Ethernet drivers on Itanium-2

3. Improving robustness and reliability of the user-mode drivers, by experimenting with the IOMMU on the ZX1 chipset of our Itanium-2 machines.

4. Measuring the reliability enhancements, by using artificial fault injection to see what problems that cause the kernel to crash are recoverable in user space.

5. User-mode filesystems.

In addition there are some housekeeping tasks to do before this infrastructure is ready for inclusion in a 2.7 kernel:

1. Replace the ad-hoc memory cache with a proper slab allocator.

2. Clean up the system call interface

9 Where d'ya Get It?

Patches against the 2.6 kernel are sent to the Linux kernel mailing list, and are on `http://www.gelato.unsw.edu.au/patches`

Sample drivers will be made available from the same website.

10 Acknowledgements

Other people on the team here did much work on the actual implementation of the user level drivers and on the benchmarking infrastructure. Prominent among them were Ben Leslie (IDE driver, port of our dp83820 into the kernel), Daniel Potts (DP83820 driver, tuntap interface), and Luke McPherson and Ian Wienand (IPbench).

References

[Chou et al., 2001] Chou, A., Yang, J., Chelf, B., Hallem, S., and Engler, D. R. (2001). An empirical study of operating systems errors. In *Symposium on Operating Systems Principles*, pages 73–88. `http://citeseer.nj.nec.com/article/chou01empirical.html`.

[ipbench, 2004] ipbench (2004). ipbench — a distributed framework for network benchmarking. `http://ipbench.sf.net/`.

[Jungo, 2003] Jungo (2003). Windriver. `http://www.jungo.com/windriver.html`.

[Keedy, 1979] Keedy, J. L. (1979). A comparison of two process structuring models. MONADS Report 4, Dept. Computer Science, Monash University.

[Leslie and Heiser, 2003] Leslie, B. and Heiser, G. (2003). Towards untrusted device drivers. Technical Report UNSW-CSE-TR-0303, Operating Systems and Distributed Systems Group, School of Computer Science and Engineering, The University of NSW. CSE techreports website, `ftp://ftp.cse.unsw.edu.au/pub/doc/papers/UNSW/0303.pdf`.

[Nakatani, 2002] Nakatani, B. (2002). ELJOnline: User mode drivers. `http://www.linuxdevices.com/articles/AT5731658926.html`.

[Swift et al., 2002] Swift, M., Martin, S., Leyand, H. M., and Eggers, S. J. (2002). Nooks: an architecture for reliable device drivers. In *Proceedings of the Tenth ACM SIGOPS European Workshop*, Saint-Emilion, France.

Big Servers—2.6 compared to 2.4

Wim A. Coekaerts
Oracle Corporation
wim.coekaerts@oracle.com

Abstract

Linux 2.4 has been around in production environments at companies for a few years now, we have been able to gather some good data on how well (or not) things scale up. Number of CPU's, amount of memory, number of processes, IO throughput, etc.

Most of the deployments in production today, are on relatively small systems, 4- to 8-ways, 8–16GB of memory, in a few cases 32GB. The architecture of choice has also been IA32. 64-bit systems are picking up in popularity rapidly, however.

Now with 2.6, a lot of the barriers are supposed to be gone. So, have they really? How much memory can be used now, how is cpu scaling these days, how good is IO throughput with multiple controllers in 2.6.

A lot of people have the assumption that 2.6 resolves all of this. We will go into detail on what we have found out, what we have tested and some of the conclusions on how good the move to 2.6 will really be.

1 Introduction

The comparison between the 2.4 and 2.6 kernel trees are not solely based on performance. A large part of the testsuites are performance benchmarks however, as you will see, they have been used to also measure stability. There are a number of features added which improve stability of the kernel under heavy workloads. The goal of comparing the two kernel releases was more to show how well the 2.6 kernel will be able to hold up in a real world production environment. Many companies which have deployed Linux over the last two years are looking forward to rolling out 2.6 and it is important to show the benefits of doing such a move. It will take a few releases before the required stability is there however it's clear so far that the 2.6 kernel has been remarkably solid, so early on.

Most of the 2.4 based tests have been run on Red Hat Enterprise Linux 3, based on Linux 2.4.21. This is the enterprise release of Red Hat's OS distribution; it contains a large number of patches on top of the Linux 2.4 kernel tree. Some of the tests have been run on the `kernel.org` mainstream 2.4 kernel, to show the benefit of having extra functionality. However it is difficult to even just boot up the mainstream kernel on the test hardware due to lack of support for drivers, or lack of stability to complete the testsuite. The interesting thing to keep in mind is that with the current Linux 2.6 main stream kernel, most of the testsuites ran through completition. A number of test runs on Linux 2.6 have been on Novell/SuSE SLES9 beta release.

2 Test Suites

The test suites used to compare the various kernels are based on an IO simulator for Oracle, called OraSim and a TPC-C like workload generator called OAST.

Oracle Simulator (OraSim) is a stand-alone tool designed to emulate the platform-critical activities of the Oracle database kernel. Oracle designed Oracle Simulator to test and characterize the input and output (I/O) software stack, the storage system, memory management, and cluster management of Oracle single instances and clusters. Oracle Simulator supports both pass-fail testing for validation, and analytical testing for debugging and tuning. It runs multiple processes, with each process representing the parameters of a particular type of system load similar to the Oracle database kernel.

OraSim is a relatively straightforward IO stresstest utility, similar to IOzone or tiobench, however it is built to be very flexible and configurable.

It has its own script language which allows one to build very complex IO patterns. The tool is not released under any open source license today because it has some code linked in which is part of the RDBMS itself. The jobfiles used for the testing are available online `http://oss.oracle.com/external/ols/jobfiles/`.

The advantage of using OraSim over a real database benchmark is mainly the simplicity. It does not require large amounts of memory or large installed software components. There is one executable which is started with the jobfile as a parameter.The jobfiles used can be easily modified to turn on certain filesystem features, such as asynchronous IO.

OraSim jobfiles were created to simulate a relatively small database. 10 files are defined as actual database datafiles and two files are used to simulate database journals.

OAST on the other hand is a complete database stress test kit, based on the TPC-C benchmark workloads. It requires a full installation of the database software and relies on an actual database environment to be created. TPC-C is an on-line transaction workload. The numbers represented during the testruns are not actual TPC-C benchmarks results and cannot or should not be used as a measure of TPC-C performance—they are TPC-C-like; however, not the same.

The database engine which runs the OAST benchmark allocates a large shared memory segment which contains the database caches for SQL and for data blocks (shared pool and buffer cache). Every client connection can run on the same server or the connection can be over TCP. In case of a local connection, for each client, 2 processes are spawned on the system. One process is a dedicated database process and the other is the client code which communicates with the database server process through IPC calls. Test run parameters include run time length in seconds and number of client connections. As you can see in the result pages, both remote and local connections have been tested.

3 Hardware

A number of hardware configurations have been used. We tried to include various CPU architectures as well as local SCSI disk versus network storage (NAS) and fibre channel (SAN).

Configuration 1 consists of an 8-way IA32 Xeon 2 GHz with 32GB RAM attached to an EMC CX300 Clariion array with 30 147GB disks using a QLA2300 fibre channel HBA. The network cards are BCM5701 Broadcom Gigabit Ethernet.

Configuration 2 consists of an 8-way Itanium 2 1.3 GHz with 8GB RAM attached to a JBOD fibre channel array with 8 36GB disks using a QLA2300 fibre channel HBA. The network cards are BCM5701 Broadcom Gigabit Ethernet.

Configuration 3 consists of a 2-way AMD64 2 GHz (Opteron 246) with 6GB RAM attached to local SCSI disk (LSI Logic 53c1030).

4 Operating System

The Linux 2.4 test cases were created using Red Hat Enterprise Linux 3 on all architectures. Linux 2.6 was done with SuSE SLES9 on all architectures; however, in a number of tests the kernel was replaced by the 2.6 mainstream kernel for comparison.

The test suites and benchmarks did not have to be recompiled to run on either RHEL3 or SLES9. Of course different executables were used on the three CPU architectures.

5 Test Results

At the time of writing a lot of changes were still happening on the 2.6 kernel. As such, the actual spreadsheets with benchmark data has been published on a website, the data is up-to-date with the current kernel tree and can be found here: `http://oss.oracle.com/external/ols/results/`

5.1 IO

If you want to build a huge database server, which can handle thousands of users, it is important to be able to attach a large number of disks. A very big shortcoming in Linux 2.4 was the fact that it could only handle 128 or 256.

With some patches SuSE got to around 3700 disks in SLES8, however that meant stealing major numbers from other components. Really large database setups which also require very high IO throughput, usually have disks attached ranging from a few hundred to a few thousand.

With the 64-bit `dev_t` in 2.6, it's now possible to attach plenty of disk. Without modifications it can easily handle tens of thousands of devices attached. This opens the world to really large scale datawarehouses, tens of terabytes of storage.

Another important change is the block IO layer, the BIO code is much more efficient when it comes to large IOs being submitted down from the running application. In 2.4, every IO got broken down into small chunks, sometimes causing bottlenecks on allocating accounting structures. Some of the tests compared 1MB `read()` and `write()` calls in 2.4 and 2.6.

5.2 Asynchronous IO and DirectIO

If there is one feature that has always been on top of the Must Have list for large database vendors, it must be async IO. Asynchronous IO allows processes to submit batches of IO operations and continue on doing different tasks in the meantime. It improves CPU utilization and can keep devices more busy. The Enterprise distributions based on Linux 2.4 all ship with the async IO patch applied on top of the mainline kernel.

Linux 2.6 has async IO out of the box. It is implemented a little different from Linux 2.4 however combined with support for direct IO it is very performant. Direct IO is very useful as it eliminates copying the userspace buffers into kernel space. On systems that are constantly overloaded, there is a nice performance im-

provement to be gained doing direct IO. Linux 2.4 did not have direct IO and async IO combined. As you can see in the performance graph on AIO+DIO, it provides a significant reduction in CPU utilization.

5.3 Virtual Memory

There has been another major VM overhaul in Linux 2.6, in fact, even after 2.6.0 was released a large portion has been re-written. This was due to large scale testing showing weaknesses as it relates to number of users that could be handled on a system. As you can see in the test results, we were able to go from around 3000 users to over 7000 users. In particular on 32-bit systems, the VM has been pretty much a disaster when it comes to deploying a system with more than 16GB of RAM. With the latest VM changes it is now possible to push a 32GB even up to 48GB system pretty reliably.

Support for large pages has also been a big winner. `HUGETLBFS` reduces TLB misses by a decent percentage. In some of the tests it provides up to a 3% performance gain. In our tests `HUGETLBFS` would be used to allocate the shared memory segment.

5.4 NUMA

Linux 2.6 is the first Linux kernel with real NUMA support. As we see high-end customers looking at deploying large SMP boxes running Linux, this became a real requirement. In fact even with the AMD64 design, NUMA support becomes important for performance even when looking at just a dual-CPU system.

NUMA support has two components; however, one is the fact that the kernel VM allocates memory for processes in a more efficient way. On the other hand, it is possible for applications to use the NUMA API and tell the OS where memory should be allocated and how. Oracle has an extention for Itanium2 to support the libnuma API from Andi Kleen. Making use of this extention showed a significant improvement, up to about 20%. It allows the database engine to be smart about memory allocations resulting in a significant performance gain.

6 Conclusion

It is very clear that many of the features that were requested by the larger corporations providing enterprise applications actually help a huge amount. The advantage of having Asynchronous IO or NUMA support in the mainstream kernel is obvious. It takes a lot of effort for distribution vendors to maintain patches on top of the mainline kernel and when functionality makes sense it helps to have it be included in mainline. Micro-optimizations are still being done and in particular the VM subsystem can improve quite a bit. Most of the stability issues are around 32-bit, where the LowMem versus HighMem split wreaks havoc quite frequently. At least with some of the features now in the 2.6 kernel it is possible to run servers with more than 16GB of memory and scale up.

The biggest surprise was the stability. It was very nice to see a new stable tree be so solid out of the box, this in contrast to earlier stable kernel trees where it took quite a few iterations to get to the same point.

The major benefit of 2.6 is being able to run on really large SMP boxes: 32-way Itanium2 or Power4 systems with large amounts of memory. This was the last stronghold of the traditional Unices and now Linux can play alongside with them even there. Very exciting times.

Multi-processor and Frequency Scaling

Making Your Server Behave Like a Laptop

Paul Devriendt
AMD Software Research and Development
paul.devriendt@amd.com

Abstract

This paper will explore a multi-processor implementation of frequency management, using an AMD Opteron™ processor 4-way server as a test vehicle.

Topics will include:

- the benefits of doing this, and why server customers are asking for this,
- the hardware, for case of the AMD Opteron processor,
- the various software components that make this work,
- the issues that arise, and
- some areas of exploration for follow on work.

1 Introduction

Processor frequency management is common on laptops, primarily as a mechanism for improving battery life. Other benefits include a cooler processor and reduced fan noise. Fans also use a non-trivial amount of power.

This technology is spreading to desktop machines, driven both by a desire to reduce power consumption and to reduce fan noise. Servers and other multiprocessor machines can equally benefit. The multiprocessor frequency management scenario offers more complexity (no surprise there). This paper discusses these complexities, based upon a test implementation on an AMD Opteron processor 4-way server. Details within this paper are AMD processor specific, but the concepts are applicable to other architectures.

The author of this paper would like to make it clear that he is just the maintainer of the AMD frequency driver, supporting the AMD Athlon™ 64 and AMD Opteron processors. This frequency driver fits into, and is totally dependent, on the CPUFreq support. The author has gratefully received much assistance and support from the CPUFreq maintainer (Dominik Brodowski).

2 Abbreviations

BKDG: The BIOS and Kernel Developer's Guide. Document published by AMD containing information needed by system software developers. See the references section, entry 4.

MSR: Model Specific Register. Processor registers, accessable only from kernel space, used for various control functions. These registers are expected to change across processor families. These registers are described in the

BKDG[4].

VRM: Voltage Regulator Module. Hardware external to the processor that controls the voltage supplied to the processor. The VRM has to be capable of supplying different voltages on command. Note that for multiprocessor systems, it is expected that each processor will have its own independent VRM, allowing each processor to change voltage independently. For systems where more than one processor shares a VRM, the processors have to be managed as a group. The current frequency driver does not have this support.

fid: Frequency Identifier. The values written to the control MSR to select a core frequency. These identifiers are processor family specific. Currently, these are six bit codes, allowing the selection of frequencies from 800 MHz to 5 Ghz. See the BKDG[4] for the mappings from fid to frequency. Note that the frequency driver does need to "understand" the mapping of fid to frequency, as frequencies are exposed to other software components.

vid: Voltage Identifier. The values written to the control MSR to select a voltage. These values are then driven to the VRM by processor logic to achieve control of the voltage. These identifiers are processor model specific. Currently these identifiers are five bit codes, of which there are two sets—a standard set and a low-voltage mobile set. The frequency driver does not need to be able to "understand" the mapping of vid to voltage, other than perhaps for debug prints.

VST: Voltage Stabilization Time. The length of time before the voltage has increased and is stable at a newly increased voltage. The driver has to wait for this time period when stepping the voltage up. The voltage has to be stable at the new level before applying a further step up in voltage, or before transitioning to a new frequency that requires the higher voltage.

MVS: Maximum Voltage Step. The maximum voltage step that can be taken when increasing the voltage. The driver has to step up voltage in multiple steps of this value when increasing the voltage. (When decreasing voltage it is not necessary to step, the driver can merely jump to the correct voltage.) A typical MVS value would be 25mV.

RVO: Ramp Voltage Offset. When transitioning frequencies, it is necessary to temporarily increase the nominal voltage by this amount during the frequency transition. A typical RVO value would be 50mV.

IRT: Isochronous Relief Time. During frequency transitions, busmasters briefly lose access to system memory. When making multiple frequency changes, the processor driver must delay the next transition for this time period to allow busmasters access to system memory. The typical value used is 80us.

PLL: Phase Locked Loop. Electronic circuit that controls an oscillator to maintain a constant phase angle relative to a reference signal. Used to synthesize new frequencies which are a multiple of a reference frequency.

PLL Lock Time: The length of time, in microseconds, for the PLL to lock.

pstate: Performance State. A combination of frequency/voltage that is supported for the operation of the processor. A processor will typically have several pstates available, with higher frequencies needing higher voltages. The processor clock can not be set to any arbitrary frequency; it may only be set to one of a limited set of frequencies. For a given frequency, there is a minimum voltage needed to operate reliably at that frequency, and this is the correct voltage, thus forming the frequency/voltage pair.

ACPI: Advanced Configuration and Power In-

terface Specification. An industry specification, initially developed by Intel, Microsoft, Phoenix and Toshiba. See the reference section, entry 5.

_PSS: Performance Supported States. ACPI object that defines the performance states valid for a processor.

_PPC: Performance Present Capabilities. ACPI object that defines which of the _PSS states are currently available, due to current platform limitations.

PSB Performance State Block. BIOS provided data structure used to pass information, to the driver, concerning the pstates available on the processor. The PSB does not support multiprocessor systems (which use the ACPI _PSS object instead) and is being deprecated. The format of the PSB is defined in the BKDG.

3 Why Does Frequency Management Affect Power Consumption?

Higher frequency requires higher voltage. As an example, data for part number ADA3200AEP4AX:

2.2 GHz @ 1.50 volts, 58 amps max – 89 watts

2.0 GHz @ 1.40 volts, 48 amps max – 69 watts

1.8 GHz @ 1.30 volts, 37 amps max – 50 watts

1.0 GHz @ 1.10 volts, 18 amps max – 22 watts

These figures are worst case current/power figures, at maximum case temperature, and include I/O power of 2.2W.

Actual power usage is determined by:

- code currently executing (idle blocks in the processor consume less power),
- activity from other processors (cache coherency, memory accesses, pass-through traffic on the HyperTransport™ connections),
- processor temperature (current increases with temperature, at constant workload and voltage),
- processor voltage.

Increasing the voltage allows operation at higher frequencies, at the cost of higher power consumption and higher heat generation. Note that relationship between frequency and power consumption is not a linear relationship—a 10% frequency increase will cost more than 10% in power consumption (30% or more).

Total system power usage depends on other devices in the system, such as whether disk drives are spinning or stopped, and on the efficiency of power supplies.

4 Why Should Your Server Behave Like A Laptop?

- Save power. It is the right thing to do for the environment. Note that power consumed is largely converted into heat, which then becomes a load on the air conditioning in the server room.
- Save money. Power costs money. The power savings for a single server are typically regarded as trivial in terms of a corporate budget. However, many large organizations have racks of many thousands of servers. The power bill is then far from trivial.
- Cooler components last longer, and this translates into improved server reliability.
- Government Regulation.

5 Interesting Scenarios

These are real world scenarios, where the application of the technology is appropriate.

5.1 Save power in an idle cluster

A cluster would typically be kept running at all times, allowing remote access on demand. During the periods when the cluster is idle, reducing the CPU frequency is a good way to reduce power consumption (and therefore also air conditioning load), yet be able to quickly transition back up to full speed (<0.1 second) when a job is submitted.

User space code (custom to the management of that cluster) can be used to offer cluster speeds of "fast" and "idle," using the `/proc` or `/sys` file systems to trigger frequency transitions.

5.2 The battery powered server

Or, the server running on a UPS.

Many production servers are connected to a battery backup mechanism (UPS—uninterruptible power supply) in case the mains power fails. Action taken on a mains power failure varies:

- Orderly shutdown.
- Stay up and running for as long as there is battery power, but orderly shutdown if mains power is not restored.
- Stay up and running, mains power will be provided by backup generators as soon as the generators can be started.

In these scenarios, transitioning to lower performance states will maximize battery life, or reduce the amount of generator/battery power capacity required.

UPS notification of mains power loss to the server for administrator alerts is well understood technology. It is not difficult to add the support for transitioning to a lower pstate. This can be done by either a cpufreq governor or by adding the simple user space controls to an existing user space daemon that is monitoring the UPS alerts.

5.3 Server At Less Than Maximum Load

As an example, a busy server may be processing 100 transactions per second, but only 5 transactions per second during quiet periods. Reducing the CPU frequency from 2.2 GHz to 1.0 GHz is not going to impact the ability of that server to process 5 transactions per second.

5.4 Processor is not the bottleneck

The bottleneck may not be the processor speed. Other likely bottlenecks are disk access and network access. Having the processor waiting faster may not improve transaction throughput.

5.5 Thermal cutback to avoid over temperature situations

The processors are the main generators of heat in a system. This becomes very apparent when many processors are in close proximity, such as with blade servers. The effectiveness of the processor cooling is impacted when the processor heat sinks are being cooled with hot air. Reducing processor frequency when idle can dramatically reduce the heat production.

5.6 Smaller Enclosures

The drive to build servers in smaller boxes, whether as standalone machines or slim rackmount machines, means that there is less space for air to circulate. Placing many slim rackmounts together in a rack (of which the most

demanding case is a blade server) aggravates the cooling problem as the neighboring boxes are also generating heat.

6 System Power Budget

The processors are only part of the system. We therefore need to understand the power consumption of the entire system to see how significant processor frequency management is on the power consumption of the whole system.

A system power budget is obviously platform specific. This sample DC (direct current) power budget is for a 4-processor AMD Opteron processor based system. The system has three 500W power supplies, of which one is redundant. Analysis shows that for many operating scenarios, the system could run on a single power supply.

This analysis is of DC power. For the system in question, the efficiency of the power supplies are approximately linear across varying loads, and thus the DC power figures expressed as percentages are meaningful as predictors of the AC (alternating current) power consumption. For systems with power supplies that are not linearly efficient across varying loads, the calculations obviously have to be factored to take account of power supply efficiency.

System components:

- 4 processors @ 89W = 356W in the maximum pstate, 4 @ 22W = 88W in the minimum pstate. These are worst case figures, at maximium case temperature, with the worst case instruction mix. The figures in Table1 are reduced from these maximums by approximately 10% to account for a reduced case temperature and for a workload that does not keep all of the processors' internal units busy.
- Two disk drives (Western Digital 250 GByte SATA), 16W read/write, 10W idle (spinning), 1.3W sleep (not spinning). Note SCSI drives typically consume more power.
- DVD Drive, 10W read, 1W idle/sleep.
- PCI 2.2 Slots – absolute max of 25W per slot, system will have a total power budget that may not account for maximum power in all slots. Estimate 2 slots occupied at a total of 20W.
- VGA video card in a PCI slot. 5W. (AGP would be more like 15W+).
- DDR DRAM, 10W max per DIMM, 40W for 4 GBytes configured as 4 DIMMs.
- Network (built in) 5W.
- Motherboard and components 30W.
- 10 fans @ 6W each. 60W.
- Keyboard + Mouse 3W

See Table 1 for the sample power budget under busy and light loads.

The light load without any frequency reduction is baselined as 100%.

The power consumption is shown for the same light load with frequency reduction enabled, and again where the idle loop incorporates the hlt instruction.

Using frequency management, the power consumption drops to 43%, and adding the use of the hlt instruction (assuming 50% time halted), the power consumption drops further to 33%.

These are significant power savings, for systems that are under light load conditions at times. The percentage of time that the system is running under reduced load has to be known to predict actual power savings.

system load	4 cpus	2 disks	dvd	pci	vga	dram	net	planar	fans	kbd mou	total
busy	320 90%	32	10	20	5	40	5	30	60	3	525W
light load	310 87%	22	1	15	5	38	5	20	60	3	479W 100%
light load, using frequency reduction	79 90%	22	1	15	5	38	5	20	20	3	208W 43%
light load, using frequency reduction and using hlt 50% of the time	32 40%	22	1	15	5	38	5	20	15	3	156 33%

Table 1: Sample System Power Budget (DC), in watts

7 Hardware—AMD Opteron

7.1 Software Interface To The Hardware

There are two MSRs, the FIDVID_STATUS MSR and the FIDVID_CONTROL MSR, that are used for frequency voltage transitions. These MSRs are the same for the single processor AMD Athlon 64 processors and for the AMD Opteron MP capable processors. These registers are not compatible with the previous generation of AMD Athlon processors, and will not be compatible with the next generation of processors.

The CPU frequency driver for AMD processors therefore has to change across processor revisions, as do the ACPI _PSS objects that describe pstates.

The status register reports the current fid and vid, as well as the maximum fid, the start fid, the maximum vid and the start vid of the particular processor.

These registers are documented in the BKDG[4].

As MSRs can only be accessed by executing code (the read msr or write msr instructions) on the target processor, the frequency driver has to use the processor affinity support to force execution on the correct processor.

7.2 Multiple Memory Controllers

In PC architectures, the memory controller is a component of the northbridge, which is traditionally a separate component from the processor. With AMD Opteron processors, the northbridge is built into the processor. Thus, in a multi-processor system there are multiple memory controllers.

See Figure 1 for a block diagram of a two processor system.

If a processor is accessing DRAM that is physically attached to a different processor, the DRAM access (and any cache coherency traffic) crosses the coherent HyperTransport interprocessor links. There is a small performance penalty in this case. This penalty is of the order of a DRAM page hit versus a DRAM page miss, about 1.7 times slower than a local access.

This penalty is minimized by the processor caches, where data/code residing in remote DRAM is locally cached. It is also minimized

by Linux's NUMA support.

Note that a single threaded application that is memory bandwidth constrained may benefit from multiple memory controllers, due to the increase in memory bandwidth.

When the remote processor is transitioned to a lower frequency, this performance penalty is worse. An upper bound to the penalty may be calculated as proportional to the frequency slowdown. I.e., taking the remote processor from 2.2 GHz to 1.0 GHz would take the 1.7 factor from above to a factor of 2.56. Note that this is an absolute worst case, an upper bound to the factor. Actual impact is workload dependent.

A worst case scenario would be a memory bound task, doing memory reads at addresses that are pathologically the worst case for the caches, with all accesses being to remote memory. A more typical scenario would see this penalty alleviated by:

- processor caches, where 64 bytes will be read and cached for a single access, so applications that walk linearly through memory will only see the penalty on 64 byte boundaries,
- memory writes do not take a penalty (as processor execution continues without waiting for a write to complete),
- memory may be interleaved,
- kernel NUMA optimizations for non-interleaved memory (which allocate memory local to the processor when possible to avoid this penalty).

7.3 DRAM Interface Speed

The DRAM interface speed is impacted by the core clock frequency. A full table is published in the processor data sheet; Table 2 shows a sample of actual DRAM frequencies for the common specified DRAM frequencies, across a range of core frequencies.

This table shows that certain DRAM speed / core speed combinations are suboptimal.

Effective memory performance is influenced by many factors:

- cache hit rates,
- effectiveness of NUMA memory allocation routines,
- load on the memory controller,
- size of penalty for remote memory accesses,
- memory speed,
- other hardware related items, such as types of DRAM accesses.

It is therefore necessary to benchmark the actual workload to get meaningful data for that workload.

7.4 UMA

During frequency transitions, and when HyperTransport LDTSTOP is asserted, DRAM is placed into self refresh mode. UMA graphics devices therefore can not access DRAM. UMA systems therefore need to limit the time that DRAM is in self refresh mode. Time constraints are bandwidth dependent, with high resolution displays needing higher memory bandwidth. This is handled by the IRT delay time during frequency transitions. When transitioning multiple steps, the driver waits an appropriate length of time to allow external devices to access memory.

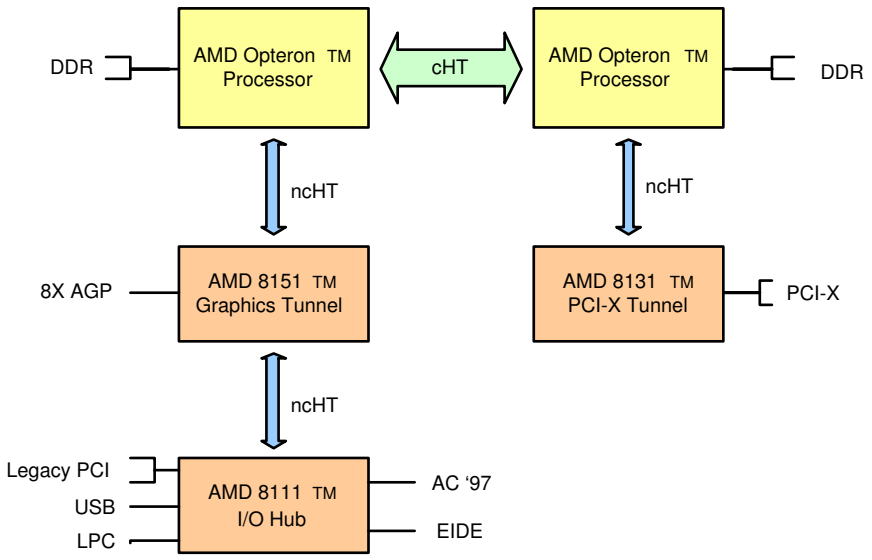


Figure 1: Two Processor System

Processor Core Frequency	100MHz DRAM spec	133MHz DRAM spec	166MHz DRAM spec	200MHz DRAM spec
800MHz	100.00	133.33	160.00	160.00
1000MHz	100.00	125.00	166.66	200.00
2000MHz	100.00	133.33	166.66	200.00
2200MHz	100.00	129.41	157.14	200.00

Table 2: DRAM Frequencies For A Range Of Processor Core Frequencies

7.5 TSC Varying

The Time Stamp Counter (TSC) register is a register that increments with the processor clock. Multiple reads of the register will see increasing values. This register increments on each core clock cycle in the current generation of processors. Thus, the rate of increase of the TSC when compared with "wall clock time" varies as the frequency varies. This causes problems in code that calibrates the TSC increments against an external time source, and then attempts to use the TSC to measure time.

The Linux kernel uses the TSC for such timings, for example when a driver calls udelay(). In this case it is not a disaster if the udelay() call waits for too long as the call is defined to allow this behavior. The case of the udelay() call returning too quickly can be fatal, and this has been demonstrated during experimentation with this code.

This particular problem is resolved by the cpufreq driver correcting the kernel TSC calibration whenever the frequency changes.

This issue may impact other code that uses the TSC register directly. It is interesting to note that it is hard to define a correct behavior. Code that calibrates the TSC against an external clock will be thrown off if the rate of increment of the TSC should change. However, other code may expect a certain code sequence to consistently execute in approximately the same number of cycles, as measured by the TSC, and this code will be thrown off if the behavior of the TSC changes relative to the processor speed.

7.6 Measurement Of Frequency Transition Times

The time required to perform a transition is a combination of the software time to execute the required code, and the hardware time to perform the transition.

Examples of hardware wait time are:

- waiting for the VRM to be stable at a newer voltage,
- waiting for the PLL to lock at the new frequency,
- waiting for DRAM to be placed into and then taken out of self refresh mode around a frequency transition.

The time taken to transition between two states is dependent on both the initial state and the target state. This is due to :

- multiple steps being required in some cases,
- certain operations are lengthier (for example, voltage is stepped up in multiple stages, but stepped down in a single step),
- difference in code execution time dependent on processor speed (although this is minor).

Measurements, taken by calibrating the frequency driver, show that frequency transitions for a processor are taking less than 0.015 seconds.

Further experimentation with multiple processors showed a worst case transition time of less than 0.08 seconds to transition all 4 processors from minimum to maximum frequency, and slightly faster to transition from maximum to minimum frequency.

Note, there is a driver optimization under consideration that would approximately halve these transition times.

7.7 Use of Hardware Enforced Throttling

The southbridge (I/O Hub, example AMD-8111™ HyperTransport I/O Hub) is capable of initiating throttling via the HyperTransport stopclock message, which will ramp down the CPU grid by the programmed amount. This may be initiated by the southbridge for thermal throttling or for other reasons.

This throttling is transparent to software, other than the performance impact.

This throttling is of greatest value in the lowest pstate, due to the reduced voltage.

The hardware enforced throttling is generally not of relevance to the software management of processor frequencies. However, a system designer would need to take care to ensure that the optimal scenarios occur—i.e., transition to a lower frequency/voltage in preference to hardware throttling in high pstates. The BIOS configurations are documented in the BKDG[4].

For maximum power savings, the southbridge would be configured to initiate throttling when the processor executes the `hlt` instruction.

8 Software

The AMD frequency driver is a small part of the software involved. The frequency driver fits into the CPUFreq architecture, which is part of the 2.6 kernel. It is also available as a patch for the 2.4 kernel, and many distributions do include it.

The CPUFreq architecture includes kernel support, the CPUFreq driver itself (`drivers/cpufreq`), an architecture specific driver to control the hardware (powernow-k8.ko is this case), and `/sys` file system code for userland access.

The kernel support code (`linux/kernel/cpufreq.c`) handles timing changes such as updating the kernel constant `loops_per_jiffies`, as well as notifiers (system components that need to be notified of a frequency change).

8.1 History Of The AMD Frequency Driver

The CPU frequency driver for AMD Athlon (the previous generation of processors) was developed by Dave Jones. This driver supports single processor transitions only, as the pstate transition capability was only enabled in mobile processors. This driver used the PSB mechanism to determine valid pstates for the processor. This driver has subsequently been enhanced to add ACPI support.

The initial AMD Athlon 64 and AMD Opteron driver (developed by me, based upon Dave's earlier work, and with much input from Dominik and others), was also PSB based. This was followed by a version of the driver that added ACPI support.

The next release is intended to add a built-in table of pstates that will allow the checking of BIOS supplied data, and also allow an override capability to provide pstate data when not supplied by BIOS.

8.2 User Interface

The deprecated `/proc/cpufreq` (and `/proc/sys`) file system offers control over all processors or individual processors. By echoing values into this file, the root user can change policies and change the limits on available frequencies.

Examples:

Constrain all processors to frequencies between 1.0 GHz and 1.6 GHz, with the performance policy (effectively chooses 1.6 GHz):

```
echo -n "1000000:16000000:
performance" > /proc/cpufreq
```

Constrain processor 2 to run at only 2.0 GHz:

```
echo -n "2:2000000:2000000:
performance" > proc/cpufreq
```

The "performance" refers to a policy, with the other policy available being "powersave." These policies simply forced the frequency to be at the appropriate extreme of the available range. With the 2.6 kernel, the choice is normally for a "userspace" governor, which allows the (root) user or any user space code (running with root privilege) to dynamically control the frequency.

With the 2.6 kernel, a new interface in the `/sys` filesystem is available to the root user, deprecating the `/proc/cpufreq` method.

The control and status files exist under `/sys/devices/system/cpu/cpuN/cpufreq`, where N varies from 0 upwards, dependent on which processors are online. Among the other files in each processor's directory, `scaling_min_freq` and `scaling_max_freq` control the minimum and maximum of the ranges in which the frequency may vary. The `scaling_governor` file is used to control the choice of governor. See `linux/Documentation/cpu-freq/userguide.txt` for more information.

Examples:

Constrain processor 2 to run only in the range 1.6 GHz to 2.0 GHz:

```
cd /sys/devices/system/cpu

cd cpu2/cpufreq

echo 1600000 > scaling_min_freq

echo 2000000 > scaling_max_freq
```

8.3 Control From User Space And User Daemons

The interface to the `/sys` filesystem allows userland control and query functionality. Some form of automation of the policy would normally be part of the desired complete implementation.

This automation is dependent on the reason for using frequency management. As an example, for the case of transitioning to a lower pstate when running on a UPS, a daemon will be notified of the failure of mains power, and that daemon will trigger the frequency change by writing to the control files in the `/sys` filesystem.

The CPUFreq architecture has thus split the implementation into multiple parts:

1. user space policy
2. kernel space driver for common functionality
3. kernel space driver for processor specific implementation.

There are multiple user space automation implementations, not all of which currently support multiprocessor systems. One that does, and that has been used in this project is cpufreqd version 1.1.2 (`http://sourceforge.net/projects/cpufreqd`).

This daemon is controlled by a configuration file. Other than making changes to the configuration file, the author of this paper has not been involved in any of the development work on cpufreqd, and is a mere user of this tool.

The configuration file specifies profiles and rules. A profile is a description of the system settings in that state, and my configuration file is setup to map the profiles to the processor

pstates. Rules are used to dynamically choose which profile to use, and my rules are setup to transition profiles based on total processor load.

My simple configuration file to change processor frequency dependent on system load is:

```
[General]
pidfile=/var/run/cpufreqd.pid
poll_interval=2
pm_type=acpi

# 2.2 GHz processor speed
[Profile]
name=hi_boost
minfreq=95%
maxfreq=100%
policy=performance

# 2.0 GHz processor speed
[Profile]
name=medium_boost
minfreq=90%
maxfreq=93%
policy=performance

# 1.0 GHz processor Speed
[Profile]
name=lo_boost
minfreq=40%
maxfreq=50%
policy=powersave

[Profile]
name=lo_power
minfreq=40%
maxfreq=50%
policy=powersave

[Rule]
#not busy 0%-40%
name=conservative
ac=on
battery_interval=0-100
cpu_interval=0-40
profile=lo_boost

#medium busy 30%-80%
[Rule]
name=lo_cpu_boost
ac=on
battery_interval=0-100
cpu_interval=30-80
profile=medium_boost

#really busy 70%-100%
[Rule]
name=hi_cpu_boost
ac=on
battery_interval=50-100
cpu_interval=70-100
profile=hi_boost
```

This approach actually works very well for multiple small tasks, for transitioning the frequencies of all the processors together based on a collective loading statistic.

For a long running, single threaded task, this approach does not work well as the load is only high on a single processor, with the others being idle. The average load is thus low, and all processors are kept at a slow speed. Such a workload scenario would require an implementation that looked at the loading of individual processors, rather than the average. See the section below on future work.

8.4 The Drivers Involved

`powernow-k8.ko` `arch/i386/kernel/cpu/cpufreq/powernow-k8.c` (the same source code is built as a 32-bit driver in the `i386` tree and as a 64-bit driver in the `x86_64` tree)

`drivers/acpi`

`drivers/cpufreq`

The Test Driver

Note that the `powernow-k8.ko` driver does not export any read, write, or ioctl interfaces. For test purposes, a second driver exists with an ioctl interface for test application use. The test driver was a big part of the test effort on `powernow-k8.ko` prior to release.

8.5 Frequency Driver Entry Points

powernowk8_init()

Driver `late_initcall`. Initialization is late as the acpi driver needs to be initialized first. Verifies that all processors in the system are capable of frequency transitions, and that all processors are supported processors. Builds a data structure with the addresses of the four entry points for cpufreq use (listed below), and calls `cpufreq_register_driver()`.

powernowk8_exit()

Called when the driver is to be unloaded. Calls `cpufreq_unregister_driver()`.

8.6 Frequency Driver Entry Points For Use By The CPUFreq driver

powernowk8_cpu_init()

This is a per-processor initialization routine. As we are not guaranteed to be executing on the processor in question, and as the driver needs access to MSRs, the driver needs to force itself to run on the correct processor by using `set_cpus_allowed()`.

This pre-processor initialization allows for processors to be taken offline or brought online dynamically. I.e., this is part of the software support that would be needed for processor hotplug, although this is not supported in the hardware.

This routine finds the ACPI pstate data for this processor, and extracts the (proprietary) data from the ACPI `_PSS` objects. This data is verified as far as is reasonable. Per-processor data tables for use during frequency transitions are constructed from this information.

powernowk8_cpu_exit()

Per-processor cleanup routine.

powernowk8_verify()

When the root user (or an application running on behalf of the root user) requests a change to the minimum/maximum frequencies, or to the policy or governor, the frequency driver's verification routine is called to verify (and correct if necessary) the input values. For example, if the maximum speed of the processor is 2.4 GHz and the user requests that the maximum range be set to 3.0 GHz, the verify routine will correct the maximum value to a value that is actually possible. The user can, however, chose a value that is less than the hardware maximum, for example 2.0 GHz in this case.

As this routine just needs to access the per-processor data, and not any MSRs, it does not matter which processor executes this code.

powernowk8_target()

This is the driver entry point that actually performs a transition to a new frequency/voltage. This entry point is called for each processor that needs to transition to a new frequency.

There is therefore an optimization possible by enhancing the interface between the frequency driver and the CPUFreq driver for the case where all processors are to be transitioned to a new, common frequency. However, it is not clear that such an optimization is worth the complexity, as the functionality to transition a single processor would still be needed.

This routine is invoked with the processor

number as a parameter, and there is no guarantee as to which processor we are currently executing on. As the mechanism for changing the frequency involves accessing MSRs, it is necessary to execute on the target processor, and the driver forces its execution onto the target processor by using `set_cpus_allowed()`.

The CPUFreq helpers are then used to determine the correct target frequency. Once a chosen target `fid` and `vid` are identified:

- the cpufreq driver is called to warn that a transition is about to occur,
- the actual transition code within powernow-k8 is called, and then
- the cpufreq driver is called again to confirm that the transition was successful.

The actual transition is protected with a semaphore that is used across all processors. This is to prevent transitions on one processor from interfering with transitions on other processors. This is due to the inter-processor communication that occurs at a hardware level when a frequency transition occurs.

8.7 CPUFreq Interface

The CPUFreq interface provides entry points, that are required to make the system function.

It also provides helper functions, which need not be used, but are there to provide common functionality across the set of all architecture specific drivers. Elimination of duplicate good is a good thing! An architecture specific driver can build a table of available frequencies, and pass this table to the CPUFreq driver. The helper functions then simplify the architecture driver code by manipulating this table.

cpufreq_register_driver()

Registers the frequency driver as being the driver capable of performing frequency transitions on this platform. Only one driver may be registered.

cpufreq_unregister_driver()

Unregisters the driver, when it is being unloaded.

cpufreq_notify_transition()

Used to notify the CPUFreq driver, and thus the kernel, that a frequency transition is occurring, and triggering recalibration of timing specific code.

cpufreq_frequency_table_target()

Helper function to find an appropriate table entry for a given target frequency. Used in the driver's target function.

cpufreq_frequency_table_verify()

Helper function to verify that an input frequency is valid. This helper is effectively a complete implementation of the driver's verify function.

cpufreq_frequency_table_cpuinfo()

Supplies the frequency table data that is used on subsequent helper function calls. Also aids with providing information as to the capabilities of the processors.

8.8 Calls To The ACPI Driver

acpi_processor_register_performance()

acpi_processor_unregister_performance()

Helper functions used at per-processor initialization time to gain access to the data from the _PSS object for that processor. This is a preferable solution to the frequency driver having to walk the ACPI namespace itself.

8.9 The Single Processor Solution

Many of the kernel system calls collapse to constants when the kernel is built without multiprocessor support. For example, `num_online_cpus()` becomes a macro with the value 1. By the careful use of the definitions in smp.h, the same driver code handles both multiprocessor and single processor machines without the use of conditional compilation. The multiprocessor support obviously adds complexity to the code for a single processor code, but this code is negligible in the case of transitioning frequencies. The driver initialization and termination code is made more complex and lengthy, but this is not frequently executed code. There is also a small penalty in terms of code space.

The author does not feel that the penalty of the multiple processor support code is noticeable on a single processor system, but this is obviously debatable. The current choice is to have a single driver that supports both single processor and multiple processor systems.

As the primary performance cost is in terms of additional code space, it is true that a single processor machine with highly constrained memory may benefit from a simplified driver without the additional multi-processor support code. However, such a machine would see greater benefit by eliminating other code that would not be necessary on a chosen platform. For example, the PSB support code could be removed from a memory constrained single processor machine that was using ACPI.

This approach of removing code unnecessary for a particular platform is not a wonderful approach when it leads to multiple variants of the driver, all of which have to be supported and enhanced, and which makes Kconfig even more complex.

8.10 Stages Of Development, Test And Debug Of The Driver

The algorithm for transitioning to a new frequency is complex. See the BKDG[4] for a good description of the steps required, including flowcharts. In order to test and debug the frequency/voltage transition code thoroughly, the author first wrote a simple simulation of the processor. This simulation maintained a state machine, verified that fid/vid MSR control activity was legal, provided fid/vid status MSR results, and wrote a log file of all activity. The core driver code was then written as an application and linked with this simulation code to allow testing of all combinations.

The driver was then developed as a skeleton using printk to develop and test the BIOS/ACPI interfaces without having the frequency/voltage transition code present. This is because attempts to actually transition to an invalid pstate often result in total system lockups that offer no debug output—if the processor voltage is too low for the frequency, successful code execution ceases.

When the skeleton was working correctly, the actual transition code was dropped into place, and tested on real hardware, both single processor and multiple processor. (The single processor driver was released many months before the multi-processor capable driver as the multiprocessor capable hardware was not available in the marketplace.) The functional driver was tested, using printk to trace activity, and using external hardware to track power usage, and using a test driver to independently verify register settings.

The functional driver was then made available to various people in the community for their feedback. The author is grateful for the extensive feedback received, which included the changed code to implement suggestions. The driver as it exists today is considerably im-

proved from the initial release, due to this feedback mechanism.

9 How To Determine Valid PStates For A Given Processor

AMD defines pstates for each processor. A performance state is a frequency/voltage pair that is valid for operation of that processor. These are specified as fid/vid (frequency identifier/voltage identifier values) pairs, and are documented in the Processor Thermal and Data Sheets (see references). The worst case processor power consumption for each pstate is also characterized. The BKDG[4] contains tables for mapping fid to frequency and vid to voltage.

Pstates are processor specific. I.e., 2.0 GHz at 1.45V may be correct for one model/revision of processor, but is not necessarily correct for a different/revision model of processor.

Code can determine whether a processor supports or does not support pstate transitions by executing the cpuid instruction. (For details, see the BKDG[4] or the source code for the Linux frequency driver). This needs to be done for each processor in an MP system.

Each processor in an MP system could theoretically have different pstates.

Ideally, the processor frequency driver would not contain hardcoded pstate tables, as the driver would then need to be revised for new processor revisions. The chosen solution is to have the BIOS provide the tables of pstates, and have the driver retrieve the pstate data from the BIOS. There are two such tables defined for use by BIOSs for AMD systems:

1. PSB, AMD's original proprietary mechanism, which does not support MP. This mechanism is being deprecated.
2. ACPI _PSS objects. Whereas the ACPI specification is a standard, the data within the _PSS objects is AMD specific (and, in fact, processor family specific), and thus there is still a proprietary nature of this solution.

The current AMD frequency driver obtains data from the ACPI objects. ACPI does introduce some limitations, which are discussed later. Experimentation is ongoing with a built-in database approach to the problem in an attempt to bypass these issues, and also to allow checking of validity of the ACPI provided data.

10 ACPI And Frequency Restrictions

ACPI[5] provides the _PPC object, that is used to constrain the pstates available. This object is dynamic, and can therefore be used in platforms for purposes such as:

- forcing frequency restrictions when operating on battery power,
- forcing frequency restrictions due to thermal conditions.

For battery / mains power transitions, an ACPI-compliant GPE (General Purpose Event) input to the chipset (I/O hub) is dedicated to assigning a SCI (System Control Interrupt) when the power source changes. The ACPI driver will then execute the ACPI control method (see the _PSR power source ACPI object), which issues a notify to the _CPUn object, which triggers the ACPI driver to re-evaluate the _PPC object. If the current pstate exceeds that allowed by this new evaluation of the _PPC object, the CPU frequency driver will be called to transition to a lower pstate.

11 ACPI Issues

ACPI as a standard is not perfect. There is variation among different implementations, and Linux ACPI support does not work on all machines.

ACPI does introduce some overhead, and some users are not willing to enable ACPI.

ACPI requires that pstates be of equivalent power usage and frequency across all processors. In a system with processors that are capable of different maximum frequencies (for example, one processor capable of 2.0 GHz and a second processor capable of 2.2 GHz), compliance with the ACPI specification means that the faster processor(s) will be restricted to the maximum speed of the slowest processor. Also, if one processor has 5 available pstates, the presence of processor with only 4 available pstates will restrict all processors to 4 pstates.

12 What Is There Today?

AMD is shipping pstate capable AMD Opteron processors (revision CG). Server processors prior to revision CG were not pstate capable. All AMD Athlon 64 processors for mobile and desktop are pstate capable.

BKDG[4] enhancements to describe the capability are in progress.

AMD internal BIOSs have the enhancements. These enhancements are rolling out to the publicly available BIOSs along with the BKDG enhancements.

The multi-processor capable Linux frequency driver has released under GPL.

The cpufreqd user-mode daemon, available for download from `http://sourceforge.net/projects/cpufreqd` supports multiple processors.

13 Other Software-directed Power Saving Mechanisms

13.1 Use Of The `HLT` Instruction

The hlt instruction is normally used when the operating system has no code for the processor to execute. This is the ACPI C1 state. Execution of instructions ceases, until the processor is restarted with an interrupt. The power savings are maximized when the hlt state is entered in the minimum pstate, due to the lower voltage. The alternative to the use of the hlt instruction is a do nothing loop.

13.2 Use of Power Managed Chipset Drivers

Devices on the planar board, such as a PCI-X bridge or an AGP tunnel, may have the capability to operate in lower power modes. Entering and leaving the lower power modes is under the control of the driver for that device.

Note that HyperTransport attached devices can transition themselves to lower power modes when certain messages are seen on the bus. However, this functionality is typically configurable, so a chipset driver (or the system BIOS during bootup) would need to enable this capability.

14 Items For Future Exploration

14.1 A Built-in Database

The theory is that the driver could have a built-in database of processors and the pstates that they support. The driver could then use this database to obtain the pstate data without dependencies on ACPI, or use it for enhanced

checking of the ACPI provided data. The disadvantage of this is the need to update the database for new processor revisions. The advantages are the ability to overcome the ACPI imposed restrictions, and also to allow the use of the technology on systems where the ACPI support is not enabled.

14.2 Kernel Scheduler—CPU Power

An enhanced scheduler for the 2.6 kernel (2.6.6-bk1) is aware of groups of processors with different processing power. The power tating of each CPU group should be dynamically adjusted using a cpufreq transition notifier as the processor frequencies are changed.

See `http://lwn.net/Articles/80601/` for a detailed acount of the scheduler changes.

14.3 Thermal Management, ACPI Thermal Zones

Publicly available BIOSs for AMD machines do not implement thermal zones. Obviously this is one way to provide the input control for frequency management based on thermal conditions.

14.4 Thermal Management, Service Processor

Servers typically have a service processor, which may be compliant to the IPMI specification. This service processor is able to accurately monitor temperature at different locations within the chassis. The 2.6 kernel includes an IPMI driver. User space code may use these thermal readings to control fan speeds and generate administrator alerts. It may make sense to also use these accurate thermal readings to trigger frequency transitions.

The interaction between thermal events from the service processor and ACPI thermal zones may be a problem.

Hiding Thermal Conditions

One concern with the use of CPU frequency manipulation to avoid overheating is that hardware problems may not be noticed. Over temperature conditions would normally cause administrator alerts, but if the processor is first taken to a lower frequency to hold temperature down, then the alert may not be generated. A failing fan (not spinning at full speed) could therefore be missed. Some hardware components fail gradually, and early warning of imminent failures is needed to perform planned maintenance. Losing this data would be badness.

15 Legal Information

16 References

1. AMD Opteron™ Processor Data Sheet, publication 23932, available from `www.amd.com`
2. AMD Opteron™ Processor Power And

Thermal Data Sheet, publication 30417, available from www.amd.com

3. AMD Athlon™ 64 Processor Power And Thermal Data Sheet, publication 30430, available from www.amd.com

4. BIOS and Kernel Developer's Guide (the BKDG) for AMD Athlon™ 64 and AMD Opteron™ Processors, publication 26094, available from www.amd.com. Chapter 9 covers frequency management.

5. ACPI 2.0b Specification, from www.acpi.info

6. Text documentation files in the kernel linux/Documentation/cpu-freq/ directory:

   - index.txt
   - user-guide.txt
   - core.txt
   - cpu-drivers.txt
   - governors.txt

Dynamic Kernel Module Support: From Theory to Practice

Matt Domsch & Gary Lerhaupt
Dell Linux Engineering
Matt_Domsch@dell.com, Gary_Lerhaupt@dell.com

Abstract

DKMS is a framework which allows individual kernel modules to be upgraded without changing your whole kernel. Its primary audience is fourfold: system administrators who want to update a single device driver rather than wait for a new kernel from elsewhere with it included; distribution maintainers, who want to release a single targeted bugfix in between larger scheduled updates; system manufacturers who need single modules changed to support new hardware or to fix bugs, but do not wish to test whole new kernels; and driver developers, who must provide updated device drivers for testing and general use on a wide variety of kernels, as well as submit drivers to kernel.org.

Since OLS2003, DKMS has gone from a good idea to deployed and used. Based on end user feedback, additional features have been added: precompiled module tarball support to speed factory installation; driver disks for Red Hat distributions; 2.6 kernel support; SuSE kernel support. Planned features include cross-architecture build support and additional distribution driver disk methods.

In addition to overviewing DKMS and its features, we explain how to create a dkms.conf file to DKMS-ify your kernel module source.

1 History

Historically, Linux distributions bundle device drivers into essentially one large kernel package, for several primary reasons:

- Completeness: The Linux kernel as distributed on kernel.org includes all the device drivers packaged neatly together in the same kernel tarball. Distro kernels follow kernel.org in this respect.
- Maintainer simplicity: With over 4000 files in the kernel `drivers/` directory, each possibly separately versioned, it would be impractical for the kernel maintainer(s) to provide a separate package for each driver.
- Quality Assurance / Support organization simplicity: It is easiest to ask a user "what kernel version are you running," and to compare this against the list of approved kernel versions released by the QA team, rather than requiring the customer to provide a long and extensive list of package versions, possibly one per module.
- End user install experience: End users don't care about which of the 4000 possible drivers they need to install, they just want it to work.

This works well as long as you are able to make the "top of the tree" contain the most current

and most stable device driver, and you are able to convince your end users to always run the "top of the tree." The `kernel.org` development processes tend to follow this model with great success.

But widely used distros cannot ask their users to always update to the top of the kernel.org tree. Instead, they start their products from the top of the kernel.org tree at some point in time, essentially freezing with that, to begin their test cycles. The duration of these test cycles can be as short as a few weeks, and as long as a few years, but 3-6 months is not unusual. During this time, the kernel.org kernels march forward, and some (but not all) of these changes are backported into the distro's kernel. They then apply the minimal patches necessary for them to declare the product finished, and move the project into the sustaining phase, where changes are very closely scrutinized before releasing them to the end users.

1.1 Backporting

It is this sustaining phase that DKMS targets. DKMS can be used to backport newer device driver versions from the "top of the tree" kernels where most development takes place to the now-historical kernels of released products.

The `PATCH_MATCH` mechanism was specifically designed to allow the application of patches to a "top of the tree" device driver to make it work with older kernels. This allows driver developers to continue to focus their efforts on keeping kernel.org up to date, while allowing that same effort to be used on existing products with minimal changes. See Section 6 for a further explanation of this feature.

1.2 Driver developers' packaging

Driver developers have recognized for a long time that they needed to provide backported versions of their drivers to match their end users' needs. Often these requirements are imposed on them by system vendors such as Dell in support of a given distro release. However, each driver developer was free to provide the backport mechanism in any way they chose. Some provided architecture-specific RPMs which contained only precompiled modules. Some provided source RPMs which could be rebuilt for the running kernel. Some provided driver disks with precompiled modules. Some provided just source code patches, and expected the end user to rebuild the kernel themselves to obtain the desired device driver version. All provided their own Makefiles rather than use the kernel-provided build system.

As a result, different problems were encountered with each developers' solution. Some developers had not included their drivers in the kernel.org tree for so long that that there were discrepancies, e.g. `CONFIG_SMP` vs `__SMP__`, `CONFIG_2G` vs. `CONFIG_3G`, and compiler option differences which went unnoticed and resulted in hard-to-debug issues.

Needless to say, with so many different mechanisms, all done differently, and all with different problems, it was a nightmare for end users.

A new mechanism was needed to cleanly handle applying updated device drivers onto an end user's system. Hence DKMS was created as the one module update mechanism to replace all previous methods.

2 Goals

DKMS has several design goals.

- Implement only mechanism, not policy.
- Allow system administrators to easily know what modules, what versions, for

what kernels, and in what state, they have on their system.

- Keep module source as it would be found in the "top of the tree" on kernel.org. Apply patches to backport the modules to earlier kernels as necessary.

- Use the kernel-provided build mechanism. This reduces the Makefile magic that driver developers need to know, thus the likelihood of getting it wrong.

- Keep additional DKMS knowledge a driver developer must have to a minimum. Only a small per-driver dkms.conf file is needed.

- Allow multiple versions of any one module to be present on the system, with only one active at any given time.

- Allow DKMS-aware drivers to be packaged in the Linux Standard Base-conformant RPM format.

- Ease of use by multiple audiences: driver developers, system administrators, Linux distros, and system vendors.

We discuss DKMS as it applies to each of these four audiences.

3 Distributions

All present Linux distributions distribute device drivers bundled into essentially one large kernel package, for reasons outlined in Section 1. It makes the most sense, most of the time.

However, there are cases where it does not make sense.

- Severity 1 bugs are discovered in a single device driver between larger scheduled updates. Ideally you'd like your affected users to be able to get the single module update without having to release and Q/A a whole new kernel. Only customers who are affected by the particular bug need to update "off-cycle."

- Solutions vendors, for change control reasons, often certify their solution on a particular distribution, scheduled update release, and sometimes specific kernel version. The latter, combined with releasing device driver bug fixes as whole new kernels, puts the customer in the untenable position of either updating to the new kernel (and losing the certification of the solution vendor), or forgoing the bug fix and possibly putting their data at risk.

- Some device drivers are not (yet) included in kernel.org nor a distro kernel, however one may be required for a functional software solution. The current support models require that the add-on driver "taint" the kernel or in some way flag to the support organization that the user is running an unsupported kernel module. Tainting, while valid, only has three dimensions to it at present: Proprietary—non-GPL licensed; Forced—loaded via `insmod -f`; and Unsafe SMP—for some CPUs which are not designed to be SMP-capable. A GPL-licensed device driver which is not yet in kernel.org or provided by the distribution may trigger none of these taints, yet the support organization needs to be aware of this module's presence. To avoid this, we expect to see the distros begin to cryptographically sign kernel modules that they produce, and taint on load of an unsigned module. This would help reduce the support organization's work for calls about "unsupported"

configurations. With DKMS in use, there is less a need for such methods, as it's easy to see which modules have been changed.

Note: this is not to suggest that driver authors should not submit their drivers to `kernel.org`*—absolutely they should.*

- The distro QA team would like to test updates to specific drivers without waiting for the kernel maintenance team to rebuild the kernel package (which can take many hours in some cases). Likewise, individual end users may be willing (and often be required, e.g. if the distro QA team can't reproduce the users's hardware and software environment exactly) to show that a particular bug is fixed in a driver, prior to releasing the fix to *all* of that distro's users.

- New hardware support via driver disks: Hardware vendors release new hardware asynchronously to any software vendor schedule, no matter how hard companies may try to synchronize releases. OS distributions provide install methods which use driver diskettes to enable new hardware for previously-released versions of the OS. Generating driver disks has always been a difficult and error-prone procedure, different for each OS distribution, not something that the casual end-user would dare attempt.

DKMS was designed to address all of these concerns.

DKMS aims to provide a clear separation between mechanism (how one updates individual kernel modules and tracks such activity) and policy (when should one update individual kernel modules).

3.1 Mechanism

DKMS provides only the mechanism for updating individual kernel modules, not policy. As such, it can be used by distributions (per their policy) for updating individual device drivers for individual users affected by Severity 1 bugs, without releasing a whole new kernel.

The first mechanism critical to a system administrator or support organization is the `status` command, which reports the name, version, and state of each kernel module under DKMS control. By querying DKMS for this information, system administrators and distribution support organizations may quickly understand when an updated device driver is in use to speed resolution when issues are seen.

DKMS's ability to generate driver diskettes gives control to both novice and seasoned system administrators alike, as they can now perform work which otherwise they would have to wait for a support organization to do for them. They can get their new hardware systems up-and-running quickly by themselves, leaving the support organizations with time to do other more interesting value-added work.

3.2 Policy

Suggested policy items include:

- Updates must pass QA. This seems obvious, but it reduces broken updates (designed to fix other problems) from being released.

- Updates must be submitted, and ideally be included already, upstream. For this we expect kernel.org and the OS distribution to include the update in their next larger scheduled update. This ensures that when the next kernel.org kernel or distro update

comes out, the short-term fix provided via DKMS is incorporated already.

- The `AUTOINSTALL` mechanism is set to `NO` for all modules which are shipped with the target distro's kernel. This prevents the DKMS autoinstaller from installing a (possibly older) kernel module onto a newer kernel without being explicitly told to do so by the system administrator. This follows from the "all DKMS updates must be in the next larger release" rule above.

- All issues for which DKMS is used are tracked in the appropriate bug tracking databases until they are included upstream, and are reviewed regularly.

- All DKMS packages are provided as DKMS-enabled RPMs for easy installation and removal, per the Linux Standard Base specification.

- All DKMS packages are posted to the distro's support web site for download by system administrators affected by the partiular issue.

4 System Vendors

DKMS is useful to System Vendors such as Dell for many of the same reasons it's useful to the Linux distributions. In addition, system vendors face additional issues:

- Critical bug fixes for distro-provided drivers: While we hope to never need such, and we test extensively with distro-provided drivers, occasionally we have discovered a critical bug after the distribution has cut their gold CDs. We use DKMS to update just the affected device drivers.

- Alternate drivers: Dell occasionally needs to provide an alternate driver for a piece of hardware rather than that provided by the distribution natively. For example, Dell provides the Intel iANS network channel bonding and failover driver for customers who have used iANS in the past, and wish to continue using it rather than upgrading to the native channel bonding driver resident in the distribution.

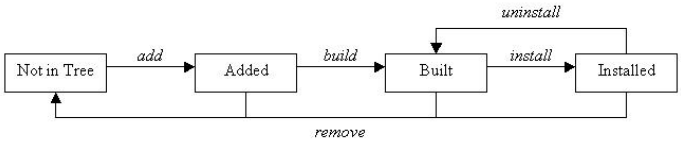
- Factory installation: Dell installs various OS distribution releases onto new hardware in its factories. We try not to update from the gold release of a distribution version to any of the scheduled updates, as customers expect to receive gold. We use DKMS to enable newer device drivers to handle newer hardware than was supported natively in the gold release, while keeping the gold kernel the same.

We briefly describe the policy Dell uses, in addition to the above rules suggested to OS distributions:

- Prebuilt DKMS tarballs are required for factory installation use, for all kernels used in the factory install process. This prevents the need for the compiler to be run, saving time through the factories. Dell rarely changes the factory install images for a given OS release, so this is not a huge burden on the DKMS packager.

- All DKMS packages are posted to support.dell.com for download by system administrators purchasing systems without Linux factory-installed.

Figure 1: DKMS state diagram.



5 System Administrators

5.1 Understanding the DKMS Life Cycle

Before diving into using DKMS to manage kernel modules, it is helpful to understand the life cycle by which DKMS maintains your kernel modules. In Figure 1, each rectangle represents a state your module can be in and each italicized word represents a DKMS action that can used to switch between the various DKMS states. In the following section we will look further into each of these DKMS actions and then continue on to discuss auxiliary DKMS functionality that extends and improves upon your ability to utilize these basic commands.

5.2 RPM and DKMS

DKMS was designed to work well with Red Hat Package Manger (RPM). Many times using DKMS to install a kernel module is as easy as installing a DKMS-enabled module RPM. Internally in these RPMs, DKMS is used to `add`, `build`, and `install` a module. By wrapping DKMS commands inside of an RPM, you get the benefits of RPM (package versioning, security, dependency resolution, and package distribution methodologies) while DKMS handles the work RPM does not, versioning and building of individual kernel modules. For reference, a sample DKMS-enabled RPM specfile can be found in the DKMS package.

5.3 Using DKMS

5.3.1 Add

DKMS manages kernel module versions at the source code level. The first requirement of using DKMS is that the module source be located on the build system and that it be located in the directory `/usr/src/<module>-<module-version>/`. It also requires that a dkms.conf file exists with the appropriately formatted directives within this configuration file to tell DKMS such things as where to install the module and how to build it. Once these two requirements have been met and DKMS has been installed on your system, you can begin using DKMS by adding a module/module-version to the DKMS tree. For example:

```
# dkms add -m megaraid2 -v 2.10.3
```

This example `add` command would add megaraid2/2.10.3 to the already existent `/var/dkms` tree, leaving it in the Added state.

5.3.2 Build

Once in the Added state, the module is ready to be built. This occurs through the DKMS `build` command and requires that the proper kernel sources are located on the system from the `/lib/module/<kernel-version>/build` symlink. The make command that is

used to compile the module is specified in the dkms.conf configuration file. Continuing with the megaraid2/2.10.3 example:

```
# dkms build -m megaraid2
   -v 2.10.3 -k 2.4.21-4.ELsmp
```

The `build` command compiles the module but stops short of installing it. As can be seen in the above example, `build` expects a kernel-version parameter. If this kernel name is left out, it assumes the currently running kernel. However, it functions perfectly well to build modules for kernels that are not currently running. This functionality is assured through use of a kernel preparation subroutine that runs before any module build is performed in order to ensure that the module being built is linked against the proper kernel symbols.

Successful completion of a `build` creates, for this example, the `/var/dkms/megaraid2/2.10.3/2.4.21-4.ELsmp/` directory as well as the log and module subdirectories within this directory. The log directory holds a log file of the module make and the module directory holds copies of the resultant binaries.

5.3.3 Install

With the completion of a `build`, the module can now be installed on the kernel for which it was built. Installation copies the compiled module binary to the correct location in the `/lib/modules/` tree as specified in the dkms.conf file. If a module by that name is already found in that location, DKMS saves it in its tree as an original module so at a later time it can be put back into place if the newer module is uninstalled. An example `install` command:

```
# dkms install -m megaraid2
   -v 2.10.3 -k 2.4.21-4.ELsmp
```

If a module by the same name is already installed, DKMS saves a copy in its tree and does so in the `/var/dkms/<module-name>/original_module/` directory. In this case, it would be saved to `/var/dkms/megaraid2/original_module/2.4.21-4.ELsmp/`.

5.3.4 Uninstall and Remove

To complete the DKMS cycle, you can also uninstall or remove your module from the tree. The `uninstall` command deletes from `/lib/modules` the module you installed and, if applicable, replaces it with its original module. In scenarios where multiple versions of a module are located within the DKMS tree, when one version is uninstalled, DKMS does not try to understand or assume which of these other versions to put in its place. Instead, if a true "original_module" was saved from the very first DKMS installation, it will be put back into the kernel and all of the other module versions for that module will be left in the Built state. An example `uninstall` would be:

```
# dkms uninstall -m megaraid2
   -v 2.10.3 -k 2.4.21-4.ELsmp
```

Again, if the kernel version parameter is unset, the currently running kernel is assumed, although, the same behavior does not occur with the `remove` command. The `remove` and `uninstall` are very similar in that `remove` will do all of the same steps as `uninstall`. However, when `remove` is employed, if the module-version being removed is the last instance of that module-version for all kernels on your system, after the uninstall portion of the remove completes, it will delete all traces of that module from the DKMS tree. To put it another way, when an `uninstall` command completes, your modules are left in the Built

state. However, when a `remove` completes, you would be left in the Not in Tree state. Here are two sample `remove` commands:

```
# dkms remove -m megaraid2
   -v 2.10.3 -k 2.4.21-4.ELsmp
# dkms remove -m megaraid2
   -v 2.10.3 --all
```

With the first example `remove` command, your module would be uninstalled and if this module/module-version were not installed on any other kernel, all traces of it would be removed from the DKMS tree all together. If, say, megaraid2/2.10.3 was also installed on the 2.4.21-4.ELhugemem kernel, the first `remove` command would leave it alone and it would remain intact in the DKMS tree. In the second example, that would not be the case. It would uninstall all versions of the megaraid2/2.10.3 module from all kernels and then completely expunge all references of megaraid2/2.10.3 from the DKMS tree. Thus, `remove` is what cleans your DKMS tree.

5.4 Miscellaneous DKMS Commands

5.4.1 Status

DKMS also comes with a fully functional status command that returns information about what is currently located in your tree. If no parameters are set, it will return all information found. Logically, the specificity of information returned depends on which parameters are passed to your status command. Each status entry returned will be of the state: "added," "built," or "installed," and if an original module has been saved, this information will also be displayed. Some example status commands include:

```
# dkms status
# dkms status -m megaraid2
# dkms status -m megaraid2 -v 2.10.3
# dkms status -k 2.4.21-4.ELsmp
# dkms status -m megaraid2
   -v 2.10.3 -k 2.4.21-4.ELsmp
```

5.4.2 Match

Another major feature of DKMS is the match command. The match command takes the configuration of DKMS installed modules for one kernel and applies this same configuration to some other kernel. When the match completes, the same module/module-versions that were installed for one kernel are also then installed on the other kernel. This is helpful when you are upgrading from one kernel to the next, but would like to keep the same DKMS modules in place for the new kernel. Here is an example:

```
# dkms match
   --templatekernel 2.4.21-4.ELsmp
   -k 2.4.21-5.ELsmp
```

As can be seen in the example, the `--templatekernel` is the "match-er" kernel from which the configuration is based, while the `-k kernel` is the "match-ee" upon which the configuration is instated.

5.4.3 dkms_autoinstaller

Similar in nature to the match command is the dkms_autoinstaller service. This service gets installed as part of the DKMS RPM in the /etc/init.d directory. Depending on whether an `AUTOINSTALL` directive is set within a module's dkms.conf configuration file, the dkms_autoinstaller will automatically build and install that module as you boot your system into new kernels which do not already have this module installed.

5.4.4 mkdriverdisk

The last miscellaneous DKMS command is `mkdriverdisk`. As can be inferred from its name, `mkdriverdisk` will take the proper

sources in your DKMS tree and create a driver disk image for use in providing updated drivers to Linux distribution installations. A sample `mkdriverdisk` might look like:

```
# dkms mkdriverdisk -d redhat
   -m megaraid2 -v 2.10.3
   -k 2.4.21-4.ELBOOT
```

Currently, the only supported distribution driver disk format is Red Hat. For more information on the extra necessary files and their formats for DKMS to create Red Hat driver disks, see `http://people.redhat.com/dledford`. These files should be placed in your module source directory.

5.5 Systems Management with DKMS Tarballs

As we have seen, DKMS provides a simple mechanism to build, install, and track device driver updates. So far, all these actions have related to a single machine. But what if you've got many similar machines under your administrative control? What if you have a compiler and kernel source on only one system (your master build system), but you need to deploy your newly built driver to all your other systems? DKMS provides a solution to this as well—in the `mktarball` and `ldtarball` commands.

The `mktarball` command rolls up copies of each device driver module file which you've built using DKMS into a compressed tarball. You may then copy this tarball to each of your target systems, and use the DKMS `ldtarball` command to load those into your DKMS tree, leaving each module in the Built state, ready to be installed. This avoids the need for both kernel source and compilers to be on every target system.

For example:

You have built the megaraid2 device driver, version 2.10.3, for two different kernel families (here 2.4.20-9 and 2.4.21-4.EL), on your master build system.

```
# dkms status
megaraid2, 2.10.3, 2.4.20-9: built
megaraid2, 2.10.3, 2.4.20-9bigmem: built
megaraid2, 2.10.3, 2.4.20-9BOOT: built
megaraid2, 2.10.3, 2.4.20-9smp: built
megaraid2, 2.10.3, 2.4.21-4.EL: built
megaraid2, 2.10.3, 2.4.21-4.ELBOOT: built
megaraid2, 2.10.3, 2.4.21-4.ELhugemem: built
megaraid2, 2.10.3, 2.4.21-4.ELsmp: built
```

You wish to deploy this version of the driver to several systems, without rebuilding from source each time. You can use the `mktarball` command to generate one tarball for each kernel family:

```
# dkms mktarball -m megaraid2
   -v 2.10.3
   -k 2.4.21-4.EL,2.4.21-4.ELsmp,
   2.4.21-4.ELBOOT,2.4.21-4.ELhugemem

Marking /usr/src/megaraid2-2.10.3 for archiving...
Marking kernel 2.4.21-4.EL for archiving...
Marking kernel 2.4.21-4.ELBOOT for archiving...
Marking kernel 2.4.21-4.ELhugemem for archiving...
Marking kernel 2.4.21-4.ELsmp for archiving...
Tarball location:
  /var/dkms/megaraid2/2.10.3/tarball/
megaraid2-2.10.3-manykernels.tgz
Done.
```

You can make one big tarball containing modules for both families by omitting the -k argument and kernel list; DKMS will include a module for every kernel version found.

You may then copy the tarball (renaming it if you wish) to each of your target systems using any mechanism you wish, and load the modules in. First, see that the target DKMS tree does not contain the modules you're loading:

```
# dkms status
Nothing found within the DKMS tree for
this status command. If your modules were
not installed with DKMS, they will not show
up here.
```

Then, load the tarball on your target system:

```
# dkms ldtarball
   --archive=megaraid2-2.10.3-manykernels.tgz

Loading tarball for module:
 megaraid2 / version: 2.10.3
Loading /usr/src/megaraid2-2.10.3...
Loading /var/dkms/megaraid2/2.10.3/2.4.21-4.EL...
Loading /var/dkms/megaraid2/2.10.3/2.4.21-4.ELBOOT...
Loading /var/dkms/megaraid2/2.10.3/2.4.21-4.ELhugemem...
Loading /var/dkms/megaraid2/2.10.3/2.4.21-4.ELsmp...
Creating /var/dkms/megaraid2/2.10.3/source symlink...
```

Finally, verify the modules are present, and in the Built state:

```
# dkms status
megaraid2, 2.10.3, 2.4.21-4.EL: built
megaraid2, 2.10.3, 2.4.21-4.ELBOOT: built
megaraid2, 2.10.3, 2.4.21-4.ELhugemem: built
megaraid2, 2.10.3, 2.4.21-4.ELsmp: built
```

DKMS `ldtarball` leaves the modules in the Built state, not the Installed state. For each kernel version you want your modules to be installed into, follow the install steps as above.

6 Driver Developers

As the maintainer of a kernel module, the only thing you need to do to get DKMS interoperability is place a small dkms.conf file in your driver source tarball. Once this has been done, any user of DKMS can simply do:

```
dkms ldtarball --archive /path/to/foo-1.0.tgz
```

That's it. We could discuss at length (which we will not rehash in this paper) the best methods to utilizing DKMS within a dkms-enabled module RPM, but for simple DKMS usability, the buck stops here. With the dkms.conf file in place, you have now positioned your source tarball to be usable by all manner and skill level of Linux users utilizing your driver. Effectively, you have widely increased your testing base without having to wade into package management or pre-compiled binaries. DKMS will handle this all for you. Along the same line, by leveraging DKMS you can now easily allow more widespread testing of your driver. Since driver versions can now be cleanly tracked outside of the kernel tree, you no longer must wait for the next kernel release in order for the community to register the necessary debugging cycles against your code. Instead, DKMS can be counted on to manage various versions of your kernel module such that any catastrophic errors in your code can be easily mitigated by a singular dkms uninstall command.

This leaves the composition of the dkms.conf as the only interesting piece left to discuss for the driver developer audience. With that in mind, we will now explicate over two dkms.conf examples ranging from that which is minimally required (Figure 2) to that which expresses maximal configuration (Figure 3).

6.1 Minimal dkms.conf for 2.4 kernels

Referring to Figure 2, the first thing that is distinguishable is the definition of the version of the package and the make command to be used to compile your module. This is only necessary for 2.4-based kernels, and lets the developer specify their desired make incantation.

Reviewing the rest of the dkms.conf, `PACKAGE_NAME` and `BUILT_MODULE_NAME[0]` appear to be duplicate in nature, but this is only the case for a package which contains only one kernel module within it. Had this example been for something like ALSA, the name of the package would be "alsa," but the `BUILT_MODULE_NAME` array would instead be populated with the names of the kernel modules within the ALSA package.

The final required piece of this minimal example is the `DEST_MODULE_LOCATION` array. This simply tells DKMS where in the /lib/modules tree it should install your module.

```
PACKAGE_NAME="megaraid2"
PACKAGE_VERSION="2.10.3"

MAKE[0]="make -C ${kernel_source_dir}
  SUBDIRS=${dkms_tree}/${PACKAGE_NAME}/${PACKAGE_VERSION}/build modules"

BUILT_MODULE_NAME[0]="megaraid2"
DEST_MODULE_LOCATION[0]="/kernel/drivers/scsi/"
```

Figure 2: A minimal dkms.conf

6.2 Minimal dkms.conf for 2.6 kernels

In the current version of DKMS, for 2.6 kernels the MAKE command listed in the dkms.conf is wholly ignored, and instead DKMS will always use:

```
make -C /lib/modules/$kernel_version/build \
  M=$dkms_tree/$module/$module_version/build
```

This jibes with the new external module build infrastructure supported by Sam Ravnborg's kernel Makefile improvements, as DKMS will always build your module in a build subdirectory it creates for each version you have installed. Similarly, an impending future version of DKMS will also begin to ignore the `PACKAGE_VERSION` as specified in dkms.conf in favor of the new modinfo provided information as implemented by Rusty Russell.

With regard to removing the requirement for `DEST_MODULE_LOCATION` for 2.6 kernels, given that similar information should be located in the install target of the Makefile provided with your package, it is theoretically possible that DKMS could one day glean such information from the Makefile instead. In fact, in a simple scenario as this example, it is further theoretically possible that the name of the package and of the built module could also be determined from the package Makefile. In effect, this would completely remove any need for a dkms.conf whatsoever, thus enabling all simple module tarballs to be automatically DKMS enabled.

Though, as these features have not been explored and as package maintainers would likely want to use some of the other dkms.conf directive features which are about to be elaborated upon, it is likely that requiring a dkms.conf will continue for the foreseeable future.

6.3 Optional dkms.conf directives

In the real-world version of the Dell's DKMS-enabled megaraid2 package, we also specify the optional directives:

```
MODULES_CONF_ALIAS_TYPE[0]=
   "scsi_hostadapter"
MODULES_CONF_OBSOLETES[0]=
   "megaraid,megaraid_2002"
REMAKE_INITRD="yes"
```

These directives tell DKMS to remake the kernel's initial ramdisk after every DKMS install or uninstall of this module. They further specify that before this happens, /etc/modules.conf (or /etc/sysconfig/kernel) should be edited intelligently so that the initrd is properly assembled. In this case, if /etc/modules.conf already contains a reference to either "megaraid" or "megaraid_2002," these will be switched to "megaraid2." If no such references are found,

then a new “scsi_hostadapter” entry will be added as the last such scsi_hostadapter number.

On the other hand, if it had also included:

```
MODULES_CONF_OBSOLETES_ONLY="yes"
```

then had no obsolete references been found, a new “scsi_hostadapter” line would not have been added. This would be useful in scenarios where you instead want to rely on something like Red Hat’s kudzu program for adding references for your kernel modules.

As well one could hypothetically also specify within the dkms.conf:

```
DEST_MODULE_NAME[0]="megaraid"
```

This would cause the resultant megaraid2 kernel module to be renamed to “megaraid” before being installed. Rather than having to propagate various one-off naming mechanisms which include the version as part of the module name in /lib/modules as has been previous common practice, DKMS could instead be relied upon to manage all module versioning to avoid such clutter. Was megaraid_2002 a version or just a special year in the hearts of the megaraid developers? While you and I might know the answer to this, it certainly confused Dell’s customers.

Continuing with hypothetical additions to the dkms.conf in Figure 2, one could also include:

```
BUILD_EXCLUSIVE_KERNEL="^2\.4.*"
BUILD_EXCLUSIVE_ARCH="i.86"
```

In the event that you know the code you produced is not portable, this is how you can tell DKMS to keep people from trying to build it elsewhere. The above restrictions would only allow the kernel module to be built on 2.4 kernels on x86 architectures.

Continuting with optional dkms.conf directives, the ALSA example in Figure 3 is taken directly from a DKMS-enabled package that Dell released to address sound issues on the Precision 360 workstation. It is slightly abridged as the alsa-driver as delivered actually installs 13 separate kernel modules, but for the sake of this example, only 9 are shown.

In this example, we have:

```
AUTOINSTALL="yes"
```

This tells the boot-time service dkms_autoinstaller that this package should be built and installed as you boot into a new kernel that DKMS has not already installed this package upon. By general policy, Dell only allows `AUTOINSTALL` to be set if the kernel modules are not already natively included with the kernel. This is to avoid the scenario where DKMS might automatically install over a newer version of the kernel module as provided by some newer version of the kernel. However, given the 2.6 modinfo changes, DKMS can now be modified to intelligently check the version of a native kernel module before clobbering it with some older version. This will likely result in a future policy change within Dell with regard to this feature.

In this example, we also have:

```
PATCH[0]="adriver.h.patch"
PATCH_MATCH[0]="2.4.[2-9][2-9]"
```

These two directives indicate to DKMS that if the kernel that the kernel module is being built for is >=2.4.22 (but still of the 2.4 family), the included adriver.h.patch should first be

```
PACKAGE_NAME="alsa-driver"
PACKAGE_VERSION="0.9.0rc6"

MAKE="sh configure --with-cards=intel8x0 --with-sequencer=yes \
  --with-kernel=/lib/modules/$kernelver/build \
  --with-moddir=/lib/modules/$kernelver/kernel/sound > /dev/null; make"
AUTOINSTALL="yes"

PATCH[0]="adriver.h.patch"
PATCH_MATCH[0]="2.4.[2-9][2-9]"

POST_INSTALL="alsa-driver-dkms-post.sh"
MODULES_CONF[0]="alias char-major-116 snd"
MODULES_CONF[1]="alias snd-card-0 snd-intel8x0"
MODULES_CONF[2]="alias char-major-14 soundcore"
MODULES_CONF[3]="alias sound-slot-0 snd-card-0"
MODULES_CONF[4]="alias sound-service-0-0 snd-mixer-oss"
MODULES_CONF[5]="alias sound-service-0-1 snd-seq-oss"
MODULES_CONF[6]="alias sound-service-0-3 snd-pcm-oss"
MODULES_CONF[7]="alias sound-service-0-8 snd-seq-oss"
MODULES_CONF[8]="alias sound-service-0-12 snd-pcm-oss"
MODULES_CONF[9]="post-install snd-card-0 /usr/sbin/alsactl restore >/dev/null 2>&1 || :"
MODULES_CONF[10]="pre-remove snd-card-0 /usr/sbin/alsactl store >/dev/null 2>&1 || :"

BUILT_MODULE_NAME[0]="snd-pcm"
BUILT_MODULE_LOCATION[0]="acore"
DEST_MODULE_LOCATION[0]="/kernel/sound/acore"

BUILT_MODULE_NAME[1]="snd-rawmidi"
BUILT_MODULE_LOCATION[1]="acore"
DEST_MODULE_LOCATION[1]="/kernel/sound/acore"

BUILT_MODULE_NAME[2]="snd-timer"
BUILT_MODULE_LOCATION[2]="acore"
DEST_MODULE_LOCATION[2]="/kernel/sound/acore"

BUILT_MODULE_NAME[3]="snd"
BUILT_MODULE_LOCATION[3]="acore"
DEST_MODULE_LOCATION[3]="/kernel/sound/acore"

BUILT_MODULE_NAME[4]="snd-mixer-oss"
BUILT_MODULE_LOCATION[4]="acore/oss"
DEST_MODULE_LOCATION[4]="/kernel/sound/acore/oss"

BUILT_MODULE_NAME[5]="snd-pcm-oss"
BUILT_MODULE_LOCATION[5]="acore/oss"
DEST_MODULE_LOCATION[5]="/kernel/sound/acore/oss"

BUILT_MODULE_NAME[6]="snd-seq-device"
BUILT_MODULE_LOCATION[6]="acore/seq"
DEST_MODULE_LOCATION[6]="/kernel/sound/acore/seq"

BUILT_MODULE_NAME[7]="snd-seq-midi-event"
BUILT_MODULE_LOCATION[7]="acore/seq"
DEST_MODULE_LOCATION[7]="/kernel/sound/acore/seq"

BUILT_MODULE_NAME[8]="snd-seq-midi"
BUILT_MODULE_LOCATION[8]="acore/seq"
DEST_MODULE_LOCATION[8]="/kernel/sound/acore/seq"

BUILT_MODULE_NAME[9]="snd-seq"
BUILT_MODULE_LOCATION[9]="acore/seq"
DEST_MODULE_LOCATION[9]="/kernel/sound/acore/seq"
```

Figure 3: An elaborate dkms.conf

applied to the module source before a module build occurs. In this way, by including various patches needed for various kernel versions, you can distribute one source tarball and ensure it will always properly build regardless of the end user target kernel. If no corresponding `PATCH_MATCH[0]` entry were specified for `PATCH[0]`, then the adriver.h.patch would always get applied before a module build. As DKMS always starts off each module build with pristine module source, you can always ensure the right patches are being applied.

Also seen in this example is:

```
MODULES_CONF[0]=
  "alias char-major-116 snd"
MODULES_CONF[1]=
  "alias snd-card-0 snd-intel8x0"
```

Unlike the previous discussion of /etc/modules.conf changes, any entries placed into the `MODULES_CONF` array are automatically added into /etc/modules.conf during a module install. These are later only removed during the final module uninstall.

Lastly, we have:

```
POST_INSTALL="alsa-driver-dkms-post.sh"
```

In the event that you have other scripts that must be run during various DKMS events, DKMS includes `POST_ADD`, `POST_BUILD`, `POST_INSTALL` and `POST_REMOVE` functionality.

7 Future

As you can tell from the above, DKMS is very much ready for deployment now. However, as with all software projects, there's room for improvement.

7.1 Cross-Architecture Builds

DKMS today has no concept of a platform architecture such as i386, x86_64, ia64, sparc, and the like. It expects that it is building kernel modules with a native compiler, not a cross compiler, and that the target architecture is the native architecture. While this works in practice, it would be convenient if DKMS were able to be used to build kernel modules for non-native architectures.

Today DKMS handles the cross-architecture build process by having separate /var/dkms directory trees for each architecture, and using the `dkmstree` option to specify a using a different tree, and the `config` option to specify to use a different kernel configuration file.

Going forward, we plan to add an `--arch` option to DKMS, or have it glean it from the kernel config file and act accordingly.

7.2 Additional distribution driver disks

DKMS today supports generating driver disks in the Red Hat format only. We recognize that other distributions accomplish the same goal using other driver disk formats. This should be relatively simple to add once we understand what the additional formats are.

8 Conclusion

DKMS provides a simple and unified mechanism for driver authors, Linux distributions, system vendors, and system administrators to update the device drivers on a target system without updating the whole kernel. It allows driver developers to keep their work aimed at the "top of the tree," and to backport that work to older kernels painlessly. It allows Linux distributions to provide updates to single device drivers asynchronous to the release of a larger

scheduled update, and to know what drivers have been updated. It lets system vendors ship newer hardware than was supported in a distribution's "gold" release without invalidating any test or certification work done on the "gold" release. It lets system administrators update individual drivers to match their environment and their needs, regardless of whose kernel they are running. It lets end users track which module versions have been added to their system.

We believe DKMS is a project whose time has come, and encourage everyone to use it.

9 References

DKMS is licensed under the GNU General Public License. It is available at

`http://linux.dell.com/dkms/`,

and has a mailing list `dkms-devel@lists.us.dell.com` to which you may subscribe at `http://lists.us.dell.com/`.

e100 Weight Reduction Program

Writing for Maintainability

Scott Feldman
Intel Corporation
scott.feldman@intel.com

Abstract

Corporate-authored device drivers are bloated/buggy with dead code, HW and OS abstraction layers, non-standard user controls, and support for complicated HW features that provide little or no value. e100 in 2.6.4 has been rewritten to address these issues and in the process lost 75% of the lines of code, with no loss of functionality. This paper gives guidelines to other corporate driver authors.

Introduction

This paper gives some basic guidelines to corporate device driver maintainers based on experiences I had while re-writing the e100 network device driver for Intel's PRO/100+ Ethernet controllers. By corporate maintainer, I mean someone employed by a corporation to provide Linux driver support for that corporation's device. Of course, these guidelines may apply to non-corporate individuals as well, but the intended audience is the corporate driver author.

The assumption behind these guidelines is that the device driver is intended for inclusion in the Linux kernel. For a driver to be accepted into the Linux kernel, it must meet both technical and non-technical requirements. This paper focuses on the non-technical requirements, specifically maintainability.

Guideline #1: Maintainability over Everything Else

Corporate marketing requirements documents specify priority order to features and performance and schedule (time-to-market), but rarely specify maintainability. However, maintainability is the *most* important requirement for Linux kernel drivers.

Why?

- You will not be the long-term driver maintainer.
- Your company will not be the long-term driver maintainer.
- Your driver will out-live your interest in it.

Driver code should be written so a like-skilled kernel maintainer can fix a problem in a reasonable amount of time without you or your resources. Here are a few items to keep in mind to improve maintainability.

- Use kernel coding style over corporate coding style
- Document how the driver/device works, at a high level, in a "Theory of Operation" comment section

old driver v2	new driver v3
VLANs tagging/ stripping	use SW VLAN support in kernel
Tx/Rx checksum offloading	use SW checksum support in kernel
interrupt moderation	use NAPI support in kernel

Table 1: Feature migration in e100

- Document hardware workarounds

Guideline #2: Don't Add Features for Feature's Sake

Consider the code complexity to support the feature versus the user's benefit. Is the device still usable without the feature? Is the device performing reasonably for the 80% use-case without the feature? Is the hardware offload feature working against ever increasing CPU/memory/IO speeds? Is there a software equivalent to the feature already provided in the OS?

If the answer is yes to any of these questions, it is better to not implement the feature, keeping the complexity in the driver low and maintainability high.

Table 1 shows features removed from the driver during the re-write of e100 because the OS already provides software equivalents.

Guideline #3: Limit User-Controls—Use What's Built into the OS

Most users will use the default settings, so before adding a user-control, consider:

1. If the driver model for your device class already provides a mechanism for the user-control, enable that support in the driver rather than adding a custom user-control.
2. If the driver model doesn't provide a user-control, but the user-control is potentially useful to other drivers, extend the driver model to include user-control.
3. If the user-control is to enable/disable a workaround, enable the workaround without the use of a user-control. (Solve the problem without requiring a decision from the user).
4. If the user-control is to tune performance, tune the driver for the 80% use-case and remove the user-control.

old driver v2	new driver v3
BundleMax	not needed – NAPI
BundleSmallFr	not needed – NAPI
IntDelay	not needed – NAPI
ucode	not needed – NAPI
RxDescriptors	ethtool -G
TxDescriptors	ethtool -G
XsumRX	not needed – checksum in OS
IFS	always enabled
e100_speed_duplex	ethtool -s

Table 2: User-control migration in e100

Table 2 shows user-controls (implemented as module parameters) removed from the driver during the re-write of e100 because the OS already provides built-in user-controls, or the user-control was no longer needed.

Guideline #4: Don't Write Code that's Already in the Kernel

Look for library code that's already used by other drivers and adapt that to your driver. Common hardware is often used between vendors' devices, so shared code will work for all (and be debugged by all).

For example, e100 has a highly MDI-compliant PHY interface, so use `mii.c` for standard PHY access and remove custom code from the driver.

For another example, e100 v2 used `/proc/net/IntelPROAdapter` to report driver information. This functionality was replaced with `ethtool`, `sysfs`, `lspci`, etc.

Look for opportunities to move code out of the driver into generic code.

Guideline #5: Don't Use OS-abstraction Layers

A common corporate design goal is to reuse driver code as much as possible between OSes. This allows a driver to be brought up on one OS and "ported" to another OS with little work. After all, the hardware interface to the device didn't change from one OS to the next, so all that is required is an OS-abstraction layer that wraps the OS's native driver model with a generic driver model. The driver is then written to the generic driver model and it's just a matter of porting the OS-abstraction layer to each target OS.

There are problems when doing this with Linux:

1. The OS-abstraction wrapper code means nothing to an outside Linux maintainer and just obfuscates the real meaning behind the code. This makes your code harder to follow and therefore harder to maintain.

2. The generic driver model may not map 1:1 with the native driver model leaving gaps in compatibility that you'll need to fix up with OS-specific code.

3. Limits your ability to back-port contributions given under GPL to non-GPL OSes.

Guideline #6: Use kcompat Techniques to Move Legacy Kernel Support out of the Driver (and Kernel)

Users may not be able to move to the latest `kernel.org` kernel, so there is a need to provide updated device drivers that can be installed against legacy kernels. The need is driven by 1) bug fixes, 2) new hardware support that wasn't included in the driver when the driver was included in the legacy kernel.

The best strategy is to:

1. Maintain your driver code to work against the latest `kernel.org` development kernel API. This will make it easier to keep the driver in the `kernel.org` kernel synchronized with your code base as changes (patches) are almost always in reference to the latest `kernel.org` kernel.

2. Provide a kernel-compat-layer (kcompat) to translate the latest API to the supported legacy kernel API. The driver code is void of any `ifdef` code for legacy kernel support. All of the `ifdef` logic moves to the kcompat layer. The kcompat layer is not included in the latest `kernel.org` kernel (by definition).

Here is an example with e100.

In driver code, use the latest API:

```
s = pci_name(pdev);
...
free_netdev(netdev);
```

In kcompat code, translate to legacy kernel API:

```
#if ( LINUX_VERSION_CODE < \
      KERNEL_VERSION(2,4,22) )
#define pci_name(x)  ((x)->slot_name)
#endif

#ifndef HAVE_FREE_NETDEV
#define free_netdev(x) kfree(x)
#endif
```

Guideline #7: Plan to Re-write the Driver at Least Once

You will not get it right the first time. Plan on rewriting the driver from scratch at least once. This will cleanse the code, removing dead code and organizing/consolidating functionality.

For example, the last e100 re-write reduced the driver size by 75% without loss of functionality.

Conclusion

Following these guidelines will result in more maintainable device drivers with better acceptance into the Linux kernel tree. The basic idea is to remove as much as possible from the driver without loss of functionality.

References

- The latest e100 driver code is available at `linux/driver/net/e100.c` (2.6.4 kernel or higher).
- An example of kcompat is here: `http://sf.net/projects/gkernel`

NFSv4 and `rpcsec_gss` for linux

J. Bruce Fields
University of Michigan
bfields@umich.edu

Abstract

The 2.6 Linux kernels now include support for version 4 of NFS. In addition to built-in locking and ACL support, and features designed to improve performance over the Internet, NFSv4 also mandates the implementation of strong cryptographic security. This security is provided by rpcsec_gss, a standard, widely implemented protocol that operates at the rpc level, and hence can also provide security for NFS versions 2 and 3.

1 The rpcsec_gss protocol

The rpc protocol, which all version of NFS and related protocols are built upon, includes generic support for authentication mechanisms: each rpc call has two fields, the credential and the verifier, each consisting of a 32-bit integer, designating a "security flavor," followed by 400 bytes of opaque data whose structure depends on the specified flavor. Similarly, each reply includes a single "verifier."

Until recently, the only widely implemented security flavor has been the auth_unix flavor, which uses the credential to pass uid's and gid's and simply asks the server to trust them. This may be satisfactory given physical security over the clients and the network, but for many situations (including use over the Internet), it is inadequate.

Thus rfc 2203 defines the rpcsec_gss protocol, which uses rpc's opaque security fields to carry cryptographically secure tokens. The cryptographic services are provided by the GSS-API ("Generic Security Service Application Program Interface," defined by rfc 2743), allowing the use of a wide variety of security mechanisms, including, for example, Kerberos.

Three levels of security are provided by rpcsec_gss:

1. Authentication only: The rpc header of each request and response is signed.
2. Integrity: The header and body of each request and response is signed.
3. Privacy: The header of each request is signed, and the body is encrypted.

The combination of a security level with a GSS-API mechanism can be designated by a 32-bit "pseudoflavor." The mount protocol used with NFS versions 2 and 3 uses a list of pseudoflavors to communicate the security capabilities of a server. NFSv4 does not use pseudoflavors on the wire, but they are still useful in internal interfaces.

Security protocols generally require some initial negotiation, to determine the capabilities of the systems involved and to choose session keys. The rpcsec_gss protocol uses calls with procedure number 0 for this purpose. Normally such a call is a simple "ping" with no side-effects, useful for measuring round-trip

latency or testing whether a certain service is running. However a call with procedure number 0, if made with authentication flavor rpcsec_gss, may use certain fields in the credential to indicate that it is part of a context-initiation exchange.

2 Linux implementation of rpcsec_gss

The Linux implementation of rpcsec_gss consists of several pieces:

1. Mechanism-specific code, currently for two mechanisms: krb5 and spkm3.
2. A stripped-down in-kernel version of the GSS-API interface, with an interface that allows mechanism-specific code to register support for various pseudoflavors.
3. Client and server code which uses the GSS-API interface to encode and decode rpc calls and replies.
4. A userland daemon, gssd, which performs context initiation.

2.1 Mechanism-specific code

The NFSv4 RFC mandates the implementation (though not the use) of three GSS-API mechanisms: krb5, spkm3, and lipkey.

Our krb5 implementation supports three pseudoflavors: krb5, krb5i, and krb5p, providing authentication only, integrity, and privacy, respectively. The code is derived from MIT's Kerberos implementation, somewhat simplified, and not currently supporting the variety of encryption algorithms that MIT's does. The krb5 mechanism is also supported by NFS implementations from Sun, Network Appliance, and others, which it interoperates with.

The Low Infrastructure Public Key Mechanism ("lipkey," specified by rfc 2847), is a public key mechanism built on top of the Simple Public Key Mechanism (spkm), which provides functionality similar to that of TLS, allowing a secure channel to be established using a server-side certificate and a client-side password.

We have a preliminary implementation of spkm3 (without privacy), but none yet of lipkey. Other NFS implementors have not yet implemented either of these mechanisms, but there appears to be sufficient interest from the grid community for us to continue implementation even if it is Linux-only for now.

2.2 GSS-API

The GSS-API interface as specified is very complex. Fortunately, rpcsec_gss only requires a subset of the GSS-API, and even less is required for per-packet processing.

Our implementation is derived by the implementation in MIT Kerberos, and initially stayed fairly close the the GSS-API specification; but over time we have pared it down to something quite a bit simpler.

The kernel gss interface also provides APIs by which code implementing particular mechanisms can register itself to the gss-api code and hence can be safely provided by modules loaded at runtime.

2.3 RPC code

The RPC code has been enhanced by the addition of a new rpcsec_gss mechanism which authenticates calls and replies and which wraps and unwraps rpc bodies in the case of integrity and privacy.

This is relatively straightforward, though somewhat complicated by the need to handle discontiguous buffers containing page data.

Caches for session state are also required on both client and server; on the client a preexisting rpc credentials cache is used, and on the server we use the same caching infrastructure used for caching of client and export information.

2.4 Userland daemon

We had no desire to put a complete implementation of Kerberos version 5 or the other mechanisms into the kernel. Fortunately, the work performed by the various GSS-API mechanisms can be divided neatly into context initiation and per-packet processing. The former is complex and is performed only once per session, while the latter is simple by comparison and needs to be performed on every packet. Therefore it makes sense to put the packet processing in the kernel, and have the context initiation performed in userspace.

Since it is the kernel that knows when context initiation is necessary, we require a mechanism allowing the kernel to pass the necessary parameters to a userspace daemon whenever context initiation is needed, and allowing the daemon to respond with the completed security context.

This problem was solved in different ways on the client and server, but both use special files (the former in a dedicated filesystem, rpc_pipefs, and the latter in the proc filesystem), which our userspace daemon, gssd, can poll for requests and then write responses back to.

In the case of Kerberos, the sequence of events will be something like this:

1. The user gets Kerberos credentials using kinit, which are cached on a local filesystem.

2. The user attempts to perform an operation on an NFS filesystem mounted with krb5 security.

3. The kernel rpc client looks for the a security context for the user in its cache; not finding any, it does an upcall to gssd to request one.

4. Gssd, on receiving the upcall, reads the user's Kerberos credentials from the local filesystem and uses them to construct a null rpc request which it sends to the server.

5. The server kernel makes an upcall which passes the null request to its gssd.

6. At this point, the server gssd has all it needs to construct a security context for this session, consisting mainly of a session key. It passes this context down to the kernel rpc server, which stores it in its context cache.

7. The server's gssd then constructs the null rpc reply, which it gives to the kernel to return to the client gssd.

8. The client gssd uses this reply to construct its own security context, and passes this context to the kernel rpc client.

9. The kernel rpc client then uses this context to send the first real rpc request to the server.

10. The server uses the new context in its cache to verify the rpc request, and to compose its reply.

3 The NFSv4 protocol

While rpcsec_gss works equally well on all existing versions of NFS, much of the work on rpcsec_gss has been motivated by NFS version 4, which is the first version of NFS to make rpcsec_gss mandatory to implement.

This new version of NFS is specified by rfc 3530, which says:

"Unlike earlier versions, the NFS version 4 protocol supports traditional file access while integrating support for file locking and the mount protocol. In addition, support for strong security (and its negotiation), compound operations, client caching, and internationalization have been added. Of course, attention has been applied to making NFS version 4 operate well in an Internet environment."

Descriptions of some of these features follow, with some notes about their implementation in Linux.

3.1 Compound operations

Each rpc request includes a procedure number, which describes the operation to be performed. The format of the body of the rpc request (the arguments) and of the reply depend on the program number. Procedure 0 is reserved as a no-op (except when it is used for rpcsec_gss context initiation, as described above).

The NFSv4 protocol only supports one non-zero procedure, procedure 1, the compound procedure.

The body of a compound is a list of operations, each with its own arguments. For example, a compound request performing a lookup might consist of 3 operations: a PUTFH, with a filehandle, which sets the "current filehandle" to the provided filehandle; a LOOKUP, with a name, which looks up the name in the directory given by the current filehandle and then modifies the current filehandle to be the filehandle of the result; a GETFH, with no arguments, which returns the new value of the current filehandle; and a GETATTR, with a bitmask specifying a set of attributes to return for the looked-up file.

The server processes these operations in order, but with no guarantee of atomicity. On encountering any error, it stops and returns the results of the operations up to and including the operation that failed.

In theory complex operations could therefore be done by long compounds which perform complex series of operations.

In practice, the compounds sent by the Linux client correspond very closely to NFSv2/v3 procedures—the VFS and the POSIX filesystem API make it difficult to do otherwise—and our server, like most NFSv4 servers we know of, rejects overly long or complex compounds.

3.2 Well-known port for NFS

RPC allows services to be run on different ports, using the "portmap" service to map program numbers to ports. While flexible, this system complicates firewall management; so NFSv4 recommends the use of port 2049.

In addition, the use of sideband protocols for mounting, locking, etc. also complicates firewall management, as multiple connections to multiple ports are required for a single NFS mount. NFSv4 eliminates these extra protocols, allowing all traffic to pass over a single connection using one protocol.

3.3 No more mount protocol

Earlier versions of NFS use a separate protocol for mount. The mount protocol exists primarily to map path names, presented to the server as

strings, to filehandles, which may then be used in the NFS protocol.

NFSv4 instead uses a single operation, PUTROOTFH, that returns a filehandle; clients can then use ordinary lookups to traverse to the filesystem they wish to mount. This changes the behavior of NFS in a few subtle ways: for example, the special status of mounts in the old protocol meant that mounting `/usr` and then looking up `local` might get you a different object than would mounting `/usr/local`; under NFSv4 this can no longer happen.

A server that exports multiple filesystems must knit them together using a single "pseudofilesystem" which links them to a common root.

On Linux's nfsd the pseudofilesystem is a real filesystem, marked by the export option "fsid=0". An adminstrator that is content to export a single filesystem can export it with "fsid=0", and clients will find it just by mounting the path "/".

The expected use for "fsid=0", however, is to designate a filesystem that is used just a collection of empty directories used as mountpoints for exported filesystems, which are mounted using `mount ---bind`; thus an administrator could export `/bin` and `/local/src` by:

```
mkdir -p /exports/home
mkdir -p /exports/bin/
mount --bind /home /exports/home
mount --bind /bin/ /exports/bin
```

and then using an exports file something like:

```
/exports *.foo.com(fsid=0,crossmnt)
/exports/home *.foo.com
/exports/bin *.foo.com
```

Clients in `foo.com` can then mount `server.foo.com:/bin` or `server.foo.com:/home`. However the relationship between the original mountpoint on the server and the mountpoint under `/exports` (which determines the path seen by the client) is arbitrary, so the administrator could just as well export `/home` as `/some/other/path` if desired.

This gives maximum flexibility at the expense of some confusion for adminstrators used to earlier NFS versions.

3.4 No more lock protocol

Locking has also been absorbed into the NFSv4 protocol. In addition to advantages enumerated above, this allows servers to support mandatory locking if desired. Previously this was impossible because it was impossible to tell whether a given read or write should be ordered before or after a lock request. NFSv4 enforces such sequencing by providing a stateid field on each read or write which identifies the locking state that the operation was performed under; thus for example a write that occurred while a lock was held, but that appeared on the server to have occurred after an unlock, can be identified as belonging to a previous locking context, and can therefore be correctly rejected.

The additional state required to manage locking is the source of much of the additional complexity in NFSv4.

3.5 String representations of user and group names

Previous versions of NFS use integers to represent users and groups; while simple to handle, they can make NFS installations difficult to manage, particularly across adminstrative domains. Version 4, therefore, uses string names of the form `user@domain`.

This poses some challenges for the kernel im-

plementation. In particular, while the protocol may use string names, the kernel still needs to deal with uid's, so it must map between NFSv4 string names and integers.

As with rpcsec_gss context initation, we solve this problem by making upcalls to a userspace daemon; with the mapping in userspace, it is easy to use mechanisms such as NIS or LDAP to do the actual mapping without introducing large amounts of code into the kernel. So as not to degrade performance by requiring a context switch every time we process a packet carrying a name, we cache the results of this mapping in the kernel.

3.6 Delegations

NFSv4, like previous versions of NFS, does not attempt to provide full cache consistency. Instead, all that is guaranteed is that if an open follows a close of the same file, then data read after the open will reflect any modifications performed before the close. This makes both open and close potentially high latency operations, since they must wait for at least one round trip before returning–in the close case, to flush out any pending writes, and in the open case, to check the attributes of the file in question to determine whether the local cache should be invalidated.

Locks provide similar semantics—writes are flushed on unlock, and cache consistency is verified on lock—and hence lock operations are also prone to high latencies.

To mitigate these concerns, and to encourage the use of NFS's locking features, delegations have been added to NFSv4. Delegations are granted or denied by the server in response to open calls, and give the client the right to perform later locks and opens locally, without the need to contact the server. A set of callbacks is provided so that the server can notify the client when another client requests an open that would confict with the open originally obtained by the client.

Thus locks and opens may be performed quickly by the client in the common case when files are not being shared, but callbacks ensure that correct close-to-open (and unlock-to-lock) semantics may be enforced when there is contention.

To allow other clients to proceed when a client holding a delegation reboots, clients are required to periodically send a "renew" operation to the server, indicating that it is still alive; a client that fails to send a renew operation within a given lease time (established when the client first contacts the server) may have all of its delegations and other locking state revoked.

Most implementations of NFSv4 delegations, including Linux's, are still young, and we haven't yet gathered good data on the performance impact.

Nevertheless, further extensions, including delegations over directories, are under consideration for future versions of the protocol.

3.7 ACLs

ACL support is integrated into the protocol, with ACLs that are more similar to those found in NT than to the POSIX ACLs supported by Linux.

Thus while it is possible to translate an arbitrary Linux ACL to an NFS4 ACL with nearly identical meaning, most NFS ACLs have no reasonable representation as Linux ACLs.

Marius Eriksen has written a draft describing the POSIX to NFS4 ACL translation. Currently the Linux implementation uses this mapping, and rejects any NFS4 ACL that isn't exactly in the image of this mapping. This en-

sures userland support from all tools that currently support POSIX ACLs, and simplifies ACL management when an exported filesystem is also used by local users, since both nfsd and the local users can use the backend filesystem's POSIX ACL implementation. However it makes it difficult to interoperate with NFSv4 implementations that support the full ACL protocol. For that reason we will eventually also want to add support for NFSv4 ACLs.

4 Acknowledgements and Further Information

This work has been sponsored by Sun Microsystems, Network Appliance, and the Accelerated Strategic Computing Initiative (ASCI). For further information, see `www.citi.umich.edu/projects/nfsv4/`.

Comparing and Evaluating epoll, select, and poll Event Mechanisms

Louay Gammo, Tim Brecht, Amol Shukla, and David Pariag
University of Waterloo
{lgammo,brecht,ashukla,db2pariag}@cs.uwaterloo.ca

Abstract

This paper uses a high-performance, event-driven, HTTP server (the μserver) to compare the performance of the select, poll, and epoll event mechanisms. We subject the μserver to a variety of workloads that allow us to expose the relative strengths and weaknesses of each event mechanism.

Interestingly, initial results show that the select and poll event mechanisms perform comparably to the epoll event mechanism in the absence of idle connections. Profiling data shows a significant amount of time spent in executing a large number of `epoll_ctl` system calls. As a result, we examine a variety of techniques for reducing `epoll_ctl` overhead including edge-triggered notification, and introducing a new system call (`epoll_ctlv`) that aggregates several `epoll_ctl` calls into a single call. Our experiments indicate that although these techniques are successful at reducing `epoll_ctl` overhead, they only improve performance slightly.

1 Introduction

The Internet is expanding in size, number of users, and in volume of content, thus it is imperative to be able to support these changes with faster and more efficient HTTP servers.

A common problem in HTTP server scalability is how to ensure that the server handles a large number of connections simultaneously without degrading the performance. An event-driven approach is often implemented in high-performance network servers [14] to multiplex a large number of concurrent connections over a few server processes. In event-driven servers it is important that the server focuses on connections that can be serviced without blocking its main process. An event dispatch mechanism such as `select` is used to determine the connections on which forward progress can be made without invoking a blocking system call. Many different event dispatch mechanisms have been used and studied in the context of network applications. These mechanisms range from `select`, `poll`, `/dev/poll`, RT signals, and epoll [2, 3, 15, 6, 18, 10, 12, 4].

The epoll event mechanism [18, 10, 12] is designed to scale to larger numbers of connections than `select` and `poll`. One of the problems with `select` and `poll` is that in a single call they must both inform the kernel of all of the events of interest and obtain new events. This can result in large overheads, particularly in environments with large numbers of connections and relatively few new events occurring. In a fashion similar to that described by Banga et al. [3] epoll separates mechanisms for obtaining events (`epoll_wait`) from those used to declare and control interest

in events (`epoll_ctl`).

Further reductions in the number of generated events can be obtained by using edge-triggered epoll semantics. In this mode events are only provided when there is a change in the state of the socket descriptor of interest. For compatibility with the semantics offered by `select` and `poll`, epoll also provides level-triggered event mechanisms.

To compare the performance of epoll with `select` and `poll`, we use the μserver [4, 7] web server. The μserver facilitates comparative analysis of different event dispatch mechanisms within the same code base through command-line parameters. Recently, a highly tuned version of the single process event driven μserver using `select` has shown promising results that rival the performance of the in-kernel TUX web server [4].

Interestingly, in this paper, we found that for some of the workloads considered `select` and `poll` perform as well as or slightly better than epoll. One such result is shown in Figure 1. This motivated further investigation with the goal of obtaining a better understanding of epoll's behaviour. In this paper, we describe our experience in trying to determine how to best use epoll, and examine techniques designed to improve its performance.

The rest of the paper is organized as follows: In Section 2 we summarize some existing work that led to the development of epoll as a scalable replacement for `select`. In Section 3 we describe the techniques we have tried to improve epoll's performance. In Section 4 we describe our experimental methodology, including the workloads used in the evaluation. In Section 5 we describe and analyze the results of our experiments. In Section 6 we summarize our findings and outline some ideas for future work.

2 Background and Related Work

Event-notification mechanisms have a long history in operating systems research and development, and have been a central issue in many performance studies. These studies have sought to improve mechanisms and interfaces for obtaining information about the state of socket and file descriptors from the operating system [2, 1, 3, 13, 15, 6, 18, 10, 12]. Some of these studies have developed improvements to `select`, `poll` and `sigwaitinfo` by reducing the amount of data copied between the application and kernel. Other studies have reduced the number of events delivered by the kernel, for example, the signal-per-fd scheme proposed by Chandra et al. [6]. Much of the aforementioned work is tracked and discussed on the web site, "The C10K Problem" [8].

Early work by Banga and Mogul [2] found that despite performing well under laboratory conditions, popular event-driven servers performed poorly under real-world conditions. They demonstrated that the discrepancy is due the inability of the `select` system call to scale to the large number of simultaneous connections that are found in WAN environments.

Subsequent work by Banga et al. [3] sought to improve on `select`'s performance by (among other things) separating the declaration of interest in events from the retrieval of events on that interest set. Event mechanisms like select and poll have traditionally combined these tasks into a single system call. However, this amalgamation requires the server to re-declare its interest set every time it wishes to retrieve events, since the kernel does not remember the interest sets from previous calls. This results in unnecessary data copying between the application and the kernel.

The `/dev/poll` mechanism was adapted from Sun Solaris to Linux by Provos et al. [15],

and improved on poll's performance by introducing a new interface that separated the declaration of interest in events from retrieval. Their `/dev/poll` mechanism further reduced data copying by using a shared memory region to return events to the application.

The kqueue event mechanism [9] addressed many of the deficiencies of `select` and `poll` for FreeBSD systems. In addition to separating the declaration of interest from retrieval, `kqueue` allows an application to retrieve events from a variety of sources including file/socket descriptors, signals, AIO completions, file system changes, and changes in process state.

The epoll event mechanism [18, 10, 12] investigated in this paper also separates the declaration of interest in events from their retrieval. The `epoll_create` system call instructs the kernel to create an event data structure that can be used to track events on a number of descriptors. Thereafter, the `epoll_ctl` call is used to modify interest sets, while the `epoll_wait` call is used to retrieve events.

Another drawback of `select` and `poll` is that they perform work that depends on the size of the interest set, rather than the number of events returned. This leads to poor performance when the interest set is much larger than the active set. The epoll mechanisms avoid this pitfall and provide performance that is largely independent of the size of the interest set.

3 Improving epoll Performance

Figure 1 in Section 5 shows the throughput obtained when using the μserver with the select, poll, and level-triggered epoll (epoll-LT) mechanisms. In this graph the x-axis shows increasing request rates and the y-axis shows the reply rate as measured by the clients that are inducing the load. This graph shows results for the one-byte workload. These results demonstrate that the μserver with level-triggered epoll does not perform as well as `select` under conditions that stress the event mechanisms. This led us to more closely examine these results. Using `gprof`, we observed that `epoll_ctl` was responsible for a large percentage of the run-time. As can been seen in Table 1 in Section 5 over 16% of the time is spent in `epoll_ctl`. The gprof output also indicates (not shown in the table) that `epoll_ctl` was being called a large number of times because it is called for every state change for each socket descriptor. We examine several approaches designed to reduce the number of `epoll_ctl` calls. These are outlined in the following paragraphs.

The first method uses epoll in an edge-triggered fashion, which requires the μserver to keep track of the current state of the socket descriptor. This is required because with the edge-triggered semantics, events are only received for transitions on the socket descriptor state. For example, once the server reads data from a socket, it needs to keep track of whether or not that socket is still readable, or if it will get another event from `epoll_wait` indicating that the socket is readable. Similar state information is maintained by the server regarding whether or not the socket can be written. This method is referred to in our graphs and the rest of the paper `epoll-ET`.

The second method, which we refer to as epoll2, simply calls `epoll_ctl` twice per socket descriptor. The first to register with the kernel that the server is interested in read and write events on the socket. The second call occurs when the socket is closed. It is used to tell epoll that we are no longer interested in events on that socket. All events are handled in a level-triggered fashion. Although this approach will reduce the number of `epoll_ctl` calls, it does have potential disadvantages.

One disadvantage of the epoll2 method is that because many of the sockets will continue to be readable or writable `epoll_wait` will return sooner, possibly with events that are currently not of interest to the server. For example, if the server is waiting for a read event on a socket it will not be interested in the fact that the socket is writable until later. Another disadvantage is that these calls return sooner, with fewer events being returned per call, resulting in a larger number of calls. Lastly, because many of the events will not be of interest to the server, the server is required to spend a bit of time to determine if it is or is not interested in each event and in discarding events that are not of interest.

The third method uses a new system call named `epoll_ctlv`. This system call is designed to reduce the overhead of multiple `epoll_ctl` system calls by aggregating several calls to `epoll_ctl` into one call to `epoll_ctlv`. This is achieved by passing an array of epoll events structures to `epoll_ctlv`, which then calls `epoll_ctl` for each element of the array. Events are generated in level-triggered fashion. This method is referred to in the figures and the remainder of the paper as epoll-ctlv.

We use `epoll_ctlv` to add socket descriptors to the interest set, and for modifying the interest sets for existing socket descriptors. However, removal of socket descriptors from the interest set is done by explicitly calling `epoll_ctl` just before the descriptor is closed. We do not aggregate deletion operations because by the time `epoll_ctlv` is invoked, the μserver has closed the descriptor and the `epoll_ctl` invoked on that descriptor will fail.

The μserver does not attempt to batch the closing of descriptors because it can run out of available file descriptors. Hence, the epoll-ctlv method uses both the `epoll_ctlv` and the `epoll_ctl` system calls. Alternatively, we could rely on the `close` system call to remove the socket descriptor from the interest set (and we did try this). However, this increases the time spent by the μserver in `close`, and does not alter performance. We verified this empirically and decided to explicitly call `epoll_ctl` to perform the deletion of descriptors from the epoll interest set.

4 Experimental Environment

The experimental environment consists of a single server and eight clients. The server contains dual 2.4 GHz Xeon processors, 1 GB of RAM, a 10,000 rpm SCSI disk, and two one Gigabit Ethernet cards. The clients are identical to the server with the exception of their disks which are EIDE. The server and clients are connected with a 24-port Gigabit switch. To avoid network bottlenecks, the first four clients communicate with the server's first Ethernet card, while the remaining four use a different IP address linked to the second Ethernet card. The server machine runs a slightly modified version of the 2.6.5 Linux kernel in uniprocessor mode.

4.1 Workloads

This section describes the workloads that we used to evaluate performance of the μserver with the different event notification mechanisms. In all experiments, we generate HTTP loads using *httperf* [11], an open-loop workload generator that uses connection timeouts to generate loads that can exceed the capacity of the server.

Our first workload is based on the widely used SPECweb99 benchmarking suite [17]. We use httperf in conjunction with a SPECweb99 file set and synthetic HTTP traces. Our traces have been carefully generated to recreate the

file classes, access patterns, and number of requests issued per (HTTP 1.1) connection that are used in the static portion of SPECweb99. The file set and server caches are sized so that the entire file set fits in the server's cache. This ensures that differences in cache hit rates do not affect performance.

Our second workload is called the one-byte workload. In this workload, the clients repeatedly request the same one byte file from the server's cache. We believe that this workload stresses the event dispatch mechanism by minimizing the amount of work that needs to be done by the server in completing a particular request. By reducing the effect of system calls such as `read` and `write`, this workload isolates the differences due to the event dispatch mechanisms.

To study the scalability of the event dispatch mechanisms as the number of socket descriptors (connections) is increased, we use *idleconn*, a program that comes as part of the httperf suite. This program maintains a steady number of idle connections to the server (in addition to the active connections maintained by httperf). If any of these connections are closed idleconn immediately re-establishes them. We first examine the behaviour of the event dispatch mechanisms without any idle connections to study scenarios where all of the connections present in a server are active. We then pre-load the server with a number of idle connections and then run experiments. The idle connections are used to increase the number of simultaneous connections in order to simulate a WAN environment. In this paper we present experiments using 10,000 idle connections, our findings with other numbers of idle connections were similar and they are not presented here.

4.2 Server Configuration

For all of our experiments, the μserver is run with the same set of configuration parameters except for the event dispatch mechanism. The μserver is configured to use `sendfile` to take advantage of zero-copy socket I/O while writing replies. We use TCP_CORK in conjunction with `sendfile`. The same server options are used for all experiments even though the use of TCP_CORK and `sendfile` may not provide benefits for the one-byte workload when compared with simply using `writev`.

4.3 Experimental Methodology

We measure the throughput of the μserver using different event dispatch mechanisms. In our graphs, each data point is the result of a two minute experiment. Trial and error revealed that two minutes is sufficient for the server to achieve a stable state of operation. A two minute delay is used between consecutive experiments, which allows the TIME_WAIT state on all sockets to be cleared before the subsequent run. All non-essential services are terminated prior to running any experiment.

5 Experimental Results

In this section we first compare the throughput achieved when using level-triggered epoll with that observed when using `select` and `poll` under both the one-byte and SPECweb99-like workloads with no idle connections. We then examine the effectiveness of the different methods described for reducing the number of `epoll_ctl` calls under these same workloads. This is followed by a comparison of the performance of the event dispatch mechanisms when the server is pre-loaded with 10,000 idle connections. Finally, we describe the results of experiments in which we tune the

accept strategy used in conjunction with epoll-LT and epoll-ctlv to further improve their performance.

We initially ran the one byte and the SPECweb99-like workloads to compare the performance of the select, poll and level-triggered epoll mechanisms.

As shown in Figure 1 and Figure 2, for both of these workloads select and poll perform as well as epoll-LT. It is important to note that because there are no idle connections for these experiments the number of socket descriptors tracked by each mechanism is not very high. As expected, the gap between epoll-LT and select is more pronounced for the one byte workload because it places more stress on the event dispatch mechanism.

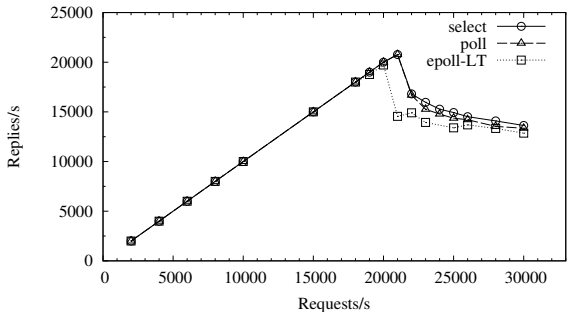


Figure 1: *µserver performance on one byte workload using select, poll, and epoll-LT*

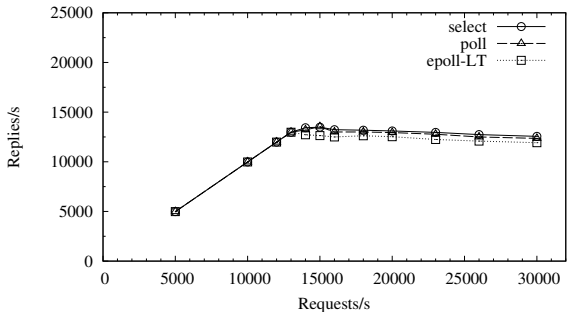


Figure 2: *µserver performance on SPECweb99-like workload using select, poll, and epoll-LT*

We tried to improve the performance of the server by exploring different techniques for using epoll as described in Section 3. The effect of these techniques on the one-byte workload is shown in Figure 3. The graphs in this figure show that for this workload the techniques used to reduce the number of `epoll_ctl` calls do not provide significant benefits when compared with their level-triggered counterpart (epoll-LT). Additionally, the performance of select and poll is equal to or slightly better than each of the epoll techniques. Note that we omit the line for poll from Figures 3 and 4 because it is nearly identical to the select line.

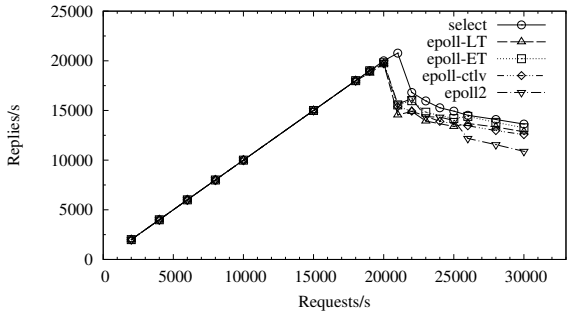


Figure 3: *µserver performance on one byte workload with no idle connections*

We further analyze the results from Figure 3 by profiling the µserver using gprof at the request rate of 22,000 requests per second. Table 1 shows the percentage of time spent in system calls (rows) under the various event dispatch methods (columns). The output for system calls and µserver functions which do not contribute significantly to the total run-time is left out of the table for clarity.

If we compare the select and poll columns we see that they have a similar breakdown including spending about 13% of their time indicating to the kernel events of interest and obtaining events. In contrast the epoll-LT, epoll-ctlv, and epoll2 approaches spend about 21 – 23% of their time on their equivalent functions (`epoll_ctl`, `epoll_ctlv` and `epoll_wait`). Despite these extra overheads the throughputs obtained using the epoll techniques compare favourably with those obtained

	select	epoll-LT	epoll-ctlv	epoll2	epoll-ET	poll
read	21.51	20.95	21.41	20.08	22.19	20.97
close	14.90	14.05	14.90	13.02	14.14	14.79
select	13.33	-	-	-	-	-
poll	-	-	-	-	-	13.32
epoll_ctl	-	16.34	5.98	10.27	11.06	-
epoll_wait	-	7.15	6.01	12.56	6.52	-
epoll_ctlv	-	-	9.28	-	-	-
setsockopt	11.17	9.13	9.13	7.57	9.08	10.68
accept	10.08	9.51	9.76	9.05	9.30	10.20
write	5.98	5.06	5.10	4.13	5.31	5.70
fcntl	3.66	3.34	3.37	3.14	3.34	3.61
sendfile	3.43	2.70	2.71	3.00	3.91	3.43

Table 1: gprof profile data for the μserver under the one-byte workload at 22,000 requests/sec

using `select` and `poll`. We note that when using `select` and `poll` the application requires extra manipulation, copying, and event scanning code that is not required in the epoll case (and does not appear in the gprof data).

The results in Table 1 also show that the overhead due to `epoll_ctl` calls is reduced in epoll-ctlv, epoll2 and epoll-ET, when compared with epoll-LT. However, in each case these improvements are offset by increased costs in other portions of the code. The epoll2 technique spends twice as much time in `epoll_wait` when compared with epoll-LT. With epoll2 the number of calls to `epoll_wait` is significantly higher, the average number of descriptors returned is lower, and only a very small proportion of the calls (less than 1%) return events that need to be acted upon by the server. On the other hand, when compared with epoll-LT the epoll2 technique spends about 6% less time on `epoll_ctl` calls so the total amount of time spent dealing with events is comparable with that of epoll-LT. Despite the significant `epoll_wait` overheads epoll2 performance compares favourably with the other methods on this workload.

Using the epoll-ctlv technique, gprof indicates that `epoll_ctlv` and `epoll_ctl` combine for a total of 1,949,404 calls compared with 3,947,769 `epoll_ctl` calls when using epoll-LT. While epoll-ctlv helps to reduce the number of user-kernel boundary crossings, the net result is no better than epoll-LT. The amount of time taken by epoll-ctlv in `epoll_ctlv` and `epoll_ctl` system calls is about the same (around 16%) as that spent by level-triggered epoll in invoking `epoll_ctl`.

When comparing the percentage of time epoll-LT and epoll-ET spend in `epoll_ctl` we see that it has been reduced using epoll-ET from 16% to 11%. Although the `epoll_ctl` time has been reduced it does not result in an appreciable improvement in throughput. We also note that about 2% of the run-time (which is not shown in the table) is also spent in the epoll-ET case checking, and tracking the state of the request (i.e., whether the server should be reading or writing) and the state of the socket (i.e., whether it is readable or writable). We expect that this can be reduced but that it wouldn't noticeably impact performance.

Results for the SPECweb99-like workload are

shown in Figure 4. Here the graph shows that all techniques produce very similar results with a very slight performance advantage going to epoll-ET after the saturation point is reached.

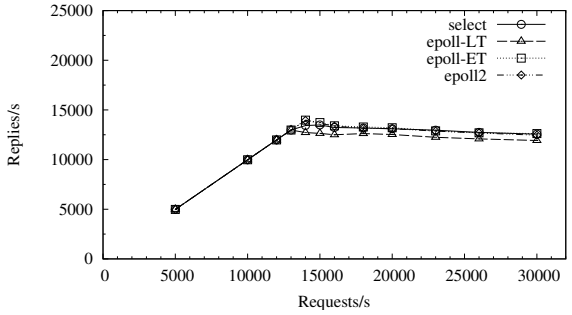


Figure 4: *µserver performance on SPECweb99-like workload with no idle connections*

5.1 Results With Idle Connections

We now compare the performance of the event mechanisms with 10,000 idle connections. The idle connections are intended to simulate the presence of larger numbers of simultaneous connections (as might occur in a WAN environment). Thus, the event dispatch mechanism has to keep track of a large number of descriptors even though only a very small portion of them are active.

By comparing results in Figures 3 and 5 one can see that the performance of select and poll degrade by up to 79% when the 10,000 idle connections are added. The performance of epoll2 with idle connections suffers similarly to select and poll. In this case, epoll2 suffers from the overheads incurred by making a large number of `epoll_wait` calls the vast majority of which return events that are not of current interest to the server. Throughput with level-triggered epoll is slightly reduced with the addition of the idle connections while edge-triggered epoll is not impacted.

The results for the SPECweb99-like workload with 10,000 idle connections are shown in Figure 6. In this case each of the event mechanisms is impacted in a manner similar to that in which they are impacted by idle connections in the one-byte workload case.

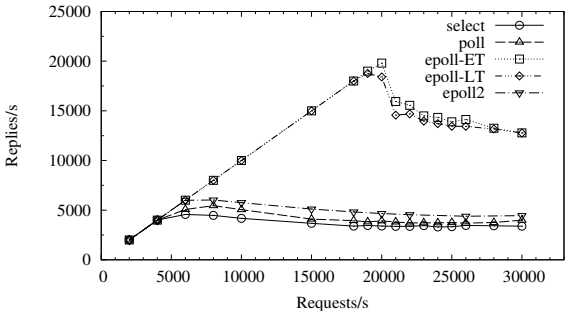


Figure 5: *µserver performance on one byte workload and 10,000 idle connections*

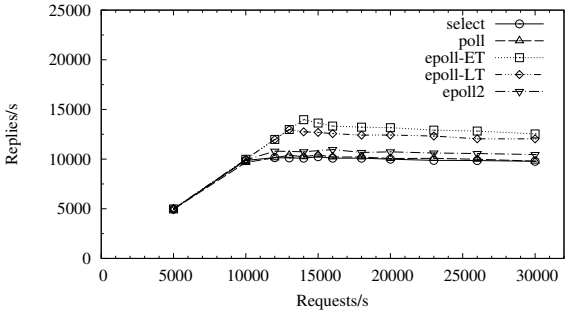


Figure 6: *µserver performance on SPECweb99-like workload and 10,000 idle connections*

5.2 Tuning Accept Strategy for epoll

The µserver's accept strategy has been tuned for use with `select`. The µserver includes a parameter that controls the number of connections that are accepted consecutively. We call this parameter the accept-limit. Parameter values range from one to infinity (Inf). A value of one limits the server to accepting at most one connection when notified of a pending connection request, while Inf causes the server to consecutively accept all currently pending connections.

To this point we have used the accept strategy that was shown to be effective for `select` by

Brecht et al. [4] (i.e., accept-limit is Inf). In order to verify whether the same strategy performs well with the epoll-based methods we explored their performance under different accept strategies.

Figure 7 examines the performance of level-triggered epoll after the accept-limit has been tuned for the one-byte workload (other values were explored but only the best values are shown). Level-triggered epoll with an accept limit of 10 shows a marked improvement over the previous accept-limit of Inf, and now matches the performance of select on this workload. The accept-limit of 10 also improves peak throughput for the epoll-ctlv model by 7%. This gap widens to 32% at 21,000 requests/sec. In fact the best accept strategy for epoll-ctlv fares slightly better than the best accept strategy for select.

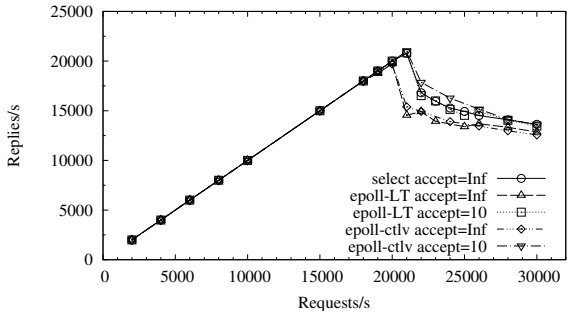


Figure 7: *µserver performance on one byte workload with different accept strategies and no idle connections*

Varying the accept-limit did not improve the performance of the edge-triggered epoll technique under this workload and it is not shown in the graph. However, we believe that the effects of the accept strategy on the various epoll techniques warrants further study as the efficacy of the strategy may be workload dependent.

6 Discussion

In this paper we use a high-performance event-driven HTTP server, the *µ*server, to compare and evaluate the performance of select, poll, and epoll event mechanisms. Interestingly, we observe that under some of the workloads examined the throughput obtained using `select` and `poll` is as good or slightly better than that obtained with epoll. While these workloads may not utilize representative numbers of simultaneous connections they do stress the event mechanisms being tested.

Our results also show that a main source of overhead when using level-triggered epoll is the large number of `epoll_ctl` calls. We explore techniques which significantly reduce the number of `epoll_ctl` calls, including the use of edge-triggered events and a system call, `epoll_ctlv`, which allows the *µ*server to aggregate large numbers of `epoll_ctl` calls into a single system call. While these techniques are successful in reducing the number of `epoll_ctl` calls they do not appear to provide appreciable improvements in performance.

As expected, the introduction of idle connections results in dramatic performance degradation when using `select` and `poll`, while not noticeably impacting the performance when using epoll. Although it is not clear that the use of idle connections to simulate larger numbers of connections is representative of real workloads, we find that the addition of idle connections does not significantly alter the performance of the edge-triggered and level-triggered epoll mechanisms. The edge-triggered epoll mechanism performs best with the level-triggered epoll mechanism offering performance that is very close to edge-triggered.

In the future we plan to re-evaluate some of

the mechanisms explored in this paper under more representative workloads that include more representative wide area network conditions. The problem with the technique of using idle connections is that the idle connections simply inflate the number of connections without doing any useful work. We plan to explore tools similar to Dummynet [16] and NIST Net [5] in order to more accurately simulate traffic delays, packet loss, and other wide area network traffic characteristics, and to re-examine the performance of Internet servers using different event dispatch mechanisms and a wider variety of workloads.

7 Acknowledgments

We gratefully acknowledge Hewlett Packard, the Ontario Research and Development Challenge Fund, and the National Sciences and Engineering Research Council of Canada for financial support for this project.

References

[1] G. Banga, P. Druschel, and J.C. Mogul. Resource containers: A new facility for resource management in server systems. In *Operating Systems Design and Implementation*, pages 45–58, 1999.

[2] G. Banga and J.C. Mogul. Scalable kernel performance for Internet servers under realistic loads. In *Proceedings of the 1998 USENIX Annual Technical Conference*, New Orleans, LA, 1998.

[3] G. Banga, J.C. Mogul, and P. Druschel. A scalable and explicit event delivery mechanism for UNIX. In *Proceedings of the 1999 USENIX Annual Technical Conference*, Monterey, CA, June 1999.

[4] Tim Brecht, David Pariag, and Louay Gammo. accept()able strategies for improving web server performance. In *Proceedings of the 2004 USENIX Annual Technical Conference (to appear)*, June 2004.

[5] M. Carson and D. Santay. NIST Net – a Linux-based network emulation tool. *Computer Communication Review*, to appear.

[6] A. Chandra and D. Mosberger. Scalability of Linux event-dispatch mechanisms. In *Proceedings of the 2001 USENIX Annual Technical Conference*, Boston, 2001.

[7] HP Labs. The userver home page, 2004. Available at `http://hpl.hp.com/research/linux/userver`.

[8] Dan Kegel. The C10K problem, 2004. Available at `http://www.kegel.com/c10k.html`.

[9] Jonathon Lemon. Kqueue—a generic and scalable event notification facility. In *Proceedings of the USENIX Annual Technical Conference, FREENIX Track*, 2001.

[10] Davide Libenzi. Improving (network) I/O performance. Available at `http://www.xmailserver.org/linux-patches/nio-improve.html`.

[11] D. Mosberger and T. Jin. httperf: A tool for measuring web server performance. In *The First Workshop on Internet Server Performance*, pages 59–67, Madison, WI, June 1998.

[12] Shailabh Nagar, Paul Larson, Hanna Linder, and David Stevens. epoll scalability web page. Available at `http://lse.sourceforge.net/epoll/index.html`.

[13] M. Ostrowski. A mechanism for scalable event notification and delivery in Linux. Master's thesis, Department of Computer Science, University of Waterloo, November 2000.

[14] Vivek S. Pai, Peter Druschel, and Willy Zwaenepoel. Flash: An efficient and portable Web server. In *Proceedings of the USENIX 1999 Annual Technical Conference*, Monterey, CA, June 1999. `http://citeseer.nj.nec.com/article/pai99flash.html`.

[15] N. Provos and C. Lever. Scalable network I/O in Linux. In *Proceedings of the USENIX Annual Technical Conference, FREENIX Track*, June 2000.

[16] Luigi Rizzo. Dummynet: a simple approach to the evaluation of network protocols. *ACM Computer Communication Review*, 27(1):31–41, 1997. `http://citeseer.ist.psu.edu/rizzo97dummynet.html`.

[17] Standard Performance Evaluation Corporation. *SPECWeb99 Benchmark*, 1999. Available at `http://www.specbench.org/osg/web99`.

[18] David Weekly. /dev/epoll – a highspeed Linux kernel patch. Available at `http://epoll.hackerdojo.com`.

The (Re)Architecture of the X Window System

James Gettys
jim.gettys@hp.com
Keith Packard
keithp@keithp.com
HP Cambridge Research Laboratory

Abstract

The X Window System, Version 11, is the standard window system on Linux and UNIX systems. X11, designed in 1987, was "state of the art" at that time. From its inception, X has been a network transparent window system in which X client applications can run on any machine in a network using an X server running on any display. While there have been some significant extensions to X over its history (e.g. OpenGL support), X's design lay fallow over much of the 1990's. With the increasing interest in open source systems, it was no longer sufficient for modern applications and a significant overhaul is now well underway. This paper describes revisions to the architecture of the window system used in a growing fraction of desktops and embedded systems

1 Introduction

While part of this work on the X window system [SG92] is "good citizenship" required by open source, some of the architectural problems solved ease the ability of open source applications to print their results, and some of the techniques developed are believed to be in advance of the commercial computer industry. The challenges being faced include:

- X's fundamentally flawed font architecture made it difficult to implement good WYSIWYG systems
- Inadequate 2D graphics, which had always been intended to be augmented and/or replaced
- Developers are loathe to adopt any new technology that limits the distribution of their applications
- Legal requirements for accessibility for screen magnifiers are difficult to implement
- Users desire modern user interface eye candy, which sport translucent graphics and windows, drop shadows, etc.
- Full integration of applications into 3 D environments
- Collaborative shared use of X (e.g. multiple simultaneous use of projector walls or other shared applications)

While some of this work has been published elsewhere, there has never been any overview paper describing this work as an integrated whole, and the compositing manager work described below is novel as of fall 2003. This work represents a long term effort that started in 1999, and will continue for several years more.

2 Text and Graphics

X's obsolete 2D bit-blit based text and graphics system problems were most urgent. The development of the Gnome and KDE GUI environments in the period 1997-2000 had shown X11's fundamental soundness, but confirmed the authors' belief that the rendering system in X was woefully inadequate. One of us participated in the original X11 design meetings where the intent was to augment the rendering design at a later date; but the "GUI Wars" of the late 1980's doomed effort in this area. Good printing support has been particularly difficult to implement in X applications, as fonts have were opaque X server side objects not directly accessible by applications.

Most applications now composite images in sophisticated ways, whether it be in Flash media players, or subtly as part of anti-aliased characters. Bit-Blit is not sufficient for these applications, and these modern applications were (if only by their use of modern toolkits) all resorting to pixel based image manipulation. The screen pixels are retrieved from the window system, composited in clients, and then restored to the screen, rather than directly composited in hardware, resulting in poor performance. Inspired by the model first implemented in the Plan 9 window system, a graphics model based on Porter/Duff [PD84] image compositing was chosen. This work resulted in the X Render extension [Pac01a].

X11's core graphics exposed fonts as a server side abstraction. This font model was, at best, marginally adequate by 1987 standards. Even WYSIWYG systems of that era found them insufficient. Much additional information embedded in fonts (e.g. kerning tables) were not available from X whatsoever. Current competitive systems implement anti-aliased outline fonts. Discovering the Unicode coverage of a font, required by current toolkits for internationalization, was causing major performance problems. Deploying new server side font technology is slow, as X is a distributed system, and many X servers are seldom (or never) updated.

Therefore, a more fundamental change in X's architecture was undertaken: to no longer use server side fonts at all, but to allow applications direct access to font files and have the window system cache and composite glyphs onto the screen.

The first implementation of the new font system [Pac01b] taught a vital lesson. Xft1 provided anti-aliased text and proper font naming/substitution support, but reverted to the core X11 bitmap fonts if the Render extension was not present. Xft1 included the first implementation what is called "subpixel decimation," which provides higher quality subpixel based rendering than Microsoft's ClearType [Pla00] technology in a completely general algorithm.

Despite these advances, Xft1 received at best a lukewarm reception. If an application developer wanted anti-aliased text universally, Xft1 did not help them, since it relied on the Render extension which had not yet been widely deployed; instead, the developer would be faced with two implementations, and higher maintenance costs. This (in retrospect obvious) rational behavior of application developers shows the high importance of backwards compatibility; X extensions intended for application developers' use must be designed in a downward compatible form whenever possible, and should enable a complete conversion to a new facility, so that multiple code paths in applications do not need testing and maintenance. These principles have guided later development.

The font installation, naming, substitution, and internationalization problems were sepa-

rated from Xft into a library named Fontconfig [Pac02], (since some printer only applications need this functionality independent of the window system.) Fontconfig provides internationalization features in advance of those in commercial systems such as Windows or OS X, and enables trivial font installation with good performance even when using thousands of fonts. Xft2 was also modified to operate against legacy X servers lacking the Render extension.

Xft2 and Fontconfig's solving of several major problems and lack of deployment barriers enabled rapid acceptance and deployment in the open source community, seeing almost universal use and uptake in less than one calendar year. They have been widely deployed on Linux systems since the end of 2002. They also "future proof" open source systems against coming improvements in font systems (e.g. OpenType), as the window system is no longer a gating item for font technology.

Sun Microsystems implemented a server side font extension over the last several years; for the reasons outlined in this section, it has not been adopted by open source developers.

While Xft2 and Fontconfig finally freed application developers from the tyranny of X11's core font system, improved performance [PG03], and at a stroke simplified their printing problems, it has still left a substantial burden on applications. The X11 core graphics, even augmented by the Render extension, lack convenient facilities for many applications for even simple primitives like splines, tasteful wide lines, stroking paths, etc, much less provide simple ways for applications to print the results on paper.

3 Cairo

The Cairo library [WP03], developed by one of the authors in conjunction with by Carl Worth of ISI, is designed to solve this problem. Cairo provides a state full user-level API with support for the PDF 1.4 imaging model. Cairo provides operations including stroking and filling Bézier cubic splines, transforming and compositing translucent images, and anti-aliased text rendering. The PostScript drawing model has been adapted for use within applications. Extensions needed to support much of the PDF 1.4 imaging operations have been included. This integration of the familiar PostScript operational model within the native application language environments provides a simple and powerful new tool for graphics application development.

Cairo's rendering algorithms use work done in the 1980's by Guibas, Ramshaw, and Stolfi [GRS83] along with work by John Hobby [Hob85], which has never been exploited in Postscript or in Windows. The implementation is fast, precise, and numerically stable, supports hardware acceleration, and is in advance of commercial systems.

Of particular note is the current development of Glitz [NR04], an OpenGL backend for Cairo, being developed by a pair of master's students in Sweden. Not only is it showing that a high speed implementation of Cairo is possible, it implements an interface very similar to the X Render extension's interface. More about this in the OpenGL section below.

Cairo is in the late stages of development and is being widely adopted in the open source community. It includes the ability to render to Postscript and a PDF back end is planned, which should greatly improve applications' printing support. Work to incorporate Cairo in the Gnome and KDE desktop environments is

well underway, as are ports to Windows and Apple's MacIntosh, and it is being used by the Mono project. As with Xft2, Cairo works with all X servers, even those without the Render extension.

4 Accessibility and Eye-Candy

Several years ago, one of us implemented a prototype X system that used image compositing as the fundamental primitive for constructing the screen representation of the window hierarchy contents. Child window contents were composited to their parent windows which were incrementally composed to their parents until the final screen image was formed, enabling translucent windows. The problem with this simplistic model was twofold—first, a naïve implementation consumed enormous resources as each window required two complete off screen buffers (one for the window contents themselves, and one for the window contents composited with the children) and took huge amounts of time to build the final screen image as it recursively composited windows together. Secondly, the policy governing the compositing was hardwired into the X server. An architecture for exposing the same semantics with less overhead seemed almost possible, and pieces of it were implemented (miext/layer). However, no complete system was fielded, and every copy of the code tracked down and destroyed to prevent its escape into the wild.

Both Mac OS X and DirectFB [Hun04] perform window-level compositing by creating off-screen buffers for each top-level window (in OS X, the window system is not nested, so there are only top-level windows). The screen image is then formed by taking the resulting images and blending them together on the screen. Without handling the nested window case, both of these systems provide the desired functionality with a simple implementation. This simple approach is inadequate for X as some desktop environments nest the whole system inside a single top-level window to allow panning, and X's long history has shown the value of separating mechanism from policy (Gnome and KDE were developed over 10 years after X11's design). The fix is pretty easy—allow applications to select which pieces of the window hierarchy are to be stored off-screen and which are to be drawn to their parent storage.

With window hierarchy contents stored in off-screen buffers, an external application can now control how the screen contents are constructed from the constituent sub-windows and whatever other graphical elements are desired. This eliminated the complexities surrounding precisely what semantics would be offered in window-level compositing within the X server and the design of the underlying X extensions. They were replaced by some concerns over the performance implications of using an external agent (the "Compositing Manager") to execute the requests needed to present the screen image. Note that every visible pixel is under the control of the compositing manager, so screen updates are limited to how fast that application can get the bits painted to the screen.

The architecture is split across three new extensions:

- Composite, which controls which sub-hierarchies within the window tree are rendered to separate buffers.
- Damage, which tracks modified areas with windows, informing the Composting Manager which areas of the off-screen hierarchy components have changed.
- Xfixes, which includes new Region objects permitting all of the above computation to be performed indirectly within the

X server, avoiding round trips.

Multiple applications can take advantage of the off screen window contents, allowing thumbnail or screen magnifier applications to be included in the desktop environment.

To allow applications other than the compositing manager to present alpha-blended content to the screen, a new X Visual was added to the server. At 32 bits deep, it provides 8 bits of red, green and blue along with 8 bits of alpha value. Applications can create windows using this visual and the compositing manager can composite them onto the screen.

Nothing in this fundamental design indicates that it is used for constructing translucent windows; redirection of window contents and notification of window content change seems pretty far removed from one of the final goals. But note the compositing manger can use whatever X requests it likes to paint the combined image, including requests from the Render extension, which does know how to blend translucent images together. The final image is constructed programmatically so the possible presentation on the screen is limited only by the fertile imagination of the numerous eye-candy developers, and not restricted to any policy imposed by the base window system. And vital to rapid deployment, most applications can be completely oblivious to this background legerdemain.

In this design, such sophisticated effects need only be applied at frame update rates on only modified sections of the screen rather than at the rate applications perform graphics; this constant behavior is highly desirable in systems.

There is very strong "pull" from both commercial and non-commercial users of X for this work and the current early version will likely be shipped as part of the next X.org Foundation X Window System release, sometime this summer. Since there has not been sufficient exposure through widespread use, further changes will certainly be required further experience with the facilities are gained in a much larger audience; as these can be made without affecting existing applications, immediate deployment is both possible and extremely desirable.

The mechanisms described above realize a fundamentally more interesting architecture than either Windows or Mac OSX, where the compositing policy is hardwired into the window system. We expect a fertile explosion of experimentation, experience (both good and bad), and a winnowing of ideas as these facilities gain wider exposure.

5 Input Transformation

In the "naïve," eye-candy use of the new compositing functions, no transformation of input events are required, as input to windows remains at the same geometric position on the screen, even though the windows are first rendered off screen. More sophisticated use, for example, screen readers or immersive environments such as Croquet [SRRK02], or Sun's Looking Glass [KJ04] requires transformation of input events from where they first occur on the visible screen to the actual position in the windows (being rendered from off screen), since the window's contents may have been arbitrarily transformed or even texture mapped onto shapes on the screen.

As part of Sun Microsystem's award winning work on accessibility in open source for screen readers, Sun has developed the XEvIE extension [Kre], which allows external clients to transform input events. This looks like a good starting point for the somewhat more general problem that 3D systems pose, and with some

modification can serve both the accessibility needs and those of more sophisticated applications.

6 Synchronization

Synchronization is probably the largest remaining challenge posed by compositing. While composite has eliminated much flashing of the screen since window exposure is eliminated, this does not solve the challenge of the compositing manager happening to copy an application's window to the frame buffer in the middle of an application painting a sequence of updates. No "tearing" of single graphics operations take place since the X server is single threaded, and all graphics operations are run to completion.

The X Synchronization extension (XSync) [GCGW92], widely available but to date seldom used, provides a general set of mechanisms for applications to synchronize with each other, with real time, and potentially with other system provided counters. XSync's original design intent intended system provided counters for vertical retrace interrupts, audio sample clocks, and similar system facilities, enabling very tight synchronization of graphics operations with these time bases. Work has begun on Linux to provide these counters at long last, when available, to flesh out the design originally put in place and tested in the early 1990's.

A possible design for solving the application synchronization problem at low overhead may be to mark sections of requests with increments of XSync counters: if the count is odd (or even) the window would be unstable/stable. The compositing manager might then copy the window only if the window is in a stable state. Some details and possibly extensions to XSync will need to be worked out, if this approach is pursued.

7 Next Steps

We believe we are slightly more than half way through the process of rearchitecting and reimplementing the X Window System. The existing prototype needs to become a production system requiring significant infrastructure work as described in this section.

7.1 OpenGL based X

Current X-based systems which support OpenGL do so by encapsulating the OpenGL environment within X windows. As such, an OpenGL application cannot manipulate X objects with OpenGL drawing commands.

Using OpenGL as the basis for the X server itself will place X objects such as pixmaps and off-screen window contents inside OpenGL objects allowing applications to use the full OpenGL command set to manipulate them.

A "proof of concept" of implementation of the X Render extension is being done as part of the Glitz back-end for Cairo, which is showing very good performance for render based applications. Whether the "core" X graphics will require any OpenGL extensions is still somewhat an open question.

In concert with the new compositing extensions, conventional X applications can then be integrated into 3D environments such as Croquet, or Sun's Looking Glass. X application contents can be used as textures and mapped onto any surface desired in those environments.

This work is underway, but not demonstrable at this date.

7.2 Kernel support for graphics cards

In current open source systems, graphics cards are supported in a manner totally unlike that of any other operating system, and unlike previous device drivers for the X Window System on commercial UNIX systems. There is no single central kernel driver responsible for managing access to the hardware. Instead, a large set of cooperating user and kernel mode systems are involved in mutual support of the hardware, including the X server (for 2D graphic), the direct-rendering infrastructure (DRI) (for accelerated 3D graphics), the kernel frame buffer driver (for text console emulation), the General ATI TV and Overlay Software (GATOS) (for video input and output) and alternate 2D graphics systems like DirectFB.

Two of these systems, the kernel frame buffer driver and the X server both include code to configure the graphics card "video mode"—the settings needed to send the correct video signals to monitors connected to the card. Three of these systems, DRI, the X server and GATOS, all include code for managing the memory space within the graphics card. All of these systems directly manipulate hardware registers without any coordination among them.

The X server has no kernel component for 2D graphics. Long-latency operations cannot use interrupts, instead the X server spins while polling status registers. DMA is difficult or impossible to configure in this environment. Perhaps the most egregious problem is that the X server reconfigures the PCI bus to correct BIOS mapping errors without informing the operating system kernel. Kernel access to devices while this remapping is going on may find the related devices mismapped.

To rationalize this situation, various groups and vendors are coordinating efforts to create a single kernel-level entity responsible for basic device management, but this effort has just begun.

7.3 Housecleaning and Latency Elimination and Latency Hiding

Serious attempts were made in the early 1990's to multi-thread the X server itself, with the discovery that the threading overhead in the X server is a net performance loss [Smi92].

Applications, however, often need to be multi-threaded. The primary C binding to the X protocol is called Xlib, and its current implementation by one of us dates from 1987. While it was partially developed on a Firefly multiprocessor workstation of that era, something almost unheard of at that date, and some consideration of multi-threaded applications were taken in its implementation, its internal transport facilities were never expected/intended to be preserved when serious multi-threaded operating systems became available. Unfortunately, rather than a full rewrite as one of us expected, multi-threaded support was debugged into existence using the original code base and the resulting code is very bug-prone and hard to maintain. Additionally, over the years, Xlib became a "kitchen sink" library, including functionality well beyond its primary use as a binding to the X protocol. We have both seriously regretted the precedents both of us set introducing extraneous functionality into Xlib, causing it to be one of the largest libraries on UNIX/Linux systems. Due to better facilities in modern toolkits and system libraries, more than half of Xlib's current footprint is obsolete code or data.

While serious work was done in X11's design to mitigate latency, X's performance, particularly over low speed networks, is often limited by round trip latency, and with retrospect much more can be done [PG03]. As this

work shows, client side fonts have made a significant improvement in startup latency, and work has already been completed in toolkits to mitigate some of the other hot spots. Much of the latency can be retrieved by some simple techniques already underway, but some require more sophisticated techniques that the current Xlib implementation is not capable of. Potentially 90the latency as of 2003 can be recovered by various techniques. The XCB library [MS01] by Bart Massey and Jamey Sharp is both carefully engineered to be multithreaded and to expose interfaces that will allow for latency hiding.

Since libraries linked against different basic X transport systems would cause havoc in the same address space, a Xlib compatibility layer (XCL) has been developed that provides the "traditional" X library API, using the original Xlib stubs, but replacing the internal transport and locking system, which will allow for much more useful latency hiding interfaces. The XCB/XCL version of Xlib is now able to run essentially all applications, and after a shakedown period, should be able to replace the existing Xlib transport soon. Other bindings than the traditional Xlib bindings then become possible in the same address space, and we may see toolkits adopt those bindings at substantial savings in space.

7.4 Mobility, Collaboration, and Other Topics

X's original intended environment included highly mobile students, and a hope, never generally realized for X, was the migration of applications between X servers.

The user should be able to travel between systems running X and retrieve your running applications (with suitable authentication and authorization). The user should be able to log out and "park" applications somewhere for later retrieval, either on the same display, or elsewhere. Users should be able to replicate an application's display on a wall projector for presentation. Applications should be able to easily survive the loss of the X server (most commonly caused by the loss of the underlying TCP connection, when running remotely).

Toolkit implementers typically did not understand and share this poorly enunciated vision and were primarily driven by pressing immediate needs, and X's design and implementation made migration or replication difficult to implement as an afterthought. As a result, migration (and replication) was seldom implemented, and early toolkits such as Xt made it even more difficult. Emacs is the only widespread application capable of both migration and replication, and it avoided using any toolkit. A more detailed description of this vision is available in [Get02].

Recent work in some of the modern toolkits (e.g. GTK+) and evolution of X itself make much of this vision demonstrable in current applications. Some work in the X infrastructure (Xlib) is underway to enable the prototype in GTK+ to be finished.

Similarly, input devices need to become full-fledged network data sources, to enable much looser coupling of keyboards, mice, game consoles and projectors and displays; the challenge here will be the authentication, authorization and security issues this will raise. The HAL and DBUS projects hosted at freedesktop.org are working on at least part of the solutions for the user interface challenges posed by hotplug of input devices.

7.5 Color Management

The existing color management facilities in X are over 10 years old, have never seen widespread use, and do not meet current needs. This area is ripe for revisiting. Marti Maria Sa-

guer's LittleCMS [Mar] may be of use here. For the first time, we have the opportunity to "get it right" from one end to the other if we choose to make the investment.

7.6 Security and Authentication

Transport security has become an burning issue; X is network transparent (applications can run on any system in a network, using remote displays), yet we dare no longer use X over the network directly due to password grabbing kits in the hands of script kiddies. SSH [BS01] provides such facilities via port forwarding and is being used as a temporary stopgap. Urgent work on something better is vital to enable scaling and avoid the performance and latency issues introduced by transit of extra processes, particularly on (Linux Terminal Server Project (LTSP [McQ02]) servers, which are beginning break out of their initial use in schools and other non security sensitive environments into very sensitive commercial environments.

Another aspect of security arises between applications sharing a display. In the early and mid 1990's efforts were made as a result of the compartmented mode workstation projects to make it much more difficult for applications to share or steal data from each other on a X display. These facilities are very inflexible, and have gone almost unused.

As projectors and other shared displays become common over the next five years, applications from multiple users sharing a display will become commonplace. In such environments, different people may be using the same display at the same time and would like some level of assurance that their application's data is not being grabbed by the other user's application.

Eamon Walsh has, as part of the SELinux project [Wal04], been working to replace the existing X Security extension with an extension that, as in SELinux, will allow multiple different security policies to be developed external to the X server. This should allow multiple different policies to be available to suit the varied uses: normal workstations, secure workstations, shared displays in conference rooms, etc.

7.7 Compression and Image Transport

Many/most modern applications and desktops, including the most commonly used application (a web browser) are now intensive users of synthetic and natural images. The previous attempt (XIE [SSF+96]) to provide compressed image transport failed due to excessive complexity and over ambition of the designers, has never been significantly used, and is now in fact not even shipped as part of current X distributions.

Today, many images are being read from disk or the network in compressed form, uncompressed into memory in the X client, moved to the X server (where they often occupy another copy of the uncompressed data). If we add general data compression to X (or run X over ssh with compression enabled) the data would be both compressed and uncompressed on its way to the X server. A simple replacement for XIE (if the complexity slippery slope can be avoided in a second attempt) would be worthwhile, along with other general compression of the X protocol.

Results in our 2003 Usenix X Network Performance paper show that, in real application workloads (the startup of a Gnome desktop), using even simple GZIP [Gai93] style compression can make a tremendous difference in a network environment, with a factor of 300(!) savings in bandwidth. Apparently the synthetic images used in many current UI's are extremely good candidates for

compression. A simple X extension that could encapsulate one or more X requests into the extension request would avoid multiple compression/uncompression of the same data in the system where an image transport extension was also present. The basic X protocol framework is actually very byte efficient relative to most conventional RPC systems, with a basic X request only occupying 4 bytes (contrast this with HTTP or CORBA, in which a simple request is more than 100 bytes).

With the great recent interest in LTSP in commercial environments, work here would be extremely well spent, saving both memory and CPU, and network bandwidth.

We are more than happy to hear from anyone interested in helping in this effort to bring X into the new millennium.

References

[BS01] Daniel J. Barrett and Richard Silverman. *SSH, The Secure Shell: The Definitive Guide*. O'Reilly & Associates, Inc., 2001.

[Gai93] Jean-Loup Gailly. *Gzip: The Data Compression Program*. iUniverse.com, 1.2.4 edition, 1993.

[GCGW92] Tim Glauert, Dave Carver, James Gettys, and David Wiggins. X Synchronization Extension Protocol, Version 3.0. X consortium standard, 1992.

[Get02] James Gettys. The Future is Coming, Where the X Window System Should Go. In *FREENIX Track, 2002 Usenix Annual Technical Conference*, Monterey, CA, June 2002. USENIX.

[GRS83] Leo Guibas, Lyle Ramshaw, and Jorge Stolfi. A kinetic framework for computational geometry. In *Proceedings of the IEEE 1983 24th Annual Symposium on the Foundations of Computer Science*, pages 100–111. IEEE Computer Society Press, 1983.

[Hob85] John D. Hobby. *Digitized Brush Trajectories*. PhD thesis, Stanford University, 1985. Also *Stanford Report STAN-CS-85-1070*.

[Hun04] A. Hundt. DirectFB Overview (v0.2 for DirectFB 0.9.21), February 2004. `http://www.directfb.org/documentation`.

[KJ04] H. Kawahara and D. Johnson. Project Looking Glass: 3D Desktop Exploration. In *X Developers Conference*, Cambridge, MA, April 2004.

[Kre] S. Kreitman. XEvIE - X Event Interception Extension. `http://freedesktop.org/~stukreit/xevie.html`.

[Mar] M. Maria. Little CMS Engine 1.12 API Definition. Technical report. `http://www.littlecms.com/lcmsapi.txt`.

[McQ02] Jim McQuillan. LTSP - Linux Terminal Server Project, Version 3.0. Technical report, March 2002. `http://www.ltsp.org/documentation/ltsp-3.0-4-en.html`.

[MS01] Bart Massey and Jamey Sharp. XCB: An X protocol c binding.

In *XFree86 Technical Conference*, Oakland, CA, November 2001. USENIX.

[NR04] Peter Nilsson and David Reveman. Glitz: Hardware Accelerated Image Compositing using OpenGL. In *FREENIX Track, 2004 Usenix Annual Technical Conference*, Boston, MA, July 2004. USENIX.

[Pac01a] Keith Packard. Design and Implementation of the X Rendering Extension. In *FREENIX Track, 2001 Usenix Annual Technical Conference*, Boston, MA, June 2001. USENIX.

[Pac01b] Keith Packard. The Xft Font Library: Architecture and Users Guide. In *XFree86 Technical Conference*, Oakland, CA, November 2001. USENIX.

[Pac02] Keith Packard. Font Configuration and Customization for Open Source Systems. In *2002 Gnome User's and Developers European Conference*, Seville, Spain, April 2002. Gnome.

[PD84] Thomas Porter and Tom Duff. Compositing Digital Images. *Computer Graphics*, 18(3):253–259, July 1984.

[PG03] Keith Packard and James Gettys. X Window System Network Performance. In *FREENIX Track, 2003 Usenix Annual Technical Conference*, San Antonio, TX, June 2003. USENIX.

[Pla00] J. Platt. Optimal filtering for patterned displays. *IEEE Signal Processing Letters*, 7(7):179–180, 2000.

[SG92] Robert W. Scheifler and James Gettys. *X Window System*. Digital Press, third edition, 1992.

[Smi92] John Smith. The Multi-Threaded X Server. *The X Resource*, 1:73–89, Winter 1992.

[SRRK02] D. Smith, A. Raab, D. Reed, and A. Kay. Croquet: The Users Manual, October 2002. `http://glab.cs.uni-magdeburg.de/~croquet/downloads/Croquet0.1.pdf`.

[SSF+96] Robert N.C. Shelley, Robert W. Scheifler, Ben Fahy, Jim Fulton, Keith Packard, Joe Mauro, Richard Hennessy, and Tom Vaughn. X Image Extension Protocol Version 5.02. X consortium standard, 1996.

[Wal04] Eamon Walsh. Integrating XFree86 With Security-Enhanced Linux. In *X Developers Conference*, Cambridge, MA, April 2004. `http://freedesktop.org/Software/XDevConf/x-security-walsh.pdf`.

[WP03] Carl Worth and Keith Packard. Xr: Cross-device Rendering for Vector Graphics. In *Proceedings of the Ottawa Linux Symposium*, Ottawa, ON, July 2003. OLS.

IA64-Linux perf tools for IO dorks

Examples of IA-64 PMU usage

Grant Grundler
Hewlett-Packard
iod00d@hp.com
grundler@parisc-linux.org

Abstract

Itanium processors have very sophisticated performance monitoring tools integrated into the CPU. McKinley and Madison Itanium CPUs have over three hundred different types of events they can filter, trigger on, and count. The restrictions on which combinations of triggers are allowed is daunting and varies across CPU implementations. Fortunately, the tools hide this complicated mess. While the tools prevent us from shooting ourselves in the foot, it's not obvious how to use those tools for measuring kernel device driver behaviors.

IO driver writers can use pfmon to measure two key areas generally not obvious from the code: MMIO read and write frequency and precise addresses of instructions regularly causing L3 data cache misses. Measuring MMIO reads has some nuances related to instruction execution which are relevant to understanding ia64 and likely ia32 platforms. Similarly, the ability to pinpoint exactly which data is being accessed by drivers enables driver writers to either modify the algorithms or add prefetching directives where feasible. I include some examples on how I used pfmon to measure NIC drivers and give some guidelines on use.

q-syscollect is a "gprof without the pain" kind of tool. While q-syscollect uses the same kernel perfmon subsystem as pfmon, the former works at a higher level. With some knowledge about how the kernel operates, q-syscollect can collect call-graphs, function call counts, and percentage of time spent in particular routines. In other words, pfmon can tell us how much time the CPU spends stalled on d-cache misses and q-syscollect can give us the call-graph for the worst offenders.

Updated versions of this paper will be available from `http://iou.parisc-linux.org/ols2004/`

1 Introduction

Improving the performance of IO drivers is really not that easy. It usually goes something like:

1. Determine which workload is relevant
2. Set up the test environment
3. Collect metrics
4. Analyze the metrics
5. Change the code based on theories about the metrics
6. Iterate on Collect metrics

This paper attempts to make the collect-analyze-change loop more efficient for three

obvious things: MMIO reads, MMIO writes, and cache line misses.

MMIO reads and writes are easier to locate in Linux code than for other OSs which support memory-mapped IO—just search for *readl()* and *writel()* calls. But `pfmon` [1] can provide statistics of actual behavior and not just where in the code MMIO space is touched.

Cache line misses are hard to detect. None of the regular performance tools I've used can precisely tell where CPU stalls are taking place. We can guess some of them based on data usage—like spin locks ping-ponging between CPUs. This requires a level of understanding that most of us mere mortals don't possess. Again, `pfmon` can help out here.

Lastly, getting an overview of system performance and getting run-time call graph usually requires compiler support that gcc doesn't provide. q-tools[4] can provide that information. Driver writers can then manually adjust the code knowing where the "hot spots" are.

1.1 `pfmon`

The author of `pfmon`, Stephane Eranian [2], describes `pfmon` as "the performance tool for IA64-Linux which exploits all the features of the IA-64 Performance Monitoring Unit (PMU)." `pfmon` uses a command line interface and does not require any special privilege to run. `pfmon` can monitor a single process, a multi-threaded process, multi-processes workloads and the entire system.

`pfmon` is the user command line interface to the kernel perfmon subsystem. perfmon does the ugly work of programming the PMU. Perfmon is versioned separately from `pfmon` command. When in doubt, use the perfmon in the latest 2.6 kernel.

There are two major types of measurements: counting and sampling. For counting, `pfmon` simply reports the number of occurrences of the desired events during the monitoring period. `pfmon` can also be configured to sample at certain intervals information about the execution of a command or for the entire system. It is possible to sample any events provided by the underlying PMU.

The information recorded by the PMU depends on what the user wants. `pfmon` contains a few preset measurements but for the most part the user is free to set up custom measurements. On Itanium2, `pfmon` provides access to all the PMU advanced features such as opcode matching, range restrictions, the Event Address Registers (EAR) and the Branch Trace Buffer.

1.2 `pfmon` command line options

Here is a summary of command line options used in the examples later in this paper:

–us-c use the US-style comma separator for large numbers.

–cpu-list=0 bind `pfmon` to CPU 0 and only count on CPU 0

–pin-command bind the command at the end of the command line to the same CPU as `pfmon`.

–resolve-addr look up addresses and print the symbols

–long-smpl-periods=2000 take a sample of every 2000th event.

–smpl-periods-random=0xfff:10 randomize the sampling period. This is necessary to avoid bias when sampling repetitive behaviors. The first value is the mask of bits to randomize (e.g., 0xfff) and the second value is initial seed (e.g., 10).

-k kernel only.

–system-wide measure the entire system (all processes and kernel)

Parameters only available on a to-be-released `pfmon` v3.1:

–smpl-module=dear-hist-itanium2 This particular module is to be used ONLY in conjunction with the Data EAR (Event Address Registers) and presents recorded samples as histograms about the cache misses. By default, the information is presented in the instruction view but it is possible to get the data view of the misses also.

-e data_ear_cache_lat64 pseudo event for memory loads with latency $\geq$ 64 cycles. The real event is `DATA_EAR_EVENT` (counts the number of times Data EAR has recorded something) and the pseudo event expresses the latency filter for the event. Use "`pfmon -ldata_ear_cache*`" to list all valid values. Valid values with McKinley CPU are powers of two ($4 - 4096$).

1.3 q-tools

The author of q-tools, David Mosberger [5], has described q-tools as "gprof without the pain."

q-tools package contains `q-syscollect`, `q-view`, `qprof`, and `q-dot`. `q-syscollect` collects profile information using kernel perfmon subsystem to sample the PMU. `q-view` will present the data collected in both flat-profile and call graph form. `q-dot` displays the call-graph in graphical form. Please see the `qprof` [6] website for details on `qprof`.

`q-syscollect` depends on the kernel perfmon subsystem which is included in all 2.6 Linux kernels. Because `q-syscollect` uses the PMU, it has the following advantages over other tools:

- no special kernel support needed (besides perfmon subsystem).
- provides call-graph of kernel functions
- can collect call-graphs of the kernel while interrupts are blocked.
- measures multi-threaded applications
- data is collected per-CPU and can be merged
- instruction level granularity (not bundles)

2 Measuring MMIO Reads

Nearly every driver uses MMIO reads to either flush MMIO writes, flush in-flight DMA, or (most obviously) collect status data from the IO device directly. While use of MMIO read is necessary in most cases, it should be avoided where possible.

2.1 Why worry about MMIO Reads?

MMIO reads are expensive—how expensive depends on speed of the IO bus, the number bridges the read (and its corresponding read return) has to cross, how "busy" each bus is, and finally how quickly the device responds to the read request. On most architectures, one can precisely measure the cost by measuring a loop of MMIO reads and calling `get_cycles()` before/after the loop.

I've measured anywhere from 1μs to 2μs per read. In practical terms:

- $\sim$ 500–600 cycles on an otherwise-idle 400 MHz PA-RISC machine.

- $\sim$ 1000 cycles on a 450 MHz Pentium machine which included crossing a PCI-PCI bridge.
- $\sim$ 900–1000 cycles on a 800 MHz IA64 HP ZX1 machine.

And for those who still don't believe me, try watching a DVD movie after turning DMA off for an IDE DVD player:

```
hdparm -d 0 /dev/cdrom
```

By switching the IDE controller to use PIO (Programmed I/O) mode, all data will be transferred to/from host memory under CPU control, byte (or word) at a time. `pfmon` can measure this. And `pfmon` looks broken when it displays three and four digit "Average Cycles Per Instruction" (CPI) output.

2.2 Eh? Memory Reads don't stall?

They do. But the CPU and PMU don't "realize" the stall until the next memory reference. The CPU continues execution until memory order is enforced by the acquire semantics in the MMIO read. This means the **Data Event Address Registers record the next stalled memory reference due to memory ordering constraints, not the MMIO read**. One has to look at the instruction stream carefully to determine which instruction actually caused the stall.

This also means the following sequence doesn't work exactly like we expect:

```
writel(CMD,addr);
readl(addr);
udelay(1);
y = buf->member;
```

The problem is the value returned by `read(x)` is never consumed. **Memory ordering imposes no constraint on non-load/store instructions.** Hence `udelay(1)` begins before the CPU stalls. The CPU will stall on `buf->member` because of memory ordering restrictions if the `udelay(1)` completes before `readl(x)` is retired. Drop the `udelay(1)` call and `pfmon` will always see the stall caused by MMIO reads on the next memory reference.

Unfortunately, the IA32 Software Developer's Manual[3] Volume 3, Chapter 7.2 "MEMORY ORDERING" is silent on the issue of how MMIO (uncached accesses) will (or will not) stall the instruction stream. This document is very clear on how "IO Operations" (e.g., IN/OUT) will stall the instruction pipeline until the read return arrives at the CPU. A direct response from Intel(R) indicated `readl()` does not stall like IN or OUT do and IA32 has the same problem. The Intel® architect who responded did hedge the above statement claiming a "udelay(10) will be as close as expected" for an example similar to mine. Anyone who has access to a frontside bus analyzer can verify the above statement by measuring timing loops between uncached accesses. I'm not that privileged and have to trust Intel® in this case.

For IA64, we considered putting an extra burden on `udelay` to stall the instruction stream until previous memory references were retired. We could use dummy loads/stores before and after the actual delay loop so memory ordering could be used to stall the instruction pipeline. That seemed excessive for something that we didn't have a bug report for.

Consensus was adding `mf.a` (memory fence) instruction to `readl()` should be sufficient. The architecture only requires `mf.a` serve as an ordering token and need not cause any delays of its own. In other words, the implementation is platform specific. `mf.a` has not been added to `readl()` yet because every-

thing was working without so far.

2.3 `pfmon -e uc_loads_retired`

IO accesses are generally the only uncached references made on IA64-linux and normally will represent MMIO reads. The basic measurement will tell us roughly how many cycles the CPU stalls for MMIO reads. Get the number of MMIO reads per sample period and then multiply by the actual cycle counts a MMIO read takes for the given device. One needs to measure MMIO read cost by using a CPU internal cycle counter and hacking the kernel to read a harmless address from the target device a few thousand times.

In order to make statements about per transaction or per interrupt, we need to know the cumulative number of transactions or interrupts processed for the sample period. `pktgen` is straightforward in this regard since `pktgen` will print transaction statistics when a run is terminated. And one can record `/proc/interrupts` contents before and after each `pfmon` run to collect interrupt events as well.

Drawbacks to the above are one assumes a homogeneous driver environment; i.e., only one type of driver is under load during the test. I think that's a fair assumption for development in most cases. Bridges (e.g., routing traffic across different interconnects) are probably the one case it's not true. One has to work a bit harder to figure out what the counts mean in that case.

For other benchmarks, like SpecWeb, we want to grab `/proc/interrupt` and networking stats before/after `pfmon` runs.

2.4 tg3 Memory Reads

In summary, Figure 1 shows tg3 is doing $2749675/(1834959 - 918505) \approx 3$ MMIO reads per interrupt and averaging about $5000000/(1834959 - 918505) \approx 5$ packets per interrupt. This is with the BCM5701 chip running in PCI mode at 66MHz:64-bit.

Based on code inspection, here is a break down of where the MMIO reads occur in temporal order:

1. `tg3_interrupt()` flushes MMIO write to `MAILBOX_INTERRUPT_0`
2. `tg3_poll()` → `tg3_enable_ints()` → `tw32(TG3PCI_MISC_HOST_CTRL)`
3. `tg3_enable_ints()` flushes MMIO write to `MAILBOX_INTERRUPT_0`

It's obvious when inspecting `tw32()`, the BCM5701 chip has a serious bug. Every call to `tw32()` on BCM5701 requires a MMIO read to follow the MMIO write. Only writes to mailbox registers don't require this and a different routine is used for mailbox writes.

Given the NIC was designed for zero MMIO reads, this is pretty poor performance. Using a BCM5703 or BCM5704 would avoid the MMIO read in tw32().

I've exchanged email with David Miller and Jeff Garzik (tg3 driver maintainers). They have valid concerns with portability. We agree tg3 could be reduced to one MMIO read after the last MMIO write (to guarantee interrupts get re-enabled).

One would need to use the "tag" field in the status block when writing the mail box register to indicate which "tag" the CPU most recently

```
gsyprf3:~# pfmon -e uc_loads_retired -k --system-wide \
-- /usr/src/pktgen-testing/pktgen-single-tg3
Adding devices to run.
Configuring devices
Running... ctrl^C to stop
 57:     918505          0  IO-SAPIC-level  eth1
Result: OK: 7613693(c7613006+d687) usec, 5000000 (64byte) 656771pps 320Mb/sec
(336266752bps)  errors: 0
 57:    1834959          0  IO-SAPIC-level  eth1
CPU0                     2749675 UC_LOADS_RETIRED
CPU1                        1175 UC_LOADS_RETIRED
}
```

Figure 1: tg3 v3.6 MMIO reads with pktgen/IRQ on same CPU

saw. Using Message Signaled Interrupts (MSI) instead of Line based IRQs would guarantee the most recent status block update (transferred via DMA writes) would be visible to the CPU before `tg3_interrupt()` gets called.

The protocol would allow correct operation without using MSI, too.

2.5 Benchmarking, `pfmon`, and CPU bindings

The purpose of binding `pktgen` to CPU1 is to verify the transmit code path is NOT doing any MMIO reads. We split the transmit code path and interrupt handler across CPUs to narrow down which code path is performing the MMIO reads. This change is not obvious from Figure 2 output since tg3 only performs MMIO reads from CPU 0 (`tg3_interrupt()`).

But in Figure 2, performance goes up 30%! Offhand, I don't know if this is due to CPU utilization (pktgen and `tg3_interrupt()` contending for CPU cycles) or if DMA is more efficient because of cache-line flows. When I don't have any deadlines looming, I'd like to determine the difference.

2.6 e1000 Memory Reads

e1000 version 5.2.52-k4 has a more efficient implementation than tg3 driver. In a nut shell, MMIO reads are pretty much irrelevant to the pktgen workload with e1000 driver using default values.

Figure 3 shows e1000 performs $173315/(703829 - 622143) \approx 2$ MMIO reads per interrupt and $5000000/(703829 - 622143) \approx 61$ packets per interrupt.

Being the curious soul I am, I tracked down the two MMIO reads anyway. One is in the interrupt handler and the second when interrupts are re-enabled. It looks like e1000 will always need at least 2 MMIO reads per interrupt.

3 Measuring MMIO Writes

3.1 Why worry about MMIO Writes?

MMIO writes are clearly not as significant as MMIO reads. Nonetheless, every time a driver writes to MMIO space, some subtle things happen. There are four minor issues to think about: memory ordering, PCI bus utilization, filling outbound write queues, and stalling MMIO reads longer than necessary.

```
gsyprf3:~# pfmon -e uc_loads_retired -k --system-wide \
-- /usr/src/pktgen-testing/pktgen-single-tg3
Adding devices to run.
Configuring devices
Running... ctrl^C to stop
 57:    5809687          0  IO-SAPIC-level  eth1
Result: OK: 5914889(c5843865+d71024) usec, 5000000 (64byte) 845451pps 412Mb/se
c (432870912bps)  errors: 0
 57:    6427969          0  IO-SAPIC-level  eth1
CPU0                    1855253 UC_LOADS_RETIRED
CPU1                        950 UC_LOADS_RETIRED
```

Figure 2: tg3 v3.6 MMIO reads with pktgen/IRQ on diff CPU

```
gsyprf3:~# pfmon -e uc_loads_retired -k --system-wide \
-- /usr/src/pktgen-testing/pktgen-single-e1000
Configuring devices
Running... ctrl^C to stop
 59:     622143          0  IO-SAPIC-level  eth3
Result: OK: 10228738(c9990105+d238633) usec, 5000000 (64byte) 488854pps 238Mb/
sec (250293248bps)  errors: 81669
 59:     703829          0  IO-SAPIC-level  eth3
CPU0                     173315 UC_LOADS_RETIRED
CPU1                       1422 UC_LOADS_RETIRED
```

Figure 3: MMIO reads for e1000 v5.2.52-k4

First, memory ordering is enforced since PCI requires strong ordering of MMIO writes. This means the MMIO write will push all previous regular memory writes ahead. This is not a serious issue but it can make a MMIO write take longer.

MMIO writes are short transactions (i.e., much less than a cache-line). The PCI bus setup time to select the device, send the target address and data, and disconnect measurably reduces PCI bus utilization. It typically results in six or more PCI bus cycles to send four (or eight) bytes of data. On systems which strongly order DMA Read Returns and MMIO Writes, the latter will also interfere with DMA flows by interrupting in-flight, outbound DMA.

If the IO bridge (e.g., PCI Bus controller) nearest the CPU has a full write queue, the CPU will stall. The bridge would normally queue the MMIO write and then tell the CPU it's done. The chip designers normally make the write queue deep enough so the CPU never needs to stall. But drivers that perform many MMIO writes (e.g., use door bells) and burst many of MMIO writes at a time, could run into a worst case.

The last concern, stalling MMIO reads longer than normal, exists because of PCI ordering rules. MMIO reads and MMIO writes are strongly ordered. E.g., if four MMIO writes are queued before a MMIO read, the read will wait until all four MMIO write transactions have completed. So instead of say 1000 CPU cycles, the MMIO read might take more than 2000 CPU cycles on current platforms.

3.2 `pfmon -e uc_stores_retired`

`pfmon` counts MMIO Writes with no surprises.

3.3 tg3 Memory Writes

Figure 4 shows tg3 does about 10M MMIO writes to send 5M packets. However, we can break the MMIO writes down into base level (feed packets onto transmit queue) and `tg3_interrupt` which handles TX (and RX) completions. Knowing which code path the MMIO writes are in helps track down usage in the source code.

Output in Figure 5 is after hacking the `pktgen-single-tg3` script to bind `pktgen` kernel thread to CPU 1 when eth1 is directing interrupts to CPU 0. The distribution between TX queue setup and interrupt handling is obvious now. CPU 0 is handling interrupts and performs $3013580/(5803789 - 5201193) \approx 5$ MMIO writes per interrupt. CPU 1 is handling TX setup and performs $5000376/5000000 \approx 1$ MMIO write per packet.

Again, as noted in section 2.5, binding pktgen thread to one CPU and interrupts to another, changes the performance dramatically.

3.4 e1000 Memory Writes

Figure 6 shows $248891/(991082 - 908366) \approx 3$ MMIO writes per interrupt and $5001303/5000000 \approx 1$ MMIO write per packet. In other words, slightly better than tg3 driver. Nonetheless, the hardware can't push as many packets. One difference is the e1000 driver is pushing data to a NIC behind a PCI-PCI Bridge.

Figure 7 shows a ≈40% improvement in throughput[1] for pktgen without a PCI-PCI Bridge in the way. Note the ratios of MMIO writes per interrupt and MMIO writes per

[1] This demonstrates how the distance between the IO device and CPU (and memory) directly translates into latency and performance.

```
gsyprf3:~# pfmon -e uc_stores_retired -k --system-wide -- /usr/src/pktgen-test
ing/pktgen-single-tg3
Adding devices to run.
Configuring devices
Running... ctrl^C to stop
 57:    4284466          0  IO-SAPIC-level  eth1
Result: OK: 7611689(c7610900+d789) usec, 5000000 (64byte) 656943pps 320Mb/sec
(336354816bps)  errors: 0
 57:    5198436          0  IO-SAPIC-level  eth1
CPU0                    9570269 UC_STORES_RETIRED
CPU1                        445 UC_STORES_RETIRED
```

Figure 4: tg3 v3.6 MMIO writes

```
gsyprf3:~# pfmon -e uc_stores_retired -k --system-wide -- /usr/src/
pktgen-testing/pktgen-single-tg3
Adding devices to run.
Configuring devices
Running... ctrl^C to stop
 57:    5201193          0  IO-SAPIC-level  eth1
Result: OK: 5880249(c5811180+d69069) usec, 5000000 (64byte) 850340pps 415Mb
/sec (435374080bps)  errors: 0
 57:    5803789          0  IO-SAPIC-level  eth1
CPU0                    3013580 UC_STORES_RETIRED
CPU1                    5000376 UC_STORES_RETIRED
```

Figure 5: tg3 v3.6 MMIO writes with pktgen/IRQ split across CPUs

```
gsyprf3:~# pfmon -e uc_stores_retired -k --system-wide -- /usr/src/
pktgen-testing/pktgen-single-e1000
Running... ctrl^C to stop
 59:     908366          0  IO-SAPIC-level  eth3
Result: OK: 10340222(c10104719+d235503) usec, 5000000 (64byte) 483558pps 236Mb
/sec (247581696bps)  errors: 82675
 59:     991082          0  IO-SAPIC-level  eth3
CPU0                     248891 UC_STORES_RETIRED
CPU1                    5001303 UC_STORES_RETIRED
```

Figure 6: MMIO writes for e1000 v5.2.52-k4

```
gsyprf3:~# pfmon -e uc_stores_retired -k --system-wide -- /usr/src/pktgen-test
ing/pktgen-single-e1000
Running... ctrl^C to stop
 71:          3          0  IO-SAPIC-level  eth7
Result: OK: 7491358(c7342756+d148602) usec, 5000000 (64byte) 667467pps 325Mb/s
ec (341743104bps)  errors: 59870
 71:      59907          0  IO-SAPIC-level  eth7
CPU0                     180406 UC_STORES_RETIRED
CPU1                    5000939 UC_STORES_RETIRED
```

Figure 7: e1000 v5.2.52-k4 MMIO writes without PCI-PCI Bridge

packet are the same. I doubt the MMIO reads and MMIO writes are the limiting factors. More likely DMA access to memory (and thus TX/RX descriptor rings) limits NIC packet processing.

4 Measuring Cache-line Misses

The Event Address Registers[2] (EAR) can only record one event at a time. What is so interesting about them is that they record precise information about data cache misses. For instance for a data cache miss, you get the:

- address of the instruction, likely a load
- address of the target data
- latency in cycles to resolve the miss

The information pinpoints the source of the miss, not the consequence (i.e., the stall).

The Data EAR (DEAR) can also tell us about MMIO reads via sampling. The DEAR can only record *loads* that miss, not stores. Of course, MMIO reads always miss because they are uncached. This is interesting if we want to track down which MMIO addresses are "hot." It's usually easier to track down usage in source code knowing which MMIO address is referenced.

Collecting with DEAR sampling requires two parameters be tweaked to statistically improve the samples. One is the frequency at which Data Addresses are recorded and the other is the threshold (how many CPU cycles latency).

Because we know the latency to L3 is about 21 cycles, setting the EAR threshold to a value higher (e.g., 64 cycles) ensures only the load misses accessing main memory will be captured. This is how to select which level of cacheline misses one samples.

While high threshholds (e.g., 64 cycles) will show us where the longest delays occur, it will not show us the worst offenders. Doing a second run with a lower threshold (e.g., 4 cycles) shows all L1, L2, and L3 cache misses and provides a much broader picture of cache utilization.

When sampling events with low threshholds, we will get saturated with events and need to reduce the number of events actually sampled to every 5000th. The appropriate value will depend on the workload and how patient one is. The workload needs to be run long enough to be statistically significant and the sampling period needs to be high enough to not significantly perturb the workload.

4.1 tg3 Data Cache misses > 64 cycles

For the output in Figure 8, I've iteratively decreased the smpl-periods until I noticed the total pktgen throughput starting to drop. Figure 8 output only shows the tg3 interrupt code path since `pfmon` is bound to CPU 0. Normally, it would be useful to run this again with `cpu-list=1`. We could then see what the TX code path and pktgen are doing.

Also, the `pin-command` option in this example doesn't do anything since `pktgen-single-tg3` directs a `pktgen` kernel thread bound CPU 1 to do the real work. I've included the option only to make people aware of it.

4.2 tg3 Data Cache misses > 4 cycles

Figure 9 puts the `lat64` output in Figure 8 into better perspective. It shows tg3 is spending more time for L1 and L2 misses than L3 misses

[2] **`pfmon` v3.1 is the first version to support EAR and is expected to be available in August, 2004.**

```
gsyprf3:~# pfmon31 --us-c --cpu-list=0 --pin-command --resolve-addr \
           --smpl-module=dear-hist-itanium2 \
           -e data_ear_cache_lat64 --long-smpl-periods=500 \
           --smpl-periods-random=0xfff:10 --system-wide \
           -k -- /usr/src/pktgen-testing/pktgen-single-tg3
added event set 0
only kernel symbols are resolved in system-wide mode
Adding devices to run.
Configuring devices
Running... ctrl^C to stop
 57:    7209769          0  IO-SAPIC-level  eth1
Result: OK: 5915877(c5845032+d70845) usec, 5000000 (64byte) 845308pps 412Mb/sec
(432797696bps)  errors: 0
 57:    7827812          0  IO-SAPIC-level  eth1
# total_samples 672
# instruction addr view
# sorted by count
# showing per per distinct value
# %L2  : percentage of L1 misses that hit L2
# %L3  : percentage of L1 misses that hit L3
# %RAM : percentage of L1 misses that hit memory
# L2   :  5 cycles load latency
# L3   : 12 cycles load latency
# sampling period: 500
#count %self   %cum    %L2    %L3    %RAM   instruction addr
  38  5.65%   5.65%  0.00%  0.00% 100.00% 0xa000000100009141 ia64_spinlock_contention
                                                            +0x21<kernel>
  36  5.36%  11.01%  0.00%  0.00% 100.00% 0xa00000020003e580 tg3_interrupt[tg3]+0xe0<kernel>
  32  4.76%  15.77%  0.00%  0.00% 100.00% 0xa000000200034770 tg3_write_indirect_reg32[tg3]
                                                            +0x90<kernel>
  32  4.76%  20.54%  0.00%  0.00% 100.00% 0xa00000020003e640 tg3_interrupt[tg3]+0x1a0<kernel>
  30  4.46%  25.00%  0.00%  0.00% 100.00% 0xa000000200034e91 tg3_enable_ints[tg3]+0x91<kernel>
  29  4.32%  29.32%  0.00%  0.00% 100.00% 0xa00000020003e510 tg3_interrupt[tg3]+0x70<kernel>
  28  4.17%  33.48%  0.00%  0.00% 100.00% 0xa00000020003d1a0 tg3_tx[tg3]+0x2e0<kernel>
  27  4.02%  37.50%  0.00%  0.00% 100.00% 0xa00000020003cfa0 tg3_tx[tg3]+0xe0<kernel>
  24  3.57%  41.07%  0.00%  0.00% 100.00% 0xa00000020003cfd1 tg3_tx[tg3]+0x111<kernel>
  21  3.12%  44.20%  0.00%  0.00% 100.00% 0xa000000200034e60 tg3_enable_ints[tg3]+0x60<kernel>
   .
   .
   .
# level 0 : counts=0 avg_cycles=0.0ms   0.00%
# level 1 : counts=0 avg_cycles=0.0ms   0.00%
# level 2 : counts=672 avg_cycles=0.0ms 100.00%
approx cost: 0.0s
```

Figure 8: tg3 v3.6 lat64 output

```
gsyprf3:~# pfmon31 --us-c --cpu-list=0 --resolve-addr --smpl-module=dear-hist-itanium2 \
-e data_ear_cache_lat4 --long-smpl-periods=5000 --smpl-periods-random=0xfff:10 \
--system-wide -k -- /usr/src/pktgen-testing/pktgen-single-tg3
added event set 0
only kernel symbols are resolved in system-wide mode
Adding devices to run.
Configuring devices
Running... ctrl^C to stop
 57:    8484552          0  IO-SAPIC-level  eth1
Result: OK: 5938001(c5866437+d71564) usec, 5000000 (64byte) 842034pps 411Mb/sec
            (431121408bps)  errors: 0
 57:    9093642          0  IO-SAPIC-level  eth1
# total_samples 795
# instruction addr view
# sorted by count
# showing per per distinct value
# %L2  : percentage of L1 misses that hit L2
# %L3  : percentage of L1 misses that hit L3
# %RAM : percentage of L1 misses that hit memory
# L2   :  5 cycles load latency
# L3   : 12 cycles load latency
# sampling period: 5000
# #count %self    %cum     %L2     %L3    %RAM   instruction addr
      95  11.95%  11.95%   0.00%  98.95%   1.05% 0xa00000020003d150 tg3_tx[tg3]+0x290<kernel>
      83  10.44%  22.39%  93.98%   4.82%   1.20% 0xa00000020003d030 tg3_tx[tg3]+0x170<kernel>
      21   2.64%  25.03%   0.00%  95.24%   4.76% 0xa0000001000180f0 ia64_handle_irq+0x170<kernel>
      20   2.52%  27.55%   5.00%  80.00%  15.00% 0xa00000020003d040 tg3_tx[tg3]+0x180<kernel>
      18   2.26%  29.81%  50.00%  11.11%  38.89% 0xa00000020003cfa0 tg3_tx[tg3]+0xe0<kernel>
      17   2.14%  31.95%   0.00%   0.00% 100.00% 0xa00000020003e671 tg3_interrupt[tg3]
                                                                    +0x1d1<kernel>
      17   2.14%  34.09%   0.00% 100.00%   0.00% 0xa00000020003e700 tg3_interrupt[tg3]
                                                                    +0x260<kernel>
      16   2.01%  36.10%  56.25%  43.75%   0.00% 0xa000000100012160 ia64_leave_kernel
                                                                    +0x180<kernel>
      16   2.01%  38.11%  62.50%   0.00%  37.50% 0xa00000020003cf60 tg3_tx[tg3]+0xa0<kernel>
      15   1.89%  40.00%  86.67%   6.67%   6.67% 0xa00000020003cfd0 tg3_tx[tg3]+0x110<kernel>
      15   1.89%  41.89%   0.00%   0.00% 100.00% 0xa000000100016041 do_IRQ+0x1a1<kernel>
      15   1.89%  43.77%   0.00%  53.33%  46.67% 0xa00000020003e370 tg3_poll[tg3]+0x350<kernel>
      .
      .
      .
# level 0 : counts=226 avg_cycles=0.0ms  28.43%
# level 1 : counts=264 avg_cycles=0.0ms  33.21%
# level 2 : counts=305 avg_cycles=0.0ms  38.36%
approx cost: 0.0s
```

Figure 9: tg3 v3.6 lat4 output

and in only two locations. Adding one prefetch to pull data from L3 into L2 would help for the top offender. One needs to figure out which bit of data each recorded access refers to and determine how early one can prefetch that data.

We can also rule out MMIO accesses as the top culprit. `tg3_interrupt+0x1d1` could be an MMIO read but it doesn't show up in Figure 8 like `tg3_write_indirect_reg32` does.

Note `smpl-periods` is 10x higher in Figure 9 than in Figure 8. Collecting 10x more samples with `lat4` definitely disturbs the workload.

5 q-tools

`q-syscollect` and `q-view` are trivial to use. An example and brief explanation for kernel usage follow.

Please remember most applications spend most of the time in user space and not in the kernel. q-tools is especially good in user space.

5.1 `q-syscollect`

```
q-syscollect -c 5000 -C 5000 -t
20 -k
```

This will collect system wide kernel data during the 20 second period. Twenty to thrity seconds is usually long enough to get sufficient accuracy[3]. However, if the workload generates a very wide call graph with even distribution, one will likely need to sample for longer periods to get accuracy in the $\pm 1\%$ range. When in doubt, try sampling for longer periods to see if the call-counts change significantly.

[3]See Page 7 of the David Mosberger's Gelato talk [4] for a nice graph on accuracy which *only applies to his example.*

The `-c` and `-C` set the call sample rate and code sample rate respectively. The call sample rate is used to collect function call counts. This is one of the key differences compared to traditional profiling tools: q-syscollect obtains call-counts in a statistical fashion, just as has been done traditionally for the execution-time profile. The code sample rate is used to collect a flat profile (`CPU_CYCLES` by default).

The `-e` option allows one to change the event used to sample for the flat profile. The default is to sample CPU_CYCLES event. This provides traditional execution time in the flat profile.

The data is stored in the current directory under `.q/` directory. The next section demonstrates how `q-view` displays the data.

5.2 `q-view`

I was running the netperf [7] TCP_RR test in the background to another server when I collected the following data. As Figure 10 shows, this particular TCP_RR test isn't costing many cycles in tg3 driver. Or, at least not ones I can measure.

`tg3_interrupt()` shows up in the flat profile with 0.314 seconds time associated with it. The time measurement is only possible because `handle_IRQ_event()` re-enables interrupts if the IRQ handler is not registered with `SA_INTERRUPT` (to indicate latency sensitive IRQ handler). `do_IRQ()` and other functions in that same call graph do NOT have any time measurements because interrupts are disabled. As noted before, the call-graph is sampled using a different part of the PMU than the part which samples the flat profile.

Lastly, I've omitted the trailing output of `q-view` which explains the fields and columns more completely. Read that first be-

```
gsyprf3:~# q-view .q/kernel-cpu0.info | more
Flat profile of CPU_CYCLES in kernel-cpu0.hist#0:
 Each histogram sample counts as 200.510u seconds
% time      self     cumul     calls self/call  tot/call name
 68.88     13.41     13.41      215k     62.5u     62.5u default_idle
  2.90      0.56     13.97      431k     1.31u     1.31u finish_task_switch
  2.50      0.49     14.46      233k     2.09u     4.89u tg3_poll
  1.77      0.35     14.80     1.38M      251n      268n ipt_do_table
  1.61      0.31     15.12      240k     1.31u     1.31u tg3_interrupt
  1.51      0.29     15.41      240k     1.22u     5.95u net_rx_action
  .
  .
  .
Call-graph table:
index %time      self  children         called     name
                                                    <spontaneous>
[176]  69.4     30.5m      13.4         -           cpu_idle
                29.5m     0.285      231k/457k      schedule [164]
                10.0m      0.00      244k/244k      check_pgt_cache [178]
                 13.4      0.00      215k/215k      default_idle [177]
-----------------------------------------------
   .
   .
   .
-----------------------------------------------
                0.293      1.14      240k           __do_softirq [40]
[56]    7.4     0.293      1.14      240k           net_rx_action
                0.487     0.649      233k/233k      tg3_poll [57]
-----------------------------------------------
                0.487     0.649      233k           net_rx_action [56]
[57]    5.9     0.487     0.649      233k           tg3_poll
                    -      0.00      229k/229k      tg3_enable_ints [133]
                97.7m     0.552      225k/225k      tg3_rx [61]
                    -      0.00      227k/227k      tg3_tx [58]
-----------------------------------------------
   .
   .
   .
-----------------------------------------------
                    -      1.88      348k           ia64_leave_kernel [10]
[11]    9.7         -      1.88      348k           ia64_handle_irq
                    -      1.52      239k/240k      do_softirq [39]
                    -     0.367      356k/356k      do_IRQ [12]
-----------------------------------------------
   .
   .
   .
```

Figure 10: `q-view` output for TCP_RR over tg3 v3.6

fore going through the rest of the output.

6 Conclusion

6.1 More pfmon examples

CPU L2 cache misses in one kernel function
`pfmon --verb -k \ --irange=sba_alloc_range \ -el2_misses --system-wide \ --session-timeout=10`
Show all L2 cache misses in `sba_alloc_range`. This is interesting since `sba_alloc_range()` walks a bitmap to look for "free" resources. One can instead specify `-el3_misses` since L3 cache misses are much more expensive.

CPU 1 memory loads
`pfmon --us-c \ --cpu-list=1 \ -e loads_retired \ -k --system-wide \ -- /tmp/pktgen-single` Only count memory loads on CPU 1. This is useful for when we can bind the interrupt to CPU 1 and the workload to a different CPU. This lets us separate interrupt path from base level code, i.e., when is the load happening (before or after DMA occurred) and which code path should one be looking more closely at.

List EAR events supported `pfmon -lear` List all EAR types supported by `pfmon`[4].

More info on Event `pfmon -i DATA_EAR_TLB_ALL` `pfmon` can provide more info on particular events it supports.

[4] EAR isn't supported until `pfmon v3.1`

6.2 And thanks to...

Special thanks to Stephane Eranian [2] for dedicating so much time to the perfmon kernel driver and associated tools. People might think the PMU does it all—but only with a lot of SW driving it. His review of this paper caught some good bloopers. This talk only happened because I sit across the aisle from him and could pester him regularly.

Thanks to David Mosberger[5] for putting together q-tools and making it so trivial to use.

In addition, in no particular order:
Christophe de Dinechin, Bjorn Helgaas, Matthew Wilcox, Andrew Patterson, Al Stone, Asit Mallick, and James Bottomley for reviewing this document or providing technical guidance.

Thanks also to the OLS staff for making this event happen every year.

My apologies if I omitted other contributors.

References

[1] perfmon homepage, `http://www.hpl.hp.com/research/linux/perfmon/`

[2] Stephane Eranian, `http://www.gelato.org/community/gelato_meeting.php?id=CU2004#talk22`

[3] The IA-32 Intel(R) Architecture Software Developer's Manuals, `http://www.intel.com/design/pentium4/manuals/253668.htm`

[4] q-tools homepage, `http://www.hpl.hp.com/`

research/linux/
q-tools/

[5] David Mosberger,
http://www.gelato.org/
community/gelato_
meeting.php?id=CU2004#
talk19

[6] qprof homepage,
http://www.hpl.hp.com/
research/linux/qprof/

[7] netperf homepage, http:
//www.netperf.org/

Carrier Grade Server Features in the Linux Kernel

Towards Linux-based Telecom Plarforms

Ibrahim Haddad
Ericsson Research
ibrahim.haddad@ericsson.com

Abstract

Traditionally, communications and data service networks were built on proprietary platforms that had to meet very specific availability, reliability, performance, and service response time requirements. Today, communication service providers are challenged to meet their needs cost-effectively for new architectures, new services, and increased bandwidth, with highly available, scalable, secure, and reliable systems that have predictable performance and that are easy to maintain and upgrade. This paper presents the technological trend of migrating from proprietary to open platforms based on software and hardware building blocks. It also focuses on the ongoing work by the Carrier Grade Linux working group at the Open Source Development Labs, examines the CGL architecture, the requirements from the latest specification release, and presents some of the needed kernel features that are not currently supported by Linux such as a Linux cluster communication mechanism, a low-level kernel mechanism for improved reliability and soft-realtime performance, support for multi-FIB, and support for additional security mechanisms.

1 Open platforms

The demand for rich media and enhanced communication services is rapidly leading to significant changes in the communication industry, such as the convergence of data and voice technologies. The transition to packet-based, converged, multi-service IP networks require a carrier grade infrastructure based on interoperable hardware and software building blocks, management middleware, and applications, implemented with standard interfaces. The communication industry is witnessing a technology trend moving away from proprietary systems toward open and standardized systems that are built using modular and flexible hardware and software (operating system and middleware) common off the shelf components. The trend is to proceed forward delivering next generation and multimedia communication services, using open standard carrier grade platforms. This trend is motivated by the expectations that open platforms are going to reduce the cost and risk of developing and delivering rich communications services. Also, they will enable faster time to market and ensure portability and interoperability between various components from different providers. One frequently asked question is: 'How can we meet tomorrow's requirements using existing infrastructures and technologies?'. Proprietary platforms are closed systems, expensive to develop, and often lack support of the current and upcoming standards. Using such closed platforms to meet tomorrow's requirements for new architectures and services is almost impossible. A uniform open software environment with the characteristics demanded by telecom

applications, combined with commercial off-the-shelf software and hardware components is a necessary part of these new architectures. The following key industry consortia are defining hardware and software high availability specifications that are directly related to telecom platforms:

1. The PCI Industrial Computer Manufacturers Group [1] (PICMG) defines standards for high availability (HA) hardware.
2. The Open Source Development Labs [2] (OSDL) Carrier Grade Linux [3] (CGL) working group was established in January 2002 with the goal of enhancing the Linux operating system, to achieve an Open Source platform that is highly available, secure, scalable and easily maintained, suitable for carrier grade systems.
3. The Service Availability Forum [4] (SA Forum) defines the interfaces of HA middleware and focusing on APIs for hardware platform management and for application failover in the application API. SA compliant middleware will provide services to an application that needs to be HA in a portable way.

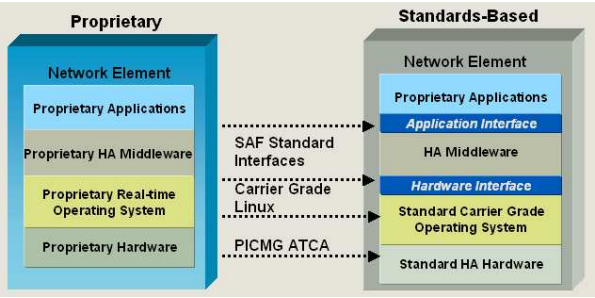


Figure 1: From Proprietary to Open Solutions

The operating system is a core component in such architectures. In the remaining of this paper, we will be focusing on CGL, its architecture and specifications.

2 The term Carrier Grade

In this paper, we refer to the term Carrier Grade on many occasions. Carrier grade is a term for public network telecommunications products that require a reliability percentage up to 5 or 6 nines of uptime.

- 5 nines refers to 99.999% of uptime per year (i.e., 5 minutes of downtime per year). This level of availability is usually associated with Carrier Grade servers.
- 6 nines refers to 99.9999% of uptime per year (i.e., 30 seconds of downtime per year). This level of availability is usually associated with Carrier Grade switches.

3 Linux versus proprietary operating systems

This section describes briefly the motivating reasons in favor of using Linux on Carrier Grade systems, versus continuing with proprietary operating systems. These motivations include:

- Cost: Linux is available free of charge in the form of a downloadable package from the Internet.
- Source code availability: With Linux, you gain full access to the source code allowing you to tailor the kernel to your needs.
- Open development process (Figure 2): The development process of the kernel is open to anyone to participate and contribute. The process is based on the concept of "release early, release often."
- Peer review and testing resources: With access to the source code, people using a

wide variety of platform, operating systems, and compiler combinations; can compile, link, and run the code on their systems to test for portability, compatibility and bugs.

- Vendor independent: With Linux, you no longer have to be locked into a specific vendor. Linux is supported on multiple platforms.
- High innovation rate: New features are usually implemented on Linux before they are available on commercial or proprietary systems.

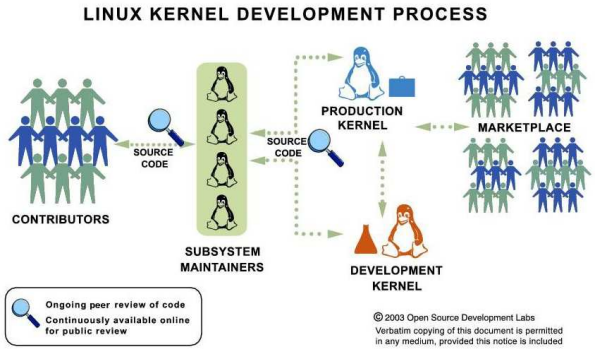


Figure 2: Open development process of the Linux kernel

Other contributing factors include Linux' support for a broad range of processors and peripherals, commercial support availability, high performance networking, and the proven record of being a stable, and reliable server platform.

4 Carrier Grade Linux

The Linux kernel is missing several features that are needed in a telecom environment. It is not adapted to meet telecom requirements in various areas such as reliability, security, and scalability. To help the advancement of Linux in the telecom space, OSDL established the CGL working group. The group specifies and helps implement an Open Source platform targeted for the communication industry that is highly available, secure, scalable and easily maintained. The CGL working group is composed of several members from network equipment providers, system integrators, platform providers, and Linux distributors. They all contribute to the requirement definition of Carrier Grade Linux, help Open Source projects to meet these requirements, and in some cases start new Open Source projects. Many of the CGL members companies have contributed pieces of technologies to Open Source in order to make the Linux Kernel a more viable option for telecom platforms. For instance, the Open Systems Lab [5] from Ericsson Research has contributed three key technologies: the Transparent IPC [6], the Asynchronous Event Mechanism [7], and the Distributed Security Infrastructure [8]. There are already Linux distributions, MontaVista [9] for instance, that are providing CGL distribution based on the CGL requirement definition. Many companies are also either deploying CGL, or at least evaluating and experimenting with it.

Consequently, CGL activities are giving much momentum for Linux in the telecom space allowing it to be a viable option to proprietary operating system. Member companies of CGL are releasing code to Open Source and are making some of their proprietary technologies open, which leads to going forward from closed platforms to open platforms that use CGL Linux.

5 Target CGL applications

The CGL Working Group has identified three main categories of application areas into which they expect the majority of applications implemented on CGL platforms to fall. These appli-

cation areas are gateways, signaling, and management servers.

- Gateways are bridges between two different technologies or administration domains. For example, a media gateway performs the critical function of converting voice messages from a native telecommunications time-division-multiplexed network, to an Internet protocol packet-switched network. A gateway processes a large number of small messages received and transmitted over a large number of physical interfaces. Gateways perform in a timely manner very close to hard real-time. They are implemented on dedicated platforms with replicated (rather than clustered) systems used for redundancy.

- Signaling servers handle call control, session control, and radio recourse control. A signaling server handles the routing and maintains the status of calls over the network. It takes the request of user agents who want to connect to other user agents and routes it to the appropriate signaling. Signaling servers require soft real time response capabilities less than 80 milliseconds, and may manage tens of thousands of simultaneous connections. A signaling server application is context switch and memory intensive due to requirements for quick switching and a capacity to manage large numbers of connections.

- Management servers handle traditional network management operations, as well as service and customer management. These servers provide services such as: a Home Location Register and Visitor Location Register (for wireless networks) or customer information (such as personal preferences including features the customer is authorized to use). Typically, management applications are data and communication intensive. Their response time requirements are less stringent by several orders of magnitude, compared to those of signaling and gateway applications.

6 Overview of the CGL working group

The CGL working group has the vision that next-generation and multimedia communication services can be delivered using Linux based open standards platforms for carrier grade infrastructure equipment. To achieve this vision, the working group has setup a strategy to define the requirements and architecture for the Carrier Grade Linux platform, develop a roadmap for the platform, and promote the development of a stable platform upon which commercial components and services can be deployed.

In the course of achieving this strategy, the OSDL CGL working group, is creating the requirement definitions, and identifying existing Open Source projects that support the roadmap to implement the required components and interfaces of the platform. When an Open Source project does not exist to support a certain requirement, OSDL CGL is launching (or support the launch of) new Open Source projects to implement missing components and interfaces of the platform.

The CGL working group consists of three distinct sub-groups that work together. These subgroups are: specification, proof-of-concept, and validation. Responsibilities of each subgroup are as follows:

1. Specifications: The specifications subgroup is responsible for defining a set of

requirements that lead to enhancements in the Linux kernel, that are useful for carrier grade implementations and applications. The group collects, categorizes, and prioritizes the requirements from participants to allow reasonable work to proceed on implementations. The group also interacts with other standard defining bodies, open source communities, developers and distributions to ensure that the requirements identify useful enhancements in such a way, that they can be adopted into the base Linux kernel.

2. Proof-of-Concept: This sub-group generates documents covering the design, features, and technology relevant to CGL. It drives the implementation and integration of core Carrier Grade enhancements to Linux as identified and prioritized by the requirement document. The group is also responsible for ensuring the integrated enhancements pass, the CGL validation test suite and for establishing and leading an open source umbrella project to coordinate implementation and integration activities for CGL enhancements.

3. Validation: This sub-group defines standard test environments for developing validation suites. It is responsible for coordinating the development of validation suites, to ensure that all of the CGL requirements are covered. This group is also responsible for the development of an Open Source project CGL validation suite.

7 CGL architecture

Figure 3 presents the scope of the CGL Working Group, which covers two areas:

- Carrier Grade Linux: Various requirements such as availability and scalability

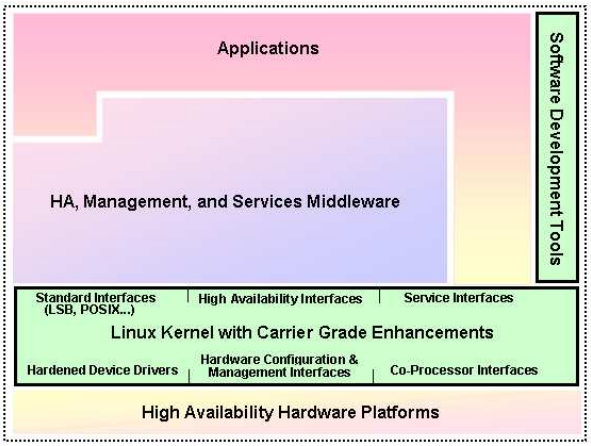


Figure 3: CGL architecture and scope

are related to the CGL enhancements to the operating system. Enhancements may also be made to hardware interfaces, interfaces to the user level or application code and interfaces to development and debugging tools. In some cases, to access the kernel services, user level library changes will be needed.

- Software Development Tools: These tools will include debuggers and analyzers.

  On October 9, 2003, OSDL announced the availability of the OSDL Carrier Grade Linux Requirements Definition, Version 2.0 (CGL 2.0). This latest requirement definition for next-generation carrier grade Linux offers major advances in security, high availability, and clustering.

8 CGL requirements

The requirement definition document of CGL version 2.0 introduced new and enhanced features to support Linux as a carrier grade platform. The CGL requirement definition divides the requirements in main categories described briefly below:

8.1 Clustering

These requirements support the use of multiple carrier server systems to provide higher levels of service availability through redundant resources and recovery capabilities, and to provide a horizontally scaled environment supporting increased throughput.

8.2 Security

The security requirements are aimed at maintaining a certain level of security while not endangering the goals of high availability, performance, and scalability. The requirements support the use of additional security mechanisms to protect the systems against attacks from both the Internet and intranets, and provide special mechanisms at kernel level to be used by telecom applications.

8.3 Standards

CGL specifies standards that are required for compliance for carrier grade server systems. Examples of these standards include:

- Linux Standard Base
- POSIX Timer Interface
- POSIX Signal Interface
- POSIX Message Queue Interface
- POSIX Semaphore Interface
- IPv6 RFCs compliance
- IPsecv6 RFCs compliance
- MIPv6 RFCs compliance
- SNMP support
- POSIX threads

8.4 Platform

OSDL CGL specifies requirements that support interactions with the hardware platforms making up carrier server systems. Platform capabilities are not tied to a particular vendor's implementation. Examples of the platform requirements include:

- Hot insert: supports hot-swap insertion of hardware components
- Hot remove: supports hot-swap removal of hardware components
- Remote boot support: supports remote booting functionality
- Boot cycle detection: supports detecting reboot cycles due to recurring failures. If the system experiences a problem that causes it to reboot repeatedly, the system will go offline. This is to prevent additional difficulties from occurring as a result of the repeated reboots
- Diskless systems: Provide support for diskless systems loading their kernel/application over the network
- Support remote booting across common LAN and WAN communication media

8.5 Availability

The availability requirements support heightened availability of carrier server systems, such as improving the robustness of software components or by supporting recovery from failure of hardware or software. Examples of these requirements include:

- RAID 1: support for RAID 1 offers mirroring to provide duplicate sets of all data on separate hard disks

- Watchdog timer interface: support for watchdog timers to perform certain specified operations when timeouts occur
- Support for Disk and volume management: to allow grouping of disks into volumes
- Ethernet link aggregation and link failover: support bonding of multiple NIC for bandwidth aggregation and provide automatic failover of IP addresses from one interface to another
- Support for application heartbeat monitor: monitor applications availability and functionality.

8.6 Serviceability

The serviceability requirements support servicing and managing hardware and software on carrier server systems. These are wide-ranging set requirements, put together, help support the availability of applications and the operating system. Examples of these requirements include:

- Support for producing and storing kernel dumps
- Support for dynamic debug to allow dynamically the insertion of software instrumentation into a running system in the kernel or applications
- Support for platform signal handler enabling infrastructures to allow interrupts generated by hardware errors to be logged using the event logging mechanism
- Support for remote access to event log information

8.7 Performance

OSDL CGL specifies the requirements that support performance levels necessary for the environments expected to be encountered by carrier server systems. Examples of these requirements include:

- Support for application (pre) loading.
- Support for soft real time performance through configuring the scheduler to provide soft real time support with latency of 10 ms.
- Support Kernel preemption.
- Raid 0 support: RAID Level 0 provides "disk striping" support to enhance performance for request-rate-intensive or transfer-rate-intensive environments

8.8 Scalability

These requirements support vertical and horizontal scaling of carrier server systems such as the addition of hardware resources to result in acceptable increases in capacity.

8.9 Tools

The tools requirements provide capabilities to facilitate diagnosis. Examples of these requirements include:

- Support the usage of a kernel debugger.
- Support for Kernel dump analysis.
- Support for debugging multi-threaded programs

9 CGL 3.0

The work on the next version of the OSDL CGL requirements, version 3.0, started in January 2004 with focus on advanced requirement areas such as manageability, serviceability, tools, security, standards, performance, hardware, clustering and availability. With the success of CGL's first two requirement documents, OSDL CGL working group anticipates that their third version will be quite beneficial to the Carrier Grade ecosystem. Official release of the CGL requirement document Version 3.0 is expected in October 2004.

10 CGL implementations

There are several enhancements to the Linux Kernel that are required by the communication industry, to help adopt Linux on their carrier grade platforms, and support telecom applications. These enhancements (Figure 4) fall into the following categories availability, security, serviceability, performance, scalability, reliability, standards, and clustering.

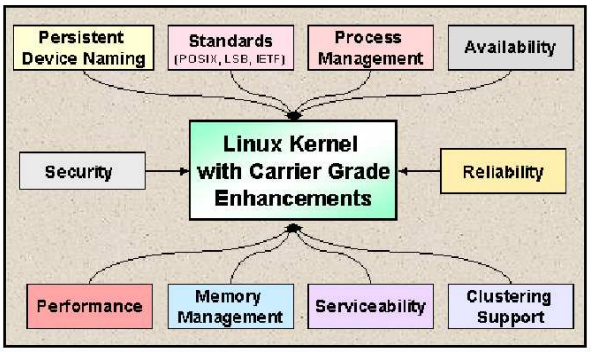


Figure 4: CGL enhancements areas

The implementations providing theses enhancements are Open Source projects and planned for integration with the Linux kernel when the implementations are mature, and ready for merging with the kernel code. In some cases, bringing some projects into maturity levels takes a considerable amount of time before being able to request its integration into the Linux kernel. Nevertheless, some of the enhancements are targeted for inclusion in kernel version 2.7. Other enhancements will follow in later kernel releases. Meanwhile, all enhancements, in the form of packages, kernel modules and patches, are available from their respective project web sites. The CGL 2.0 requirements are in-line with the Linux development community. The purpose of this project is to form a catalyst to capture common requirements from end-users for a CGL distribution. With a common set of requirements from the major Network Equipment Providers, developers can be much more productive and efficient within development projects. Many individuals within the CGL initiative are also active participants and contributors in the Open Source development community.

11 Examples of needed features in the Linux Kernel

In this section, we provide some examples of missing features and mechanisms from the Linux kernel that are necessary in a telecom environment.

11.1 Transparent Inter-Process and Inter-Processor Communication Protocol for Linux Clusters

Today's telecommunication environments are increasingly adopting clustered servers to gain benefits in performance, availability, and scalability. The resulting benefits of a cluster are greater or more cost-efficient than what a single server can provide. Furthermore, the telecommunications industry interest in clustering originates from the fact that clusters address carrier grade characteristics such as guaranteed service availability, reliability and

scaled performance, using cost-effective hardware and software. Without being absolute about these requirements, they can be divided in these three categories: short failure detection and failure recovery, guaranteed availability of service, and short response times. The most widely adopted clustering technique is use of multiple interconnected loosely coupled nodes to create a single highly available system.

One missing feature from the Linux kernel in this area is a reliable, efficient, and transparent inter-process and inter-processor communication protocol. Transparent Inter Process Communication (TIPC) [6] is a suitable Open Source implementation that fills this gap and provides an efficient cluster communication protocol. This leverages the particular conditions present within loosely coupled clusters. It runs on Linux and is provided as a portable source code package implementing a loadable kernel module.

TIPC is unique because there seems to be no other protocol providing a comparable combination of versatility and performance. It includes some original innovations such as the functional addressing, the topology subscription services, and the reactive connection concept. Other important TIPC features include full location transparency, support for lightweight connections, reliable multicast, signaling link protocol, topology subscription services and more.

TIPC should be regarded as a useful toolbox for anyone wanting to develop or use Carrier Grade or Highly Available Linux clusters. It provides the necessary infrastructure for cluster, network and software management functionality, as well as a good support for designing site-independent, scalable, distributed, high-availability and high-performance applications.

It is also worthwhile to mention that the ForCES (Forwarding and Control Element WG) [11] working group within IETF has agreed that their router internal protocol (the ForCES protocol) must be possible to carry over different types of transport protocols. There is consensus on that TCP is the protocol to be used when ForCES messages are transported over the Internet, while TIPC is the protocol to be used in closed environments (LANs), where special characteristics such as high performance and multicast support is desirable. Other protocols may also be added as options.

TIPC is a contribution from Ericsson [5] to the Open Source community. TIPC was announced on LKML on June 28, 2004; it is licensed under a dual GPL and BSD license.

11.2 IPv4, IPv6, MIPv6 forwarding tables fast access and compact memory with multiple FIB support

Routers are core elements of modern telecom networks. They propagate and direct billion of data packets from their source to their destination using air transport devices or through high-speed links. They must operate as fast as the medium in order to deliver the best quality of service and have a negligible effect on communications. To give some figures, it is common for routers to manage between 10.000 to 500.000 routes. In these situations, good performance is achievable by handling around 2000 routes/sec. The actual implementation of the IP stack in Linux works fine for home or small business routers. However, with the high expectation of telecom operators and the new capabilities of telecom hardware, it appears as barely possible to use Linux as an efficient forwarding and routing element of a high-end router for large network (core/border/access router) or a high-end server with routing capabilities.

One problem with the networking stack in Linux is the lack of support for multiple forward-ing information bases (multi-FIB) wit h overlapping interface's IP address, and the lack of appropriate interfaces for addressing FIB. Another problem with the curren t implementation is the limited scalability of the routing table.

The solution to these problems is to provide support for multi-FIB with overlapping IP address. As such, we can have on differe nt VLAN or different physical interfaces, independent network in the same Linux box. For example, we can have two HTTP servers serving two different networks with potentially the same IP address. One HTTP server will serve the network/FIB 10, and the othe r HTTP server will serves the network/FIB 20. The advantage gained is to have one Linux box serving two different customers usi ng the same IP address. ISPs adopt this approach by providing services for multiple customers sharing the same server (server pa rtitioning), instead of using a server per customer.

The way to achieve this is to have an ID (an identifier that identifies the customer or user of the service) to completely separ ate the routing table in memory. Two approaches exist: the first is to have a separate routing tables, each routing table is looked up by their ID and within tha t table the lookup is done one the prefix. The second approach is to have one table, and the lookup is done on the combined key = prefix + ID.

A different kind of problem arises when we are not able to predict access time, with the chaining in the hash table of the routi ng cache (and FIB). This problem is of particular inter-est in an environment that requires predictable performance.

Another aspect of the problem is that the route cache and the routing table are not kept synchronized most of the time (path MTU, just to name one). The route cache flush is executed regularly; therefore, any updates on the cache are lost. For example, if you have a routing cache flush, you have to rebuild every route that you are currently talking to, by going for every route in the hash/try table and rebuilding the information. First, you have to lookup in the routing cache, and if you have a miss, then you need to go in the hash/try table. This process is very slow and not predictable since the hash/try table is implemented wi th linked list and there is high potential for collisions when a large number of routes are present. This design is suitable fo r a home PC with a few routes, but it is not scalable for a large server.

To support the various routing requirements of server nodes operating in high performance and mission critical envrionments, Linux should support the following:

- Implementation of multi-FIB using tree (radix, patricia, etc.): It is very important to have predictable performance in insert/delete/lookup from 10.000 to 500.000 routes. In addition, it is favourable to have the same data structure for both IPv4 and IPv6.
- Socket and ioctl interfaces for addressing multi-FIB.
- Multi-FIB support for neighbors (arp).

Providing these implementations in Linux will affect a large part of net/core, net/ipv4 and net/ipv6; these subsystems (mostly network layer) will need to be re-written. Other areas will have minimal impact at the source code level, mostly at the transport layer (socket, TCP, UDP, RAW, NAT, IPIP, IGMP, etc.).

As for the availability of an Open Source project that can provide these functionalities,

there exists a project called "Linux Virtual Routing and Forwarding" [12]. This project aims to implement a flexible and scalable mechanism for providing multiple routing instances within the Linux kernel. The project has some potential in providing the needed functionalities, however no progress has been made since 2002 and the project seems to be inactive.

11.3 Run-time Authenticity Verification for Binaries

Linux has generally been considered immune to the spread of viruses, backdoors and Trojan programs on the Internet. However, with the increasing popularity of Linux as a desktop platform, the risk of seeing viruses or Trojans developed for this platform are rapidly growing. To alleviate this problem, the system should prevent on run time the execution of un-trusted software. One solution is to digitally sign the trusted binaries and have the system check the digital signature of binaries before running them. Therefore, untrusted (not signed) binaries are denied the execution. This can improve the security of the system by avoiding a wide range of malicious binaries like viruses, worms, Trojan programs and backdoors from running on the system.

DigSig [13] is a Linux kernel module that checks the signature of a binary before running it. It inserts digital signatures inside the ELF binary and verifies this signature before loading the binary. It is based on the Linux Security Module hooks (LSM has been integrated with the Linux kernel since 2.5.X and higher).

Typically, in this approach, vendors do not sign binaries; the control of the system remains with the local administrator. The responsible administrator is to sign all binaries they trust with their private key. Therefore, DigSig guarantees two things: (1) if you signed a binary, nobody else other than yourself can modify that binary without being detected. (2) Nobody can run a binary which is not signed or badly signed.

There has already been several initiatives in this domain, such as Tripwire [14], BSign [15], Cryptomark [16], but we believe the DigSig project is the first to be both easily accessible to all (available on SourceForge, under the GPL license) and to operate at kernel level on run time. The run time is very important for Carrier Grade Linux as this takes into account the high availability aspects of the system.

The DigSig approach has been using existing solutions like GnuPG [17] and BSign (a Debian package) rather than reinventing the wheel. However, in order to reduce the overhead in the kernel, the DigSig project only took the minimum code necessary from GnuPG. This helped much to reduce the amount of code imported to the kernel in source code of the original (only 1/10 of the original GnuPG 1.2.2 source code has been imported to the kernel module).

DigSig is a contribution from Ericsson [5] to the Open Source community. It was released under the GPL license and it is available from [8].

DigSig has been announced on LKML [18] but it not yet integrated in the Linux Kernel.

11.4 Efficient Low-Level Asynchronous Event Mechanism

Carrier grade systems must provide a 5-nines availability, a maximum of five minutes per year of downtime, which includes hardware, operating system, software upgrade and maintenance. Operating systems for such systems must ensure that they can deliver a high response rate with minimum downtime. In addition, carrier-grade systems must take into account such characteristics such as scalabil-

ity, high availability and performance. In carrier grade systems, thousands of requests must be handled concurrently without affecting the overall system's performance, even under extremely high loads. Subscribers can expect some latency time when issuing a request, but they are not willing to accept an unbounded response time. Such transactions are not handled instantaneously for many reasons, and it can take some milliseconds or seconds to reply. Waiting for an answer reduces applications abilities to handle other transactions.

Many different solutions have been envisaged to improve Linux's capabilities in this area using different types of software organization, such as multithreaded architectures, implementing efficient POSIX interfaces, or improving the scalability of existing kernel routines.

One possible solution that is adequate for carrier grade servers is the Asynchronous Event Mechanism (AEM), which provides asynchronous execution of processes in the Linux kernel. AEM implements a native support for asynchronous events in the Linux kernel and aims to bring carrier-grade characteristics to Linux in areas of scalability and soft real-time responsiveness. In addition, AEM offers event-based development framework, scalability, flexibility, and extensibility.

Ericsson [5] released AEM to Open Source in February 2003 under the GPL license. AEM was announced on the Linux Kernel Mailing List (LKML) [20], and received feedback that resulted in some changes to the design and implementation. AEM is not yet integrated with the Linux kernel.

12 Conclusion

There are many challenges accompanying the migration from proprietary to open platforms. The main challenge remains to be the availability of the various kernel features and mechanisms needed for telecom platforms and integrating these features in the Linux kernel.

References

[1] PCI Industrial Computer Manufacturers Group, `http://www.picmg.org`

[2] Open Source Development Labs, `http://www.osdl.org`

[3] Carrier Grade Linux, `http://osdl.org/lab_activities`

[4] Service Availability Forum, `http://www.saforum.org`

[5] Open System Lab, `http://www.linux.ericsson.ca`

[6] Transparent IPC, `http://tipc.sf.net`

[7] Asynchronous Event Mechanism, `http://aem.sf.net`

[8] Distributed Security Infrastructure, `http://disec.sf.net`

[9] MontaVista Carrier Grade Edition, `http://www.mvista.com/cge`

[10] Make Clustering Easy with TIPC, LinuxWorld Magazine, April 2004

[11] IETF ForCES working group, `http://www.sstanamera.com/~forces`

[12] Linux Virtual Routing and Forwarding project, `http://linux-vrf.sf.net`

[13] Stop Malicious Code Execution at Kernel Level, LinuxWorld Magazine, January 2004

[14] Tripwire,
http://www.tripwire.com

[15] Bsign,
http://packages.debian.org/bsign

[16] Cryptomark,
http://immunix.org/cryptomark.html

[17] GnuPG,
http://www.gnupg.org

[18] DigSig announcement on LKML,
http://lwn.net/Articles/51007

[19] An Event Mechanism for Linux, Linux Journal, July 2003

[20] AEM announcement on LKML,
http://lwn.net/Articles/45633

Acknowledgments

Thank you to Ludovic Beliveau, Mathieu Giguere, Magnus Karlson, Jon Maloy, Mats Naslund, Makan Pourzandi, and Frederic Rossi, for their valuable contributions and reviews.

Demands, Solutions, and Improvements for Linux Filesystem Security

Michael Austin Halcrow
International Business Machines, Inc.
mike@halcrow.us

Abstract

Securing file resources under Linux is a team effort. No one library, application, or kernel feature can stand alone in providing robust security. Current Linux access control mechanisms work in concert to provide a certain level of security, but they depend upon the integrity of the machine itself to protect that data. Once the data leaves that machine, or if the machine itself is physically compromised, those access control mechanisms can no longer protect the data in the filesystem. At that point, data privacy must be enforced via encryption.

As Linux makes inroads in the desktop market, the need for transparent and effective data encryption increases. To be practically deployable, the encryption/decryption process must be secure, unobtrusive, consistent, flexible, reliable, and efficient. Most encryption mechanisms that run under Linux today fail in one or more of these categories. In this paper, we discuss solutions to many of these issues via the integration of encryption into the Linux filesystem. This will provide access control enforcement on data that is not necessarily under the control of the operating environment. We also explore how stackable filesystems, Extended Attributes, PAM, GnuPG web-of-trust, supporting libraries, and applications (such as GNOME/KDE) can all be orchestrated to provide robust encryption-based access control over filesystem content.

1 Development Efforts

This paper is motivated by an effort on the part of the IBM Linux Technology Center to enhance Linux filesystem security through better integration of encryption technology. The author of this paper is working together with the external community and several members of the LTC in the design and development of a transparent cryptographic filesystem layer in the Linux kernel. The "we" in this paper refers to immediate members of the author's development team who are working together on this project, although many others outside that development team have thus far had a significant part in this development effort.

2 The Filesystem Security

2.1 Threat Model

Computer users tend to be overly concerned about protecting their credit card numbers from being sniffed as they are transmitted over the Internet. At the same time, many do not think twice when sending equally sensitive information in the clear via an email message. A thief who steals a removable device, laptop, or server can also read the confidential files on those devices if they are left unprotected. Nevertheless, far too many users neglect to take the necessary steps to protect their files from such an event. Your liability limit for unauthorized

charges to your credit card is $50 (and most credit card companies waive that liability for victims of fraud); on the other hand, confidentiality cannot be restored once lost.

Today, we see countless examples of neglect to use encryption to protect the integrity and the confidentiality of sensitive data. Those who are trusted with sensitive information routinely send that information as unencrypted email attachments. They also store that information in clear text on disks, USB keychain drives, backup tapes, and other removable media. GnuPG[7] and OpenSSL[8] provide all the encryption tools necessary to protect this information, but these tools are not used nearly as often as they ought to be.

If required to go through tedious encryption or decryption steps every time they need to work with a file or share it, people will select insecure passwords, transmit passwords in an insecure manner, fail to consider or use public key encryption options, or simply stop encrypting their files altogether. If security is overly obstructive, people will remove it, work around it, or misuse it (thus rendering it less effective). As Linux gains adoption in the desktop market, we need integrated file integrity and confidentiality that is seamless, transparent, easy to use, and effective.

2.2 Integration of File Encryption into the Filesystem

Several solutions exist that solve separate pieces of the problem. In one example highlighting transparency, employees within an organization that uses IBM™ Lotus Notes™ [9] for its email will not even notice the complex PKI or the encryption process that is integrated into the product. Encryption and decryption of sensitive email messages is seamless to the end user; it involves checking an "Encrypt" box, specifying a recipient, and sending the message. This effectively addresses a significant file in-transit confidentiality problem. If the local replicated mailbox database is also encrypted, then it also addresses confidentiality on the local storage device, but the protection is lost once the data leaves the domain of Notes (for example, if an attached file is saved to disk). The process must be seamlessly integrated into *all* relevant aspects of the user's operating environment.

In Section 4, we discuss filesystem security in general under Linux, with an emphasis on confidentiality and integrity enforcement via cryptographic technologies. In Section 6, we propose a mechanism to integrate encryption of files at the filesystem level, including integration of GnuPG[7] web-of-trust, PAM[10], a stackable filesystem model[2], Extended Attributes[6], and libraries and applications, in order to make the entire process as transparent as possible to the end user.

3 A Team Effort

Filesystem security encompasses more than just the filesystem itself. It is a team effort, involving the kernel, the shells, the login processes, the filesystems, the applications, the administrators, and the users. When we speak of "filesystem security," we refer to the security of the files in a filesystem, no matter what ends up providing that security.

For any filesystem security problem that exists, there are usually several different ways of solving it. Solutions that involve modifications in the kernel tend to introduce less overhead. This is due to the fact that context switches and copying of data between kernel and user memory is reduced. However, changes in the kernel may reduce the efficiency of the kernel's VFS while making it both harder to maintain and more bug-prone. As notable exceptions,

Erez Zadok's stackable filesystem framework, FiST[3], and Loop-aes, require no change to the current Linux kernel VFS. Solutions that exist entirely in userspace do not complicate the kernel, but they tend to have more overhead and may be limited in the functionality they are able to provide, as they are limited by the interface to the kernel from userspace. Since they are in userspace, they are also more prone to attack.

4 Aspects of Filesystem Security

Computer security can be decomposed into several areas:

- Identifying who you are and having the machine recognize that identification (*authentication*).
- Determining whether or not you should be granted access to a resource such as a sensitive file (*authorization*). This is often based on the permissions associated with the resource by its owner or an administrator (*access control*).
- Transforming your data into an encrypted format in order to make it prohibitively costly for unauthorized users to decrypt and view (*confidentiality*).
- Performing checksums, keyed hashes, and/or signing of your data to make unauthorized modifications of your data detectable (*integrity*).

4.1 Filesystem Integrity

When people consider filesystem security, they traditionally think about access control (file permissions) and confidentiality (encryption). File integrity, however, can be just as important as confidentiality, if not more so. If a script that performs an administrative task is altered in an unauthorized fashion, the script may perform actions that violate the system's security policies. For example, many rootkits modify system startup and shutdown scripts to facilitate the attacker's attempts to record the user's keystrokes, sniff network traffic, or otherwise infiltrate the system.

More often than not, the value of the data stored in files is greater than that of the machine that hosts the files. For example, if an attacker manages to insert false data into a financial report, the alteration to the report may go unnoticed until substantial damage has been done; jobs could be at stake and in more extreme cases even criminal charges against the user could result . If trojan code sneaks into the source repository for a major project, the public release of that project may contain a backdoor.[1]

Many security professionals foresee a nightmare scenario wherein a widely propagated Internet worm quietly alters the contents of word processing and spreadsheet documents. Without any sort of integrity mechanism in place in the vast majority of the desktop machines in the world, nobody would know if any data that traversed vulnerable machines could be trusted. This threat could be very effectively addressed with a combination of a kernel-level mandatory access control (MAC)[11] protection profile and a filesystem that provides integrity and auditing capabilities. Such a combination would be resistant to damage done by a root compromise, especially if aided by a Trusted Platform Module (TPM)[13] using attestation.

[1] A high-profile example of an attempt to do this occurred with the Linux kernel last year. Fortunately, the source code management process used by the kernel developers allowed them to catch the attempted insertion of the trojan code before it made it into the actual kernel.

One can approach filesystem integrity from two angles. The first is to have strong authentication and authorization mechanisms in place that employ sufficiently flexible policy languages. The second is to have an auditing mechanism, to detect unauthorized attempts at modifying the contents of a filesystem.

4.1.1 Authentication and Authorization

The filesystem must contain support for the kernel's security structure, which requires stateful security attributes on each file. Most GNU/Linux applications today use PAM[10] (see Section 4.1.2 below) for authentication and process credentials to represent their authorization; policy language is limited to what can be expressed using the file owner and group, along with the owner/group/world read/write/execute attributes of the file. The administrator and the current owner have the authority to set the owner of the file or the read/write/execute policies for that file. In many filesystems, files may also contain additional security flags, such as an immutable or append-only flag.

Posix Access Control Lists (ACL's)[6] provide for more stringent delegations of access authority on a per-file basis. In an ACL, individual read/write/execute permissions can be assigned to the owner, the owning group, individual users, or groups. Masks can also be applied that indicate the maximum effective permissions for a class.

For those who require even more flexible access control, SE Linux[15] uses a powerful policy language that can express a wide variety of access control policies for files and filesystem operations. In fact, Linux Security Module (LSM)[14] hooks (see Section 4.1.3 below) exist for most of the security-relevant filesystem operations, which makes it easier to implement custom filesystem-agnostic security models. Authentication and authorization are pretty well covered with a combination of existing filesystem, kernel, and user-space solutions that are part of most GNU/Linux distributions. Many distributions could, however, do a better job of aiding both the administrator and the user in understanding and using all the tools that they have available to them.

Policies that safeguard sensitive data should include timeouts, whereby the user must periodically re-authenticate in order to continue to access the data. In the event that the authorized users neglect to lock down the machine before leaving work for the day, timeouts help to keep the custodial staff from accessing the data when they come in at night to clean the office. As usual, this must be implemented in such a way as to be unobtrusive to the user. If a user finds a security mechanism overly imposing or inconvenient, he will usually disable or circumvent it.

4.1.2 PAM

Pluggable Authentication Modules (PAM)[10] implement authentication-related security policies. PAM offers discretionary access control (DAC)[12]; applications must defer to PAM in order to authenticate a user. If the authenticating PAM function that is called returns an affirmative answer, then the application can use that response to authorize the action, and vice versa. The exact mechanism that the PAM function uses to evaluate the authentication is dependent on the module called.[2]

In the case of filesystem security and encryption, PAM can be employed to obtain and forward keys to a filesystem encryption layer in kernel space. This would allow seamless inte-

[2] This is parameterizable in the configuration files found under */etc/pam.d/*

gration with any key retrieval mechanism that can be coded as a Pluggable Authentication Module.

4.1.3 LSM

Linux Security Modules (LSM) can provide customized security models. One possible use of LSM is to allow decryption of certain files only when a physical device is connected to the machine. This could be, for example, a USB keychain device, a Smartcard, or an RFID device. Some devices of these classes can also be used to house the encryption keys (retrievable via PAM, as previously discussed).

4.1.4 Auditing

The second angle to filesystem integrity is auditing. Auditing should only fill in where authentication and authorization mechanisms fall short. In a utopian world, where security systems are perfect and trusted people always act trustworthily, auditing does not have much of a use. In reality, code that implements security has defects and vulnerabilities. Passwords can be compromised, and authorized people can act in an untrustworthy manner. Auditing can involve keeping a log of all changes made to the attributes of the file or to the file data itself. It can also involve taking snapshots of the attributes and/or contents of the file and comparing the current state of the file with what was recorded in a prior snapshot.

Intrusion detection systems (IDS), such as Tripwire[16], AIDE[17], or Samhain[18], perform auditing functions. As an example, Tripwire periodically scans the contents of the filesystem, checking file attributes, such as the size, the modification time, and the cryptographic hash of each file. If any attributes for the files being checked are found to be altered, Tripwire will report it. This approach can work fairly well in cases where the files are not expected to change very often, as is the case with most system scripts, shared libraries, executables, or configuration files. However, care must be taken to assure that the attacker cannot also modify Tripwire's database when he modifies a system file; the integrity of the IDS system itself must also be assured.

In cases where a file changes often, such as a database file or a spreadsheet file in an active project, we see a need for a more dynamic auditing solution - one which is perhaps more closely integrated with the filesystem itself. In many cases, the simple fact that the file has changed does not imply a security violation. We must also know who made the change. More robust security requirements also demand that we know what parts of the file were changed and when the changes were made. One could even imagine scenarios where the context of the change must also be taken into consideration (i.e., who was logged in, which processes were running, or what network activity was taking place at the time the change was made).

File integrity, particularly in the area of auditing, is perhaps the security aspect of Linux filesystems that could use the most improvement. Most efforts in secure filesystem development have focused on confidentiality more so than integrity, and integrity has been regulated to the domain of userland utilities that must periodically scan the entire filesystem. Sometimes, just knowing that a file has been changed is insufficient. Administrators would like to know exactly how the attacker made the changes and under what circumstances they were made.

Cryptographic hashes are often used. These can detect unauthorized circumvention of the filesystem itself, as long as the attacker forgets

(or is unable) to update the hashes when making unauthorized changes to the files. Some auditing solutions, such as the Linux Auditing System (LAuS)[3] that is part of SuSE Linux Enterprise Server, can track system calls that affect the filesystem. Another recent addition to the 2.6 Linux kernel is the Light-weight Auditing Framework written by Rik Faith[28]. These are implemented independently of the filesystem itself, and the level of detail in the records is largely limited to the system call parameters and return codes. It is advisable that you keep your log files on a separate machine than the one being audited, since the attacker could modify the audit logs themselves once he has compromised the machine's security.

4.1.5 Improvements on Integrity

Extended Attributes provide for a convenient way to attach metadata relating to a file to the file itself. On the premise that possession of a secret equates to authentication, every time an authenticated subject makes an authorized write to a file, a hash over the concatenation of that secret to the file contents (keyed hashing; HMAC is one popular standard) can be written as an Extended Attribute on that file. Since this action would be performed on the filesystem level, the user would not have to conscientiously re-run userspace tools to perform such an operation every time he wants to generate an integrity verifier on the file.

This is an expensive operation to perform over large files, and so it would be a good idea to define extent sizes over which keyed hashes are formed, with the Extended Attributes including extent descriptors along with the keyed hashes. That way, a small change in the middle of a large file would only require the keyed hash to be re-generated over the extent in which the change occurs. A keyed hash over the sequential set of the extent hashes would also keep an attacker from swapping around extents undetected.

[3]Note that LAuS is being covered in more detail in the 2004 Ottawa Linux Symposium by Doc Shankar, Emily Ratliff, and Olaf Kirch as part of their presentation regarding CAPP/EAL3+ Certification.

4.2 File Confidentiality

Confidentiality means that only authorized users can read the contents of a file. Sometimes the names of the files themselves or a directory structure can be sensitive. In other cases, the sizes of the files or the modification times can betray more information than one might want to be known. Even the security policies protecting the files can reveal sensitive information. For example, "Only employees of Novell and SuSE can read this file" would imply that Novell and SuSE are collaborating on something, and neither of them may want this fact to be public knowledge as of yet. Many interesting protocols have been developed that can address these sorts of issues; some of them are easier to implement than others.

When approaching the question of confidentiality, we assume that the block device that contains the file is vulnerable to physical compromise. For example, a laptop that contains sensitive material might be lost, or a database server might be stolen in a burglary. In either event, the data on the hard drive must not be readable by an unauthorized individual. If any individual must be authenticated before he is able to access to the data, then the data is protected against unauthorized access.

Surprisingly, many users surrender their own data's confidentiality (and more often than not they do so unwittingly). It has been my personal observation that most people do not fully understand the lack of confidentiality afforded their data when they send it over the Internet. To compound this problem, comprehend-

ing and even using most encryption tools takes considerable time and effort on the part of most users. If sensitive files could be *encrypted by default*, only to be decrypted by those authorized at the time of access, then the user would not have to expend so much effort toward protecting the data's confidentiality.

By putting the encryption at the filesystem layer, this model becomes possible without any modifications to the applications or libraries. A policy at that layer can dictate that certain processes, such as the mail client, are to receive the encrypted version any files that are read from disk.

4.2.1 Encryption

File confidentiality is most commonly accomplished through encryption. For performance reasons, secure filesystems use symmetric key cryptography, like AES or Triple-DES, although an asymmetric public/private keypair may be used to encrypt the symmetric key in some key management schemes. This hybrid approach is in common use through SSL and PGP encryption protocols.

One of our proposals to extend Cryptfs is to mirror the techniques used in GnuPG encryption. If the symmetric key that protects the contents of a file is encrypted with the public key of the intended recipient of the file and stored as an Extended Attribute of the file, then that file can be transmitted in multiple ways (e.g., physical device such as removable storage); as long as the Extended Attributes of the file are preserved across filesystem transfers, then the recipient with the corresponding private key has all the information that his Cryptfs layer needs to transparently decrypt the contents of the file.

4.2.2 Key Management

Key management will make or break a cryptographic filesystem.[5] If the key can be easily compromised, then even the strongest cipher will provide weak protection. If your key is accessible in an unencrypted file or in an unprotected region of memory, or if it is ever transmitted over the network in the clear, a rogue user can capture that key and use it later. Most passwords have poor entropy, which means that an attacker can have pretty good success with a brute force attack against the password. Thus the weakest link in the chain for password-based encryption is usually the password itself. The Cryptographic Filesystem (CFS)[22] mandates that the user choose a password with a length of at least 16 characters.[4]

Ideally, the key would be kept in password-encrypted form on a removable device (like a USB keychain drive) that is stored separately from the files that the key is used to encrypt. That way, an attacker would have to both compromise the password and gain physical access to the removable device before he could decrypt your files.

Filesystem encryption is one of the most exciting applications for the Trusted Computing Platform. Given that the attacker has physical access to a machine with a Trusted Platform Module, it is significantly more difficult to compromise the key. By using secret sharing (otherwise known as *key splitting*)[4], the actual key used to decrypt a file on the filesystem can be contained as both the user's key and the machine's key (as contained in the TPM). In order to decrypt the files, an attacker must not

[4] The subject of secure password selection, although an important one, is beyond the scope of this article. Recommended reading on this subject is at `http://www.alw.nih.gov/Security/Docs/passwd.html`.

only compromise the user key, but he must also have access to the machine on which the TPM chip is installed. This "binds" the encrypted files to the machine. This is especially useful for protecting files on removable backup media.

4.2.3 Cryptanalysis

All block ciphers and most stream ciphers are, to various degrees, vulnerable to successful cryptanalysis. If a cipher is used improperly, then it may become even easier to discover the plaintext and/or the key. For example, with certain ciphers operating in certain modes, an attacker could discover information that aids in cryptanalysis by getting the filesystem to re-encrypt an already encrypted block of data. Other times, a cryptanalyst can deduce information about the type of data in the encrypted file when that data has predictable segments of data, like a common header or footer (thus allowing for a known-plaintext attack).

4.2.4 Cipher Modes

A block encryption mode that is resistant to cryptanalysis can involve dependencies among chains of bytes or blocks of data. Cipher-block-chaining (CBC) mode, for example, provides adequate encryption in many circumstances. In CBC mode, a change to one block of data will require that all subsequent blocks of data be re-encrypted. One can see how this would impact performance for large files, as a modification to data near the beginning of the file would require that all subsequent blocks be read, decrypted, re-encrypted, and written out again.

This particular inefficiency can be effectively addressed by defining chaining extents. By limiting regions of the file that encompass chained blocks, it is feasible to decrypt and re-encrypt the smaller segments. For example, if the block size for a cipher is 64 bits (8 bytes) and the block size, which is (we assume) the minimum unit of data that the block device driver can transfer at a time (512 bytes) then one could limit the number of blocks in any extent to 64 blocks. Depending on the plaintext (and other factors), this may be too few to effectively counter cryptanalysis, and so the extent size could be set to a small multiple of the page size without severely impacting overall performance. The optimal extent size largely depends on the access patterns and data patterns for the file in question; we plan on benchmarking against varying extent lengths under varying access patterns.

4.2.5 Key Escrow

The proverbial question, "What if the sysadmin gets hit by a bus?" is one that no organization should ever stop asking. In fact, sometimes no one person should alone have independent access to the sensitive data; multiple passwords may be required before the data is decrypted. Shareholders should demand that no single person in the company have full access to certain valuable data, in order to mitigate the damage to the company that could be done by a single corrupt administrator or executive. Methods for secret sharing can be employed to assure that multiple keys be required for file access, and (m,n)-threshold schemes [4] can ensure that the data is retrievable, even if a certain number of the keys are lost. Secret sharing would be easily implementable as part of any of the existing cryptographic filesystems.

4.3 File Resilience

The loss of a file can be just as devastating as the compromise of a file. There are many

well-established solutions to performing backups of your filesystem, but some cryptographic filesystems preclude the ability to efficiently and/or securely use them. Backup tapes tend to be easier to steal than secure computer systems are, and if unencrypted versions of secure files exist on the tapes, that constitutes an often-overlooked vulnerability.

The Linux 2.6 kernel cryptoloop device[5] filesystem is an all-or-nothing approach. Most backup utilities must be given free reign on the unencrypted directory listings in order to perform incremental backups. Most other encrypted filesystems keep sets of encrypted files in directories in the underlying filesystem, which makes incremental backups possible without giving the backup tools access to the unencrypted content of the files.

The backup utilities must, however, maintain backups of the metadata in the directories containing the encrypted files in addition to the files themselves. On the other hand, when the filesystem takes the approach of storing the cryptographic metadata as Extended Attributes for each file, then backup utilities need only worry about copying just the file in question to the backup medium (preserving the Extended Attributes, of course).

4.4 Advantages of FS-Level, EA-Guided Encryption

Most encrypted filesystem solutions either operate on the entire block device or operate on entire directories. There are several advantages to implementing filesystem encryption at the filesystem level and storing encryption metadata in the Extended Attributes of each file:

- Granularity: Keys can be mapped to individual files, rather than entire block devices or entire directories.
- Backup Utilities: Incremental backup tools can correctly operate without having to have access to the decrypted content of the files it is backing up.
- Performance: In most cases, only certain files need to be encrypted. System libraries and executables, in general, do not need to be encrypted. By limiting the actual encryption and decryption to only those files that really need it, system resources will not be taxed as much.
- Transparent Operation: Individual encrypted files can be easily transfered off of the block device without any extra transformation, and others with authorization will be able to decrypt those files. The userspace applications and libraries do not need to be modified and recompiled to support this transparency.

Since all the information necessary to decrypt a file is contained in the Extended Attributes of the file, it is possible for a user on a machine that is not running Cryptfs to use userland utilities to access the contents of the file. This also applies to other security-related operations, like verifying keyed hashes. This addresses compatibility issues with machines that are not running the encrypted filesystem layer.

5 Survey of Linux Encrypted Filesystems

5.1 Encrypted Loopback Filesystems

5.1.1 Loop-aes

The most well-known method of encrypting a filesystem is to use a loopback en-

[5] Note that this is deprecated and is in the process of being replaced with the Device Mapper crypto target.

crypted filesystem.[6] Loop-aes[20] is part of the 2.6 Linux kernel (`CONFIG_BLK_DEV_CRYPTOLOOP`). It performs encryption at the block device level. With Loop-aes, the administrator can choose whatever cipher he wishes to use with the filesystem. The *mount* package on most popular GNU/Linux distributions contains the *losetup* utility, which can be used to set up the encrypted loopback mount (you can choose whatever cipher that the kernel supports; we use blowfish in this example):

```
root# modprobe cryptoloop
root# modprobe blowfish
root# dd if=/dev/urandom of=encrypted.img \
      bs=4k count=1000
root# losetup -e blowfish /dev/loop0 \
      encrypted.img
root# mkfs.ext3 /dev/loop0
root# mkdir /mnt/unencrypted-view
root# mount /dev/loop0 /mnt/unencrypted-view
```

The loopback encrypted filesystem falls short in the fact that it is an all-or-nothing solution. It is impossible for most standard backup utilities to perform incremental backups on sets of encrypted files without being given access to the unencrypted files. In addition, remote users will need to use IPSec or some other network encryption layer when accessing the files, which must be exported from the unencrypted mount point on the server. Loop-aes is, however, the best performing encrypted filesystem that is freely available and integrated with most GNU/Linux distributions. It is an adequate solution for many who require little more than basic encryption of their entire filesystems.

5.1.2 BestCrypt

BestCrypt[23] is a non-free product that uses a loopback approach, similar to Loop-aes.

[6] Note that Loop-aes is being deprecated, in favor of Device Mapping (DM) Crypt, which also does encryption at the block device layer.

5.1.3 PPDD

PPDD[21] is a block device driver that encrypts and decrypts data as it goes to and comes from another block device. It works very much like Loop-aes; in fact, in the 2.4 kernel, it uses the loopback device, as Loop-aes does. PPDD has not been ported to the 2.6 kernel. Loop-aes takes the same approach, and Loop-aes ships with the 2.6 kernel itself.

5.2 CFS

The Cryptographic Filesystem (CFS)[22] by Matt Blaze is a well established transparent encrypted filesystem, originally written for BSD platforms. CFS is implemented entirely in userspace and operates similarly to NFS. A userspace daemon, cfsd, acts as a pseudo-NFS server, and the kernel makes RPC calls to the daemon. The CFS daemon performs transparent encryption and decryption when writing and reading data. Just as NFS can export a directory from any exportable filesystem, CFS can do the same, while managing the encryption on top of that filesystem.

In the background, CFS stores the metadata necessary to encrypt and decrypt files with the files being encrypted or decrypted on the filesystem. If you were to look at those directories directly, you would see a set of files with encrypted values for filenames, and there would be a handful of metadata files mixed in. When accessed through CFS, those metadata files are hidden, and the files are transparently encrypted and decrypted for the user applications (with the proper credentials) to freely work with the data.

While CFS is capable of acting as a remote NFS server, this is not recommended for many reasons, some of which include performance and security issues with plaintext passwords and unencrypted data being transmitted over

the network. You would be better off, from a security perspective (and perhaps also performance, depending on the number of clients), to use a regular NFS server to handle remote mounts of the encrypted directories, with local CFS mounts off of the NFS mounts.

Perhaps the most attractive attribute of CFS is the fact that it does not require any modifications to the standard Linux kernel. The source code for CFS is freely obtainable. It is packaged in the Debian repositories and is also available in RPM form. Using apt, CFS is perhaps the easiest encrypted filesystem for a user to set up and start using:

```
root# apt-get install cfs
user# cmkdir encrypted-data
user# cattach encrypted-data unencrypted-view
```

The user will be prompted for his password at the requisite stages. At this point, anything the user writes to or reads from */crypt/unencrypted-view* will be transparently encrypted to and decrypted from files in *encrypted-data*. Note that any user on the system can make a new encrypted directory and attach it. It is not necessary to initialize and mount an entire block device, as is the case with Loop-aes.

5.3 TCFS

TCFS[24] is a variation on CFS that includes secure integrated remote access and file integrity features. TCFS assumes the client's workstation is trusted, and the server cannot necessarily be trusted. Everything sent to and from the server is encrypted. Encryption and decryption take place on the client side.

Note that this behavior can be mimicked with a CFS mount on top of an NFS mount. However, because TCFS works within the kernel (thus requiring a patch) and does not necessitate two levels of mounting, it is faster than an NFS+CFS combination.

TCFS is no longer an actively maintained project. The last release was made three years ago for the 2.0 kernel.

5.4 Cryptfs

As a proof-of-concept for the FiST stackable filesystem framework, Erez Zadok, et. al. developed Cryptfs[1]. Under Cryptfs, symmetric keys are associated with groups of files within a single directory. The key is generated with a password that is entered at the time that the filesystem is mounted. The Cryptfs mount point provides an unencrypted view of the directory that contains the encrypted files.

The authors of this paper are currently working on extending Cryptfs to provide seamless integration into the user's desktop environment (see Section 6).

5.5 Userspace Encrypted Filesystems

EncFS[25] utilizes the Filesystem in Userspace (FUSE) library and kernel module to implement an encrypted filesystem in userspace. Like CFS, EncFS encrypts on a per-file basis.

CryptoFS[26] is similar to EncFS, except it uses the Linux Userland Filesystem (LUFS) library instead of FUSE.

SSHFS[27], like CryptoFS, uses the LUFS kernel module and userspace daemon. It limits itself to encrypting the files via SFTP as they are transfered over a network; the files stored on disk are unencrypted. From the user perspective, all file accesses take place as though they were being performed on any regular filesystem (opens, read, writes, etc.). SSHFS transfers the files back and forth via SFTP with the file server as these operations occur.

5.6 Reiser4

ReiserFS version 4 (Reiser4)[29], while still in the development stage, features pluggable security modules. There are currently proposed modules for Reiser4 that will perform encryption and auditing.

5.7 Network Filesystem Security

Much research has taken place in the domain of networking filesystem security. CIFS, NFSv4, and other networking filesystems face special challenges in relation to user identification, access control, and data secrecy. The NFSv4 protocol definition in RFC 3010 contains descriptions of security mechanisms in section 3[30].

6 Proposed Extensions to Cryptfs

Our proposal is to place file encryption metadata into the Extended Attributes (EA's) of the file itself. Extended Attributes are a generic interface for attaching metadata to files. The Cryptfs layer will be extended to extract that information and to use the information to direct the encrypting and decrypting of the contents of the file. In the event that the filesystem does not support Extended Attributes, another filesystem layer can provide that functionality. The stackable framework effectively allows Cryptfs to operate on top of *any* filesystem.

The encryption process is very similar to that of GnuPG and other public key cryptography programs that use a hybrid approach to encrypting data. By integrating the process into the filesystem, we can achieve a greater degree of transparency, without requiring any changes to userspace applications or libraries.

Under our proposed design, when a new file is created as an encrypted file, the Cryptfs layer generates a new symmetric key K_s for the encryption of the data that will be written. File creation policy enacted by Cryptfs can be dictated by directory attributes or globally defined behavior. The owner of the file is automatically authorized to access the file, and so the symmetric key is encrypted with the public key of the owner of the file K_u, which was passed into the Cryptfs layer at the time that the user logged in by a Pluggable Authentication Module linked against libcryptfs. The encrypted symmetric key is then added to the Extended Attribute set of the file:

$$\{K_s\}K_u$$

Suppose that the user at this point wants to grant Alice access to the file. Alice's public key, K_a, is in the user's GnuPG keyring. He can run a utility that selects Alice's key, extracts it from the GnuPG keyring, and passes it to the Cryptfs layer, with instructions to add Alice as an authorized user for the file. The new key list in the Extended Attribute set for the file then contains two copies of the symmetric key, encrypted with different public keys:

$$\{K_s\}K_u$$
$$\{K_s\}K_a$$

Note that this is not an access control directive; it is rather a confidentiality enforcement mechanism that extends beyond the local machine's access control. Without either the user's or Alice's private key, no entity will be able to access the decrypted contents of the file. The machine that harbors such keys will enact its own access control over the decrypted file, based on standard UNIX file permissions and/or ACL's.

When that file is copied to a removable media or attached to an email, as long as the Extended Attributes are preserved, Alice will have all the information that she needs in order to retrieve the symmetric key for the file and de-

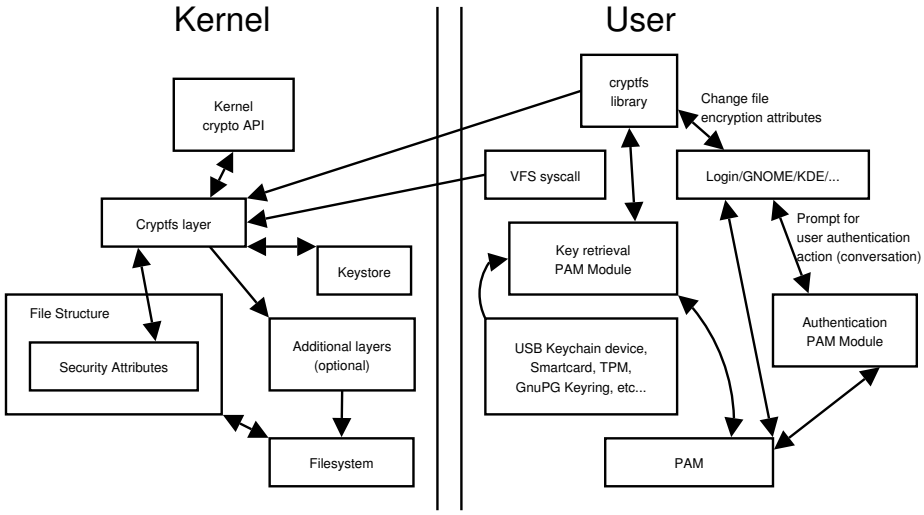


Figure 1: Overview of proposed extended Cryptfs architecture

crypt it. If Alice is also running Cryptfs, when she launches an application that accesses the file, the decryption process is entirely transparent to her, since her Cryptfs layer received her private key from PAM at the time that she logged in.

If the user requires the ability to encrypt a file for access by a group of users, then the user can associate sets of public keys with groups and refer to the groups when granting access. The userspace application that links against libcryptfs can then pass in the public keys to Cryptfs for each member of the group and instruct Cryptfs to add the associated key record to the Extended Attributes. Thus no special support for groups is needed within the Cryptfs layer itself.

6.1 Kernel-level Changes

No modifications to the 2.6 kernel itself are necessary to support the stackable Cryptfs layer. The Cryptfs module's logical divisions include a sysfs interface, a keystore, and the VFS operation routines that perform the encryption and the decryption on reads and writes.

By working with a userspace daemon, it would be possible for Cryptfs to export public key cryptographic operations to userspace. In order to avoid the need for such a daemon while using public key cryptography, the kernel cryptographic API must be extended to support it.

6.2 PAM

At login, the user's public and private keys need to find their way into the kernel Cryptfs layer. This can be accomplished by writing a Pluggable Authentication Module, pam_cryptfs.so. This module will link against libcryptfs and will extract keys from the user's GnuPG keystore. The libcryptfs library will use the sysfs interface to pass the user's keys into the Cryptfs layer.

6.3 libcryptfs

The libcryptfs library works with the Cryptfs's sysfs interface. Userspace utilities, such as pam_cryptfs.so, GNOME/KDE, or stand-alone utilities, will link against this library and use it

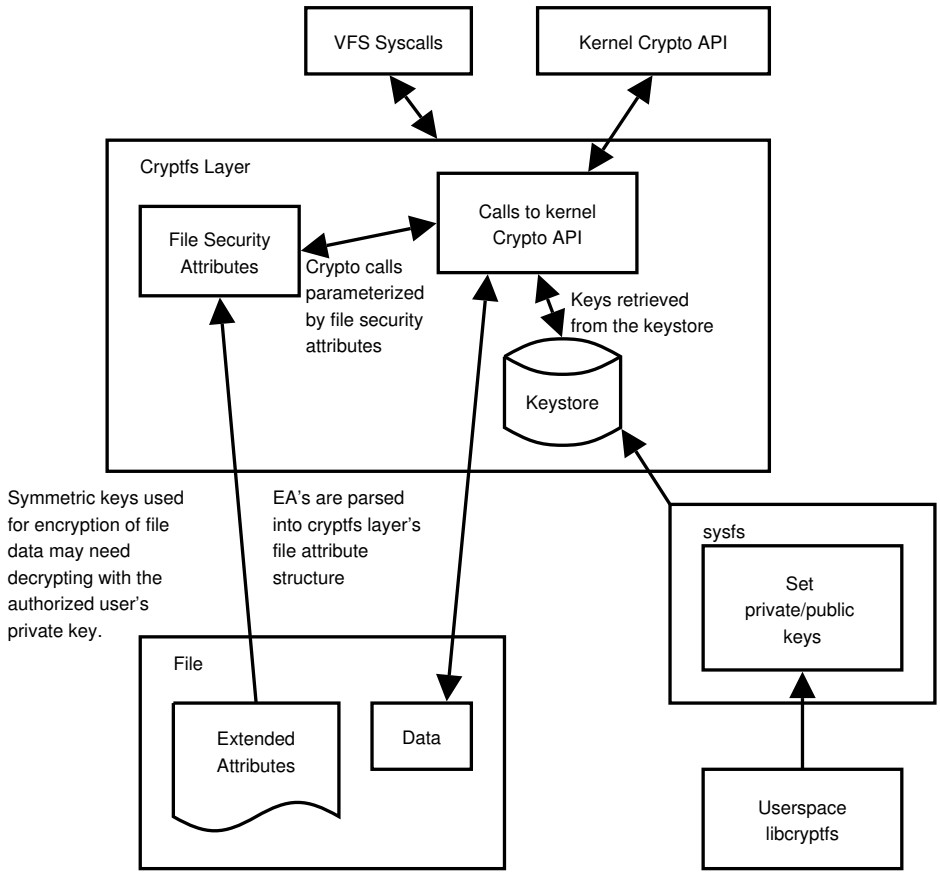


Figure 2: Structure of Cryptfs layer in kernel

to communicate with the kernel Cryptfs layer.

6.4 User Interface

Desktop environments such as GNOME or KDE can link against libcryptfs to provide users with a convenient interface through which to work with the files. For example, by right-clicking on an icon representing the file and selecting "Security", the user will be presented with a window that can be used to control the encryption status of the file. Such options will include whether or not the file is encrypted, which users should be able to encrypt and decrypt the file (identified by their public keys from the user's GnuPG keyring), what cipher is used, what keylength is used, an optional password that encrypts the symmetric key, whether or not to use keyed hashing over extents of the file for integrity, the hash algorithm to use, whether accesses to the file when no key is available should result in an error or in the encrypted blocks being returned (perhaps associated with UID's - good for backup utilities), and other properties that are controlled by the Cryptfs layer.

6.5 Example Walkthrough

When a file's encryption attribute is set, the first thing that the Cryptfs layer will do will be to generate a new symmetric key, which will be used for all encryption and decryption of the file in question. Any data in that file is then immediately encrypted with that key. When using public key-enforced access control, that

key will be encrypted with the process owner's private key and stored as an EA of the file. When the process owner wishes to allow others to access the file, he encrypts the symmetric key with the their public keys. From the user's perspective, this can be done by right-clicking on an icon representing the file, selecting "Security→Add Authorized User Key", and having the user specify the authorized user while using PAM to retrieve the public key for that user.

When using password-enforced access control, the symmetric key is instead encrypted using a key generated from a password. The user can then share that password with everyone who he authorized to access the file. In either case (public key-enforced or password-enforced access control), revocation of access to future versions of the file will necessitate regeneration and re-encryption of the symmetric key.

Suppose the encrypted file is then copied to a removable device and delivered to an authorized user. When that user logged into his machine, his private key was retrieved by the key retrieval Pluggable Authentication Module and sent to the Cryptfs keystore. When that user launches any arbitrary application and attempts to access the encrypted file from the removable media, Cryptfs retrieves the encrypted symmetric key correlating with that user's public key, uses the authenticated user's private key to decrypt the symmetric key, associates that symmetric key with the file, and then proceeds to use that symmetric key for reading and writing the file. This is done in an entirely transparent manner from the perspective of the user, and the file maintains its encrypted status on the removable media throughout the entire process. No modification to the application or applications accessing the file are necessary to implement such functionality.

In the case where a file's symmetric key is encrypted with a password, it will be necessary for the user to launch a daemon that listens for password queries from the kernel cryptfs layer. Without such a daemon, the user's initial attempt to access the file will be denied, and the user will have to use a password set utility to send the password to the cryptfs layer in the kernel.

6.6 Other Considerations

Sparse files present a challenge to encrypted filesystems. Under traditional UNIX semantics, when a user seeks more than a block beyond the end of a file to write, then that space is not stored on the block device at all. These missing blocks are known as "holes."

When holes are later read, the kernel simply fills in zeros into the memory without actually reading the zeros from disk (recall that they do not exist on the disk at all; the filesystem "fakes it"). From the point of view of whatever is asking for the data from the filesystem, the section of the file being read appears to be all zeros. This presents a problem when the file is supposed to be encrypted. Without taking sparse files into consideration, the encryption layer will naïvely assume that the zeros being passed to it from the underlying filesystem are actually encrypted data, and it will attempt to decrypt the zeros. Obviously, this will result in something other that zeros being presented above the encryption layer, thus violating UNIX sparse file semantics.

One solution to this problem is to abandon the concept of "holes" altogether at the Cryptfs layer. Whenever we seek past the end of the file and write, we can actually encrypt blocks of zeros and write them out to the underlying filesystem. While this allows Cryptfs to adhere to UNIX semantics, it is much less efficient. One possible solution might be to store a "hole bitmap" as an Extended Attribute of the

file. Each bit would correspond with a block of the file; a "1" might indicate that the block is a "hole" and should be zero'd out rather than decrypted, and a "0" might indicate that the block should be normally decrypted.

Our proposed extensions to Cryptfs in the near future do not currently address the issues of directory structure and file size secrecy. We recognize that this type of confidentiality is important to many, and we plan to explore ways to integrate such features into Cryptfs, possibly by employing extra filesystem layers to aid in the process.

Extended Attribute content can also be sensitive. Technically, only enough information to retrieve the symmetric decryption key need be accessible by authorized individuals; all other attributes can be encrypted with that key, just as the contents of the file are encrypted.

Processes that are not authorized to access the decrypted content will either be denied access to the file or will receive the encrypted content, depending on how the Cryptfs layer is parameterized. This behavior permits incremental backup utilities to function properly, without requiring access to the unencrypted content of the files they are backing up.

At some point, we would like to include file integrity information in the Extended Attributes. As previously mentioned, this can be accomplished via sets of keyed hashes over extents within the file:

$$H_0 = H\{O_0, D_0, K_s\}$$
$$H_1 = H\{O_1, D_1, K_s\}$$
$$\ldots$$
$$H_n = H\{O_n, D_n, K_s\}$$
$$H_f = H\{H_0, H_1, \ldots, H_n, n, s, K_s\}$$

where n is the number of extents in the file, s is the extent size (also contained as another EA), O_i is the offset number i within the file, D_i is the data from offset O_i to $O_i + s$, K_s is the key that one must possess in order to make authorized changes to the file, and H_f is the hash of the hashes, the number of extents, the extent size, and the secret key, to help detect when an attacker swaps around extents or alters the extent size.

Keyed hashes prove that whoever modified the data had access to the shared secret, which is, in this case, the symmetric key. Digital signatures can also be incorporated into Cryptfs. Executables downloaded over the Internet can often be of questionable origin or integrity. If you trust the person who signed the executable, then you can have a higher degree of certainty that the executable is safe to run if the digital signature is verifiable. The verification of the digital signature can be dynamically performed at the time of execution.

As previously mentioned, in addition to the extensions to the Cryptfs stackable layer, this effort is requiring the development of a cryptfs library, a set of PAM modules, hooks into GNOME and KDE, and some utilities for managing file encryption. Applications that copy files with Extended Attributes must take steps to make sure that they preserve the Extended Attributes.[7]

7 Conclusion

Linux currently has a comprehensive framework for managing filesystem security. Standard file security attributes, process credentials, ACL, PAM, LSM, Device Mapping (DM) Crypt, and other features together provide good security in a contained environment. To extend access control enforcement over individual files beyond the local environment, you must use encryption in a way that can be easily

[7] See `http://www.suse.de/~agruen/ea-acl-copy/`

applied to individual files. The currently employed processes of encrypting and decrypting files, however, is inconvenient and often obstructive.

By integrating the encryption and the decryption of the individual files into the filesystem itself, associating encryption metadata with the individual files, we can extend Linux security to provide seamless encryption-enforced access control and integrity auditing.

8 Recognitions

We would like to express our appreciation for the contributions and input on the part of all those who have laid the groundwork for an effort toward transparent filesystem encryption. This includes contributors to FiST and Cryptfs, GnuPG, PAM, and many others from which we are basing our development efforts, as well as several members of the kernel development community.

9 Legal Statement

References

[1] E. Zadok, L. Badulescu, and A. Shender. Cryptfs: A stackable vnode level encryption file system. *Technical Report CUCS-021-98, Computer Science Department*, Columbia University, 1998.

[2] J.S. Heidemann and G.J. Popek. File system development with stackable layers. *ACM Transactions on Computer Systems*, 12(1):58–89, February 1994.

[3] E. Zadok and J. Nieh. FiST: A Language for Stackable File Systems. *Proceedings of the Annual USENIX Technical Conference*, pp. 55–70, San Diego, June 2000.

[4] S.C. Kothari, Generalized Linear Threshold Scheme, *Advances in Cryptology: Proceedings of CRYPTO 84*, Springer-Verlag, 1985, pp. 231–241.

[5] Matt Blaze. "Key Management in an Encrypting File System," Proc. *Summer '94 USENIX Tech. Conference*, Boston, MA, June 1994.

[6] For more information on Extended Attributes (EA's) and Access Control Lists (ACL's), see `http://acl.bestbits.at/` or `http://www.suse.de/~agruen/acl/chapter/fs_acl-en.pdf`

[7] For more information on GnuPG, see `http://www.gnupg.org/`

[8] For more information on OpenSSL, see `http://www.openssl.org/`

[9] For more information on IBM Lotus Notes, see `http://www-306.ibm.com/software/lotus/`. Information on Notes security can be obtained from `http://www-10.lotus.com/ldd/today.nsf/f01245ebfc115aaf8525661a006b86b9/232e604b847d2cad88256ab90074e298?OpenDocument`

[10] For more information on Pluggable Authentication Modules (PAM), see `http://www.kernel.org/pub/linux/libs/pam/`

[11] For more information on Mandatory Access Control (MAC), see http://csrc.nist.gov/publications/nistpubs/800-7/node35.html

[12] For more information on Discretionary Access Control (DAC), see http://csrc.nist.gov/publications/nistpubs/800-7/node25.html

[13] For more information on the Trusted Computing Platform Alliance (TCPA), see http://www.trustedcomputing.org/home

[14] For more information on Linux Security Modules (LSM's), see http://lsm.immunix.org/

[15] For more information on Security-Enhanced Linux (SE Linux), see http://www.nsa.gov/selinux/index.cfm

[16] For more information on Tripwire, see http://www.tripwire.org/

[17] For more information on AIDE, see http://www.cs.tut.fi/~rammer/aide.html

[18] For more information on Samhain, see http://la-samhna.de/samhain/

[19] For more information on Logcrypt, see http://www.lunkwill.org/logcrypt/

[20] For more information on Loop-aes, see http://sourceforge.net/projects/loop-aes/

[21] For more information on PPDD, see http://linux01.gwdg.de/~alatham/ppdd.html

[22] For more information on CFS, see http://sourceforge.net/projects/cfsnfs/

[23] For more information on BestCrypt, see http://www.jetico.com/index.htm#/products.htm

[24] For more information on TCFS, see http://www.tcfs.it/

[25] For more information on EncFS, see http://arg0.net/users/vgough/encfs.html

[26] For more information on CryptoFS, see http://reboot.animeirc.de/cryptofs/

[27] For more information on SSHFS, see http://lufs.sourceforge.net/lufs/fs.html

[28] For more information on the Light-weight Auditing Framework, see http://lwn.net/Articles/79326/

[29] For more information on Reiser4, see http://www.namesys.com/v4/v4.html

[30] NFSv4 RFC 3010 can be obtained from http://www.ietf.org/rfc/rfc3010.txt

Hotplug Memory and the Linux VM

Dave Hansen, Mike Kravetz, with Brad Christiansen
IBM Linux Technology Center
haveblue@us.ibm.com, kravetz@us.ibm.com, bradc1@us.ibm.com

Matt Tolentino
Intel
matthew.e.tolentino@intel.com

Abstract

This paper will describe the changes needed to the Linux memory management system to cope with adding or removing RAM from a running system. In addition to support for physically adding or removing DIMMs, there is an ever-increasing number of virtualized environments such as UML or the IBM pSeries™ Hypervisor which can transition RAM between virtual system images, based on need. This paper will describe techniques common to all supported platforms, as well as challenges for specific architectures.

1 Introduction

As Free Software Operating Systems continue to expand their scope of use, so do the demands placed upon them. One area of continuing growth for Linux is the adaptation to incessantly changing hardware configurations at runtime. While initially confined to commonly removed devices such as keyboards, digital cameras or hard disks, Linux has recently begun to grow to include the capability to hot-plug integral system components. This paper describes the changes necessary to enable Linux to adapt to dynamic changes in one of the most critical system resource—system RAM.

2 Motivation

The underlying reason for wanting to change the amount of RAM is very simple: availability. The systems that support memory hot-plug operations are designed to fulfill mission critical roles; significant enough that the cost of a reboot cycle for the sole purpose of adding or replacing system RAM is simply too expensive. For example, some large ppc64 machines have been reported to take well over thirty minutes for a simple reboot. Therefore, the downtime necessary for an upgrade may compromise the five nine uptime requirement critical to high-end system customers [1].

However, memory hotplug is not just important for big-iron. The availability of high speed, commodity hardware has prompted a resurgence of research into virtual machine monitors—layers of software such as Xen [2], VMWare [3], and conceptually even User Mode Linux that allow for multiple operating system instances to be run in isolated, virtual domains. As computing hardware density has increased, so has the possibility of splitting up that computing power into more manageable pieces. The capability for an operating system to expand or contract the range of physical

memory resources available presents the possibility for virtual machine implementations to balance memory requirements and improve the management of memory availability between domains[1]. This author currently leases a small User Mode Linux partition for small Internet tasks such as DNS and low-traffic web serving. Similar configurations with an approximately 100 MHz processor and 64 MB of RAM are not uncommon. Imagine, in the case of an accidental Slashdotting, how useful radically growing such a machine could be.

3 Linux's Hotplug Shortcomings

Before being able to handle the full wrath of Slashdot. we have to consider Linux's current design. Linux only has two data structures that absolutely limit the amount of RAM that Linux can handle: the page allocator bitmaps, and `mem_map[]` (on contiguous memory systems). The page allocator bitmaps are very simple in concept, have a bit set one way when a page is available, and the opposite when it has been allocated. Since there needs to be one bit available for each page, it obviously has to scale with the size of the system's total RAM. The bitmap memory consumption is approximately 1 bit of memory for each page of system RAM.

4 Resizing `mem_map[]`

The `mem_map[]` structure is a bit more complicated. Conceptually, it is an array, with one `struct page` for each physical page which the system contains. These structures contain bookkeeping information such as flags indicating page usage and locking structures. The complexity with the `struct page`s is associated when their size. They have a size of 40 bytes each on i386 (in the 2.6.5 kernel). On a system with 4096 byte hardware pages, this implies that about 1% of the total system memory will be consumed by `struct page`s alone. This use of 1% of the system memory is not a problem in and of itself. But, it does other problems.

The Linux page allocator has a limitation on the maximum amounts of memory that it can allocate to a single request. On i386, this is 4MB, while on ppc64, it is 16MB. It is easy to calculate that anything larger than a 4GB i386 system will be unable to allocate its `mem_map[]` with the normal page allocator. Normally, this problem with `mem_map` is avoided by using a boot-time allocator which does not have the same restrictions as the allocator used at runtime. However, memory hotplug requires the ability to grow the amount of `mem_map[]` used at runtime. It is not feasible to use the same approach as the page allocator bitmaps because, in contrast, they are kept to small-enough sizes to not impinge on the maximum size allocation limits.

4.1 `mem_map[]` preallocation

A very simple way around the runtime allocator limitations might be to allocate sufficient memory form `mem_map[]` at boot-time to account for any amount of RAM that could possibly be added to the system. But, this approach quickly breaks down in at least one important case. The `mem_map[]` must be allocated in low memory, an area on i386 which is approximately 896MB in total size. This is very important memory which is commonly exhausted [4],[5],[6]. Consider an 8GB system which could be expanded to 64GB in the future. Its normal `mem_map[]` use would be around 84MB, an acceptable 10% use of low memory. However, had `mem_map[]` been preallocated to handle a total capacity of 64GB of system memory, it would use an astound-

[1] err, I could write a lot about this, so I won't go any further

ing 71% of low memory, giving any 8GB system all of the low memory problems associated with much larger systems.

Preallocation also has the disadvantage of imposing limitations possibly making the user decide how large they expect the system to be, either when the kernel is compiled, or when it is booted. Perhaps the administrator of the above 8GB machine knows that it will never get any larger than 16GB. Does that make the low memory usage more acceptable? It would likely solve the immediate problem, however, such limitations and user intervention are becoming increasingly unacceptable to Linux vendors, as they drastically increase possible user configurations, and support costs along with it.

4.2 Breaking `mem_map[]` up

Instead of preallocation, another solution is to break up `mem_map[]`. Instead of needing massive amounts of memory, smaller ones could be used to piece together `mem_map[]` from more manageable allocations Interestingly, there is already precedent in the Linux kernel for such an approach. The discontiguous memory support code tries to solve a different problem (large holes in the physical address space), but a similar solution was needed. In fact, there has been code released to use the current discontigmem support in Linux to implement memory hotplug. But, this has several disadvantages. Most importantly, it requires hijacking the NUMA code for use with memory hotplug. This would exclude the use of NUMA and memory hotplug on the same system, which is likely an unacceptable compromise due to the vast performance benefits demonstrated from using the Linux NUMA code for its intended use [6].

Using the NUMA code for memory hotplug is a very tempting proposition because in addition to splitting up `mem_map[]` the NUMA support also handles discontiguous memory. Discontiguous memory simply means that the system does not lay out all of its physical memory in a single block, rather there are holes. Handling these holes with memory hotplug is very important, otherwise the only memory that could be added or removed would be on the end.

Although an approch similar to this “node hotplug” approach will be needed when adding or removing entire NUMA nodes, using it on a regular SMP hotplug system could be disastrous. Each discontiguous area is represented by several data structures but each has at least one `structzone`. This structure is the basic unit which Linux uses to pool memory. When the amounts of memory reach certain low levels, Linux will respond by trying to free or swap memory. Artificially creating too many zones causes these events to be triggered much too early, degrading system performance and under-utilizing available RAM.

5 CONFIG_NONLINEAR

The solution to both the `mem_map[]` and discontiguous memory problems comes in a single package: nonlinear memory. First implemented by Daniel Phillips in April of 2002 as an alternative to discontiguous memory, nonlinear solves a similar set of problems.

Laying out `mem_map[]` as an array has several advantages. One of the most important is the ability to quickly determine the physical address of any arbitrary `struct page`. Since `mem_map[N]` represents the Nth page of physical memory, the physical address of the memory represented by that `struct page` can be determined by simple pointer arithmetic:

Once `mem_map[]` is broken up these simple

```
physical_address = (&mem_map[N] - &mem_map[0]) * sizeof(struct page)

struct page N = mem_map[(physical_address / sizeof(struct page)]
```

Figure 1: Physical Address Calculations

calculations are no longer possible, thus another approach is required. The nonlinear approach is to use a set of two lookup tables, each one complementing the above operations: one for converting `struct page` to physical addresses, the other for doing the opposite. While it would be possible to have a table with an entry for every single page, that approach wastes far too much memory. As a result, nonlinear handles pages in uniformly sized sections, each of which has its own `mem_map[]` and an associated physical address range. Linux has some interesting conventions about how addresses are represented, and this has serious implications for how the nonlinear code functions.

5.1 Physical Address Representations

There are, in fact, at least three different ways to represent a physical address in Linux: a physical address, a `struct page`, and a page frame number (pfn). A pfn is traditionally just the physical address divided by the size of a physical page (the *N* in the above in Figure 1). Many parts of the kernel prefer to use a pfn as opposed to a `struct page` pointer to keep track of pages because pfn's are easier to work with, being conceptually just array indexes. The page allocator bitmaps discussed above are just such a part of the kernel. To allocate or free a page, the page allocator toggles a bit at an index in one of the bitmaps. That index is based on a pfn, not a `struct page` or a physical address.

Being so easily transposed, that decision does not seem horribly important. But it does cause a serious problem for memory hotplug. Consider a system with 100 1GB DIMM slots that support hotplug. When the system is first booted, only one of these DIMM slots is populated. Later on, the owner decides to hotplug another DIMM, but puts it in slot 100 instead of slot 2. Now, nonlinear has a bit of a problem: the new DIMM happens to appear at a physical address 100 times higher address than the first DIMM. The `mem_map[]` for the new DIMM is split up properly, but the allocator bitmap's length is directly tied to the pfn, and thus the physical address of the memory.

Having already stated that the allocator bitmap stays at manageable sizes, this still does not seem like much of an issue. However, the physical address of that new memory *could* have an even greater range than 100 GB; it has the capability to have many, many terabytes of range, based on the hardware. Keeping allocator bitmaps for terabytes of memory could conceivably consume all system memory on a small machine, which is quite unacceptable. Nonlinear offers a solution to this by introducing a new way to represent a physical address: a fourth addressing scheme. With three addressing schemes already existing, a fourth seems almost comical, until its small scope is considered. The new scheme is isolated to use inside of a small set of core allocator functions a single place in the memory hotplug code itself. A simple lookup table converts these new "linear" pfns into the more familiar physical pfns.

5.2 Issues with CONFIG_NONLINEAR

Although it greatly simplifies several issues, nonlinear is not without its problems. Firstly, it does require the consultation of a small number of lookup tables during critical sections of code. Random access of these tables is likely to cause cache overhead. The more finely grained the units of hotplug, the larger these tables will grow, and the worse the cache effects.

Another concern arises with the size of the nonlinear tables themselves. While they allow pfns and `mem_map[]` to have nonlinear relationships, the nonlinear structures themselves remain normal, everyday, linear arrays. If hardware is encountered with sufficiently small hotplug units, and sufficiently large ranges of physical addresses, an alternate scheme to the arrays may be required. However, it is the authors' desire to keep the implementation simple, until such a need is actually demonstrated.

6 Memory Removal

While memory addition is a relatively black-and-white problem, memory removal has many more shades of gray. There are many different ways to use memory, and each of them has specific challenges for *un*using it. We will first discuss the kinds of memory that Linux has which are relevant to memory removal, along with strategies to go about unusing them.

6.1 "Easy" User Memory

Unusing memory is a matter of either moving data or simply throwing it away. The easiest, most straightforward kind of memory to remove is that whose contents can just be discarded. The two most common manifestations of this are clean page cache pages and swapped pages. Page cache pages are either dirty (containing information which has not been written to disk) or clean pages, which are simply a copy of something that *is* present on the disk. Memory removal logic that encounters a clean page cache page is free to have it discarded, just as the low memory reclaim code does today. The same is true of swapped pages; a page of RAM which has been written to disk is safe to discard. (Note: there is usually a brief period between when a page is written to disk, and when it is actually removed from memory.) Any page that *can* be swapped is also an easy candidate for memory removal, because it can easily be turned into a swapped page with existing code.

6.2 Swappable User Memory

Another type of memory which is very similar to the two types above is something which is only used by user programs, but is for some reason not a candidate for swapping. This at least includes pages which have been `mlock()`'d (which is a system call to prevent swapping). Instead of discarding these pages out of RAM, they must instead be moved. The algorithm to accomplish this should be very similar to the algorithm for a complete page swapping: freeze writes to the page, move the page's contents to another place in memory, change all references to the page, and re-enable writing. Notice that this is the same process as a complete swap cycle except that the writes to the disk are removed.

6.3 Kernel Memory

Now comes the hard part. Up until now, we have discussed memory which is being used by user programs. There is also memory that Linux sets aside for its own use and this comes in many more varieties than that used by user programs. The techniques for dealing with this memory are largely still theoretical, and do not have existing implementations.

Remember how the Linux page allocator can only keep track of pages in powers of two? The Linux slab cache was designed to make up for that [6], [7]. It has the ability to take those powers of two pages, and chop them up into smaller pieces. There are some fixed-size groups for common allocations like 1024, 1532, or 8192 bytes, but there are also caches for certain kinds of data structures. Some of these caches have the ability to attempt to shrink themselves when the system needs some memory back, but even that is relatively worthless for memory hotplug.

6.4 Removing Slab Cache Pages

The problem is that the slab cache's shrinking mechanism does not concentrate on shrinking any particular memory, it just concentrates on shrinking, period. Plus, there's currently no mechanism to tell *which* slab a particular page belongs to. It could just as easily be a simply discarded dcache entry as it could be a completely immovable entry like a `pte_chain`. Linux will need mechanisms to allow the slab cache shrinking to be much more surgical.

However, there will always be slab cache memory which is not covered by any of the shrinking code, like for generic `kmalloc()` allocations. The slab cache could also make efforts to keep these "mystery" allocations away from those for which it knows how to handle.

While the record-keeping for some slab-cache pages is sparse, there is memory with even more mysterious origins. Some is allocated early in the boot process, while other uses pull pages directly out of the allocator never to be seen again. If hot-removal of these areas is required, then a different approach must be employed: direct replacement. Instead of simply reducing the usage of an area of memory until it is unused, a one-to-one replacement of this memory is required. With the judicious use of page tables, the best that can be done is to preserve the virtual address of these areas. While this is acceptable for most use, it is not without its pitfalls.

6.5 Removing DMA Memory

One unacceptable place to change the physical address of some data is for a device's DMA buffer. Modern disk controllers and network devices can transfer their data directly into the system's memory without the CPU's direct involvement. However, since the CPU is not involved, the devices lack access to the CPU's virtual memory architecture. For this reason, all DMA-capable devices' transfers are based on the physical address of the memory to which they are transferring. Every user of DMA in Linux will either need to be guaranteed to not be affected by memory replacement, or to be notified of such a replacement so that it can take corrective action. It should be noted, however, that the virtualization layer on ppc64 can properly handle this remapping in its IOMMU. Other architectures with IOMMUs should be able to employ similar techniques.

6.6 Removal and the Page Allocator

The Linux page allocator works by keeping lists of groups of pages in sizes that are powers of two times the size of a page. It keeps a list of groups that are available for each power of two. However, when a request for a page is made, the only real information provided is for the *size* required, there is no component for specifically specifying which particular memory is required.

The first thing to consider before removing memory is to make sure that no other part of the system is using that piece of memory. Thankfully, that's exactly what a normal allocation does: make sure that it is alone in

its use of the page. So, making the page allocator support memory removal will simply involve walking the same lists that store the page groups. But, instead of simply taking the first available pages, it will be more finicky, only "allocating" pages that are among those about to be removed. In addition, the allocator should have checks in the `free_pages()` path to look for pages which were selected for removal.

1. Inform allocator to catch any pages in the area being removed.
2. Go into allocator, and remove any pages in that area.
3. Trigger page reclaim mechanisms to trigger `free()`s, and hopefully unuse all target pages.
4. If not complete, goto 3.

6.7 Page Groupings

As described above, the page allocator is the basis for all memory allocations. However, when it comes time to remove memory a fixed size block of memory is what is removed. These blocks correspond to the sections defined in the the implementation of nonlinear memory. When removing a section of memory, the code performing the remove operation will first try to essentially allocate all the pages in the section. To remove the section, all pages within the section must be made free of use by some mechanism as described above. However, it should be noted that some pages will not be able to be made available for removal. For example, pages in use for kernel allocations, DMA or via the slab-cache. Since the page allocator makes no attempt to group pages based on usage, it is possible in a worst case situation that every section contains one in-use page that can not be removed. Ideally, we would like to group pages based on their usage to allow the maximum number of sections to be removed.

Currently, the definition of zones provides some level of grouping on specific architectures. For example, on i386, three zones are defined: DMA, NORMAL and HIGHMEM. With such definitions, one would expect most non-removable pages to be allocated out of the DMA and NORMAL zones. In addition, one would expect most HIGHMEM allocations to be associated with userspace pages and thus removable. Of course, when the page allocator is under memory pressure it is possible that zone preferences will be ignored and allocations may come from an alternate zone. It should also be noted that on some architectures, such as ppc64, only one zone (DMA) is defined. Hence, zones can not provide grouping of pages on every architecture. It appears that zones do provide some level of page grouping, but possibly not sufficient for memory hotplug.

Ideally, we would like to experiment with teaching the page allocator about the use of pages it is handing out. A simple thought would be to introduce the concept of sections to the allocator. Allocations of a specific type are directed to a section that is primarily used for allocations of that same type. For example, when allocations for use within the kernel are needed the allocator will attempt to allocate the page from a section that contains other internal kernel allocations. If no such pages can be found, then a new section is marked for internal kernel allocations. In this way pages which can not be easily freed are grouped together rather than spread throughout the system. In this way the page allocator's use of sections would be analogous to the slab caches use of pages.

7 Conclusion

The prevalence of hotplug-capable Linux systems is only expanding. Support for these systems will make Linux more flexible and will make additional capabilities available to other parts of the system.

Legal Statement

References

[1] *Five Nine at the IP Edge* `http://www.iec.org/online/tutorials/five-nines`

[2] Barham, Paul, et al. *Xen and the Art of Virtualization* Proceedings of the ACM Symposium on Operating System Principles (SOSP), October 2003.

[3] Waldspurger, Carl *Memory Resource Management in VMware ESX Server* Proceedings of the USENIX Association Symposium on Operating System Design and Implementation, 2002. pp 181–194.

[4] Dobson, Matthew and Gaughen, Patricia and Hohnbaum, Michael. *Linux Support for NUMA Hardware* Proceedings of the Ottawa Linux Symposium. July 2003. pp 181–196.

[5] Gorman, Mel *Understanding the Linux Virtual Memory Manager* Prentice Hall, NJ. 2004.

[6] Martin Bligh and Dave Hansen *Linux Memory Management on Larger Machines* Proceeedings of the Ottawa Linux Symposium 2003. pp 53–88.

[7] Bonwick, Jeff *The Slab Allocator: An Object-Caching Kernel Memory Allocator* Proceedings of USENIX Summer 1994 Technical Conference `http://www.usenix.org/publications/library/proceedings/bos94/bonwick.html`

kobjects and krefs

lockless reference counting for kernel structures

Greg Kroah-Hartman *
Linux Technology Center
IBM Corp.
greg@kroah.com
gregkh@us.ibm.com

Abstract

This paper will describe the current kobject and kref kernel structures in detail. It will cover why they were created, how to use them, and how the internals work. It will also cover a few directions that these structures might be taking in the future.

1 Introduction

The Linux kernel file `Documentation/CodingStyle` has the following statement about reference counting:

> Data structures that have visibility outside the single-threaded environment they are created and destroyed in should always have reference counts. In the kernel, garbage collection doesn't exist (and outside the kernel garbage collection is slow and inefficient), which means that you absolutely _have_ to reference count all your uses.

This requirement of providing proper reference counting for kernel structures has caused developers to create their own logic and functions to implement this feature. During the development of the Linux Kernel Driver model[4], a simple structure, `struct kobject`, was created that provided automatic reference counting for any user of the object. Unfortunately, `struct kobject` is closely tied to the kernel driver model, and for any data structure that does not want to show up in sysfs, and participate in the global kernel "web woven by a spider on drugs"[2], using a struct kobject only for reference counting is a big waste of memory resources and is much more complex than needed. To this end, the data structure, `struct kref`, was created to provide a simple, and hopefully failproof method of adding proper reference counting to any kernel data structure.

2 How to use it

To use the `struct kref` structure, simply embed it within the structure that reference counting is needed for. For example, to add reference counting to a structure called `struct foo` then it would be defined as:

```
struct foo {
    ...
    struct kref kref;
    ...
```

*This work represents the view of the author and does not necessarily represent the view of IBM.

```
};
```

It is not important that the `struct kref` structure be the first or last element of the structure that it is embedded in. The only requirement is that the whole `struct kref` structure be in the structure being reference counted, not a pointer to the a `struct kref` structure.

When the `struct foo` structure is initialized, the `kref` variable must also be initialized before reference counting can be used. This is done with a call to the `kref_init` function:

```
struct foo *foo;
foo = kmalloc(sizeof(*foo),
              GFP_KERNEL);
kref_init(&foo->kref,
          foo_release);
```

The parameter `foo_release` is a pointer The first parameter of `kref_init` is a pointer to the `struct kref` structure that is to be initialized. The second parameter is a pointer to the release function for the structure. This release function is described in detail below.

After the kref structure has been initialized, the internal reference count of the structure is set to 1. Now the reference count can be incremented and decremented at will.

To increment the reference count of a kref structure, the function `kref_get` is called:

```
/* get a new reference to our
   foo structure */
kref_get(&foo->kref);
```

When a user of the structure is finished with it, the `kref_put` function should be called to release the reference:

```
/* finished with this
   foo structure */
kref_put(&foo->kref);
```

This function should also be called after the original creator of the structure that the kref variable is in, is finished with the structure. The `kfree` function must *NOT* be directly called because other portions of the kernel could have valid references to this structure.

After the `kref_put` function is called, the structure can not be referred to by any future code, as the memory for that structure could be now gone.

When the last reference count is released, the function that was passed to the original `kref_init` function is called to release the memory used by the structure. The prototype of this function must accept a pointer to a `struct kref`:

```
void foo_release(struct kref
                 *kref)
{
    struct foo *foo;

    foo = container_of(foo,
                       struct foo,
                       kref);
    kfree(foo);
}
```

As the above example function shows, to get back to the original `struct foo` structure location, the `container_of` macro is used. For a complete description of how the `container_of` macro works, please see[1].

As there are not any locks within the `kref` structure, there are three rules that need to be followed when using this reference counting logic:

- If the code accessing the variable already has a valid reference to the structure, it is

safe, and required to increment that reference with a call to `kref_get` in order to give the variable to any other piece of code.

- If the code accessing the variable already has a valid reference to the structure, then it is safe to release that reference with a call to `kref_put`.
- If the code wanting to access the variable, does not have a valid reference, then it needs to serialize with a place within the code where the last call to `kref_put` put could happen.

This last rule can not be emphasized enough. The only reason that the `struct kref` can work without any internal locks is because a call to `kref_get` can not happen at the same time that `kref_put` is happening. In order to ensure this, a simple lock for the driver or subsystem that owns the specific `struct kref` reference can be used.

An example of using such a lock can be seen in Figure 1.

So, with the three simple functions, `kref_init`, `kref_get`, and `kref_put`, combined with a release function that the caller provides, complete reference counting can be added to any kernel structure.

3 How it works

`struct kref` is a very tiny structure with only two elements:

```
struct kref {
    atomic_t refcount;
    void (*release)(struct kref *kref);
};
```

The `refcount` variable is an atomic counter that is used to hold the reference count of the structure. The `release` variable is a pointer to a function that will be called when the last user of the structure is finished with the structure.

The `kref_init` function is a mere three lines long:

```
void kref_init(struct kref *kref,
               void (*release)
               (struct kref *kref))
{
    WARN_ON(release == NULL);
    atomic_set(&kref->refcount,1);
    kref->release = release;
}
```

First a warning is printed out to the syslog if a `release` callback is not provided, as this is not allowed. Then the `refcount` variable is initialized to 1 as the structure needs to have a single initial reference count. After that the `release` function pointer is stored in the `release` variable in the structure.

The `kref_get` function is also only three lines of code:

```
struct kref *kref_get(struct kref *kref)
{
    WARN_ON(!atomic_read(&kref->refcount));
    atomic_inc(&kref->refcount);
    return kref;
}
```

Again, a warning is printed out to the syslog if the `refcount` variable is zero. This catches the very common error of calling `kref_get` without first calling `kref_init`. After that, the `refcount` variable is incremented, and then a pointer to the same structure is returned. This return type makes it easier for code to do things pass the result of `kref_get` as a function parameter:

```
do_foo(kref_get(my_kref));
```

Keeping with the tradition of tiny functions, the `kref_put` function weighs in at a whopping two lines:

```
/* prevent races between open() and disconnect() */
static DECLARE_MUTEX (disconnect_sem);

static int skel_open(struct inode *inode, struct file *file)
{
    struct usb_skel *dev;
    struct usb_interface *interface;

    /* prevent disconnects */
    down (&disconnect_sem);

    interface = usb_find_interface(&skel_driver, iminor(inode));
    dev = usb_get_intfdata(interface);

    /* increment our usage count for the device */
    kref_get(&dev->kref);
    up(&disconnect_sem);

    ...
}

static void skel_disconnect(struct usb_interface *interface)
{
    struct usb_skel *dev;
    int minor = interface->minor;

    /* prevent skel_open() from racing skel_disconnect() */
    down (&disconnect_sem);

    dev = usb_get_intfdata(interface);
    usb_set_intfdata(interface, NULL);

    /* give back our minor */
    usb_deregister_dev(interface, &skel_class);

    /* decrement our usage count */
    kref_put(&dev->kref);

    up(&disconnect_sem);
}
```

Figure 1: Using a lock to ensure safe access to `kref_put`

```
void kref_put(struct kref *kref)            kref->release(kref);
{                                       }
    if (atomic_dec_and_test
              (&kref->refcount))
```

This function decrements the value stored in the `refcount` variable, and if the result is zero, this was the last reference to the structure, so the function stored in the `release` variable is called to clean up the memory used by this structure.

4 kref vs. kobject

This paper has focused on on how `struct kref` works, and ignored `struct kobject`. For the most part, both structures work identically, with the following minor differences:

- `struct kobject` does not contain a `release` function. When a `struct kobject`'s last reference count is decremented, the release function of the `struct kset` that is associated with the `struct kobject` is called. For more details on how `struct kobject` and `struct kset` is related, please see [3].
- A `struct kobject` can be initialized with two different functions, `kobject_register` or `kobject_init`. `kobject_register` calls `kobject_init` and then calls `kobject_add` to add the kobject to the sysfs hierarchy. If a `struct kobject` is to not be used within the sysfs hierarchy, then `kobject_add` should never be called.
- A `struct kobject` can have its reference count incremented with a call to `kobject_get` and decremented with a call to `kobject_put`. But if the kobject was initialized with the sysfs core with a call to either `kobject_add` or `kobject_register`, then it needs to be removed from it with a call to `kobject_del`, which will also call `kobject_put` on the `struct kobject`. After a `struct kobject` has had `kobject_del` called for it, the `kboject_get` function can not be called on the variable without having a previous reference count already on the variable. This is the same as the previously mentioned issue for calling `kref_put` without serializing the access.
- Before using a `struct kobject`, the structure must be initialized to zero by using `memset` before `kobject_init` or `kobject_register` is called. If not, a warning will be printed out to the syslog.

5 Future

In future releases of the Linux kernel, the `struct kobject` will probably loose its internal reference count and use the `struct kref` instead. If this happens, `struct kref` might have to be changed in order to support passing the `release` callback as a parameter to the `kref_put` function, in order to save the storage size of the function pointer from the structure.

Other kernel uses of a `atomic_t` variable will probably be converted to use the `struct kref` interface instead of providing their own logic to handle reference counting.

6 Legal Statement

References

[1] Greg Kroah-Hartman. The Driver Model Core, part 1. `http://www.linuxjournal.com/modules.php?op=modload&name=NS-lj-issues/is%sue110&file=6717s2`, June 2003.

[2] Linux Weekly News - Driver porting: Device model overview. `http://lwn.net/Articles/31185/`.

[3] Linux Weekly News - The zen of kobjects. `http://lwn.net/Articles/51437/`.

[4] Patrick Mochel. The Linux Kernel Device Model. In *Linux.conf.au*, Perth, Australia, January 2003.

The Cursor Wiggles Faster: Measuring Scheduler Performance

Rick Lindsley
IBM Linux Technology Center
ricklind@us.ibm.com

Abstract

Trying to pin down whether changes to the 2.5 and 2.6 scheduler have helped or hurt performance, especially on interactive programs, has been both difficult to quantify and very subjective. One favored test has been to create your favorite load and then move your cursor around and observe how slow or fast it is. Another one is to drag a window across your desktop and see how quickly it gets redrawn. And I would certainly be skewered if I didn't mention what is probably the favorite: playing your favorite music while under load and listening intently for skips.

Unfortunately, all these measurements are subjective, and even, at times, argumentative. With scheduler statistics installed, one can accurately measure such things as the amount of time processes are spending on the processor or the amount of time they are waiting for the processor. This means that on SMP and NUMA machines, load balancing efforts can be objectively evaluated, and process migration decisions more effectively reviewed. And all of this can be done with no measurable impact to the system.

This paper will describe what information can be captured, use that information to characterize some simple loads, and describe how that same information may be coordinated with other system measurements both to characterize new loads, and to more clearly identify scheduler shortcomings.

1 Introduction

As the 2.5 code revisions came out in mid- to late 2003, the scheduler, like much of the 2.5 release, became more and more stable. True, there was still work to be done in some areas, like SMP and NUMA. Although an increasing number of dual-CPU desktops and even laptops introduced more users to the world of SMP, it was the high end users with 16, 32, 128, or even more CPUs that really were stretching the existing SMP and NUMA code. The increasing load on the existing infrastructure was causing developers to realize that code paths they previously thought "impossible" were really "rarely," and paths deemed "infrequent" were unfortunately morphing to "once or twice a day."

And an odd thing happened on the way to better code for the high end machines. Those pesky desktop and laptop users got in the way. With every fix that would demonstrably improve the situation for the big iron, dozens of desktop and laptop owners would immediately pick up the new code, try it out, and more often than not, pronounce it faulty. Why? Because *their* 2-proc SMP machines were used very differently than the file servers and web servers that the 128-proc systems had become.

The testing and measurements that had gone into verifying the patch did not test the system the same way the desktop users did. Consequently, these desktop users saw very different results, and formed very different opinions about the correctness and usefulness of these high-end SMP fixes.

And while their opinions mattered, of course, addressing their concerns was difficult. They were using human eyes and ears—notoriously unreliable biological components known to be fraught with frequent failure and highly subjective readouts—to detect problems with code. These observations needed to be backed up with numbers somehow.

2 Why is the wiggle so important?

So why weren't the big iron folks seeing the same problems as the desktop people if they were both utilizing the same code? The answer lay in usage patterns. People with laptops and desktops did not run two dozen instances of a server daemon that depended on ultra fast cache and great amounts of parallelism. They did not have petabytes of disk, and typically did not have gigabytes of memory either. They didn't read terabytes of disk per minute, nor expect to fully utilize their bus bandwidth on a regular basis.

These folks browsed the web, sorted mail, and compiled kernels while, in the background, they listened to their favorite playlist. While doing this, they would notice that with the new scheduler mods, their windows took longer to redraw. Or their cursor moved more sluggishly under this relatively heavy load. Or their music skipped now and then because their music player didn't get back on the CPU soon enough to catch the next few notes.

That's not to make light of their complaints; they were uncovering real problems that existing testing was inadequate to find. In fact, there were two main problems that needed to be solved. One was to close the testing hole by reliably repeating the tests that the desktop users were running, and repeating them on as wide a variety of hardware as the original patches had been run on. The other was that even the desktop users quibbled among themselves, sometimes, about whether wiggles, skips, and redraws had degraded. It was important to find a way to measure this "wiggle effect" in some quantifiable, objective way so you could reliably tell whether a new patch worsened it or improved it.

Server software, for its part, didn't need music to function, didn't need cursors to point with, and it sure didn't care how fast windows were redrawn. These highly interactive activities had no place in server evaluations. It was typically all about *throughput*, and placing stress on some subsystem or another: disk, memory, or network, typically. Stress on the scheduler was a given. Even though dozens of benchmarks exist for measuring the throughput of high-end machines, producing megabytes or even gigabytes of analysis and data, there was no easy way to automate the type of subjective human observation that desktop users were using. There was no way to have weekly regression tests pick it up, nor any way to precisely duplicate the environment in which these observations were being made. In short, there was no way to quantify the observations being made, so no existing tests could detect regressions in this area.

Previous scheduler modifications had labeled applications that tended to spend a lot of time waiting for I/O as "interactive," and attempted to give scheduler bonuses to those tasks when the I/O they had been waiting for completed. This was *supposed* to provide the exact behavior the desktops were *not* seeing. The suspicion was that either these types of applications were

not being correctly recognized, or they were not being given sufficient bonuses.

3 Isolating the wiggle

The first part of the solution was recognizing that the "wiggle effect" comes from tasks not regaining the CPU fast enough. The second part was recognizing that the audible stutter from a music player, or the delay in redrawing a window, were showing the same problem as wiggling the cursor.

In the case of a cursor, coordinates from a serial mouse are presented as a stream of input to the windowing system. If the task that moves the cursor is not brought to a CPU quickly enough, there will be a lag between the time the movement is initiated and the time it appears on the screen. With all the input consumed, the task again goes to sleep even though a split second later more input appears as the mouse continues to move. While this is an efficient way to handle a serial mouse, it is dependent on hitting the processor quickly enough to guarantee the input stream doesn't back up too much. If the consuming task does not get to run quickly enough, the cursor will appear to move across the screen in a staccato fashion, even though the mouse itself is being moved smoothly.

In the case of a music player, the application (say, *xmms*) will read a certain amount of input from a file, but it will take longer to play it to the speaker. Even though this is, in general, a very I/O-intensive task, there are times when *xmms* will go to sleep either waiting for output to drain to the speaker or input to come from the file. Waking up too slowly from these self-imposed interruptions is what causes the music to pause or stutter.

Slow window redrawing is a case of applications taking too long after notification to wake up and redraw. This *might* also be attributed to slow interprocess communication or slow signal delivery, but it should be easy to rule out these causes if we were to measure the time a task spent in a queue waiting for a processor.

A patch for scheduler statistics has been available since 2.5.59[1]. However, it was with the 2.6.0-test5 release in September of 2003 that it was updated to include code to measure task latency. The task is given a new timestamp when it is placed in a run queue, placed on a processor, or removed from a processor. This makes it trivial to determine how long the task spent in the run queue before making it to the processor. It has the side effect of allowing us to also measure, on average, how long a task remains on the processor before relinquishing it, usually voluntarily. This allows us to easily characterize the kind of load a benchmark may place on a system.

Adding statistics counting to the scheduler path was a dicey task. This is one of the most heavily used paths in the system, and anything that slows down this path can have a catastrophic effect on the system as a whole. Consequently, the statistics patch tries to do what it can to gather accurate statistics without the use of a lock.

- Per-CPU counters are used, and incremented only by their respective CPU. This makes update collisions (and loss of data) impossible.
- Even so, when possible, these counters are incremented while a per-CPU runqueue lock is already acquired.
- Counters are only incremented, so minor variations from unflushed caches that may be observed while reading another CPU's counters can be safely ignored. (The

[1] http://oss.software.ibm.com/developerworks/opensource/linux/patches/?patch_id=730

counters are declared unsigned long, so user-level utilities on 32-bit architectures must take note that the counters could wrap. While theoretically possible on 64-bit machines, wrapping is far less likely than on 32-bit machines.)

Measurements were taken across several different releases using several different benchmarks to see if any statistical impact could be found on the benchmarks when scheduler statistics were utilized. To date, none have been found.

After the patch is applied, the counters can be obtained by reading `/proc/schedstat`. A full description of the statistics collected can be found in `Documentation/schedstats.txt` in the kernel source. The patch itself introduces a config option SCHEDSTATS that is on by default; if it is turned off, all the additional code is compiled out. There are three important fields:

timestamp *N*
: This line indicates a timestamp, in jiffies, of when this output was produced. The statistics are most effectively utilized when collected at small regular intervals, since this allows you to more accurately see how the behavior of a load or benchmark may change over its lifetime. Any process reading this file, however, is subject to the same scheduler delays it is trying to measure. Consequently, a simple script like

```
while true
do
    sleep 10
    cat /proc/schedstats >> \
        /tmp/stat.out
done
```

may find it collects statistics roughly every 10 seconds when the system is lightly loaded, but every 15-20 seconds or more when the system is heavily loaded. The code to note the timestamp is just a few lines before the data is totaled in the kernel, and on a non-preemptible kernel is an inexpensive way of identifying the time at which the snapshot was *actually* taken.

cpu*N n n n n n n n n ...*
: These are the values of the counters for cpu ***N***. The precise meaning of these counters will vary depending on the version of scheduler statistics being utilized. A few examples of data collected are:

1. number of times some functions were called
2. number of times certain functions were called under certain circumstances (i.e., were the runqueues unbalanced? was this processor idle?)
3. total number of milliseconds that tasks on this processor have used, not including the current one
4. total number of milliseconds that tasks that ran here had to wait in queue

version *N*
: identifies the version of output being produced. Since the meaning of fields (and the number of fields) in the **cpu*N*** line, above, can vary in different versions of scheduler statistics, this allows tools to be as flexible or inflexible as desired when processing input.

A sample of the output from `/proc/schedstat` is provided in Appendix A.

4 What would I use statistics for?

Scheduler statistics can serve three basic purposes. In many cases, they are doing no more than providing some detailed code path and profiling data. Knowing, for instance, that a particular function was called 50,000 times during a benchmark run may be key if it is expected to be called a dozen times—or a million. Similarly, knowing that 22,000 of those calls were made while the processor was idle, or made on just one of eight CPUs, may also be quite informative. About half the counters provide this sort of information, and it must be coupled with a knowledge of what to expect given your workload in order to detect anomalies.

Another purpose is to provide information beyond just counting. There is a counter that sums the imbalance found when queues are inspected. Combine this with the number of times you called this function and you can determine the average imbalance between runqueues. In most cases you wouldn't want this to exceed 1. Truth is, though, that a flurry of forking or even I/O completions might suddenly cause a processor to suddenly find itself with significantly more runnable tasks than other processors. Seeing where these spikes happen during the test run, and how often they happen, may help to suggest better "default" behavior in the scheduler or even tuning in the benchmark itself.

The last purpose has already been mentioned—task latency. We already need to note when a task is queued on a processor and when it acquires a processor. By noting one more thing—when it leaves the processor—we can also determine what I call the *runslice*.

The runslice is the amount of time a task spends *on* the processor before yielding it. In contrast, the *timeslice* alloted by the scheduler indicates how long the task may run before it is *forced* off. Processes are usually given generous timeslices (100 ms is the default) but typically don't use all of them at one shot. A task may need to put itself to sleep, perhaps to wait for input, before it has used up that full 100 ms. It will have any unused amount available to it when the event awakens it, but how long it spends on the processor can be an important characteristic of the system load. If a task spends only a few milliseconds before giving up the processor, it may be I/O-bound. By the same token, if it uses its full timeslice every time before being kicked off, then it is CPU-bound.

While many benchmarks are already characterized as CPU- or I/O-bound, they are rarely that way from beginning to end. Seeing this behavior graphed over a period of time can be very informative to a person trying to tune the system or the benchmark.

5 Diagnostic examples

The data that the scheduler statistics collect can be utilized in several different ways.

5.1 Using the function counts to characterize behavior

Recently a colleague remarked that he was running a benchmark that he expected to fully load a machine; yet profiling was reporting that the system was in the idle routine 50% of the time. He increased the load significantly on the machine and idle time only dropped to 49%. He couldn't believe the machine still had spare cycles, so we used the scheduler statistics to determine what was happening.

From the beginning of the benchmark, we captured the counters in `/proc/schedstat` every 10 seconds with a shell script. When the benchmark exited, we killed the shell script.

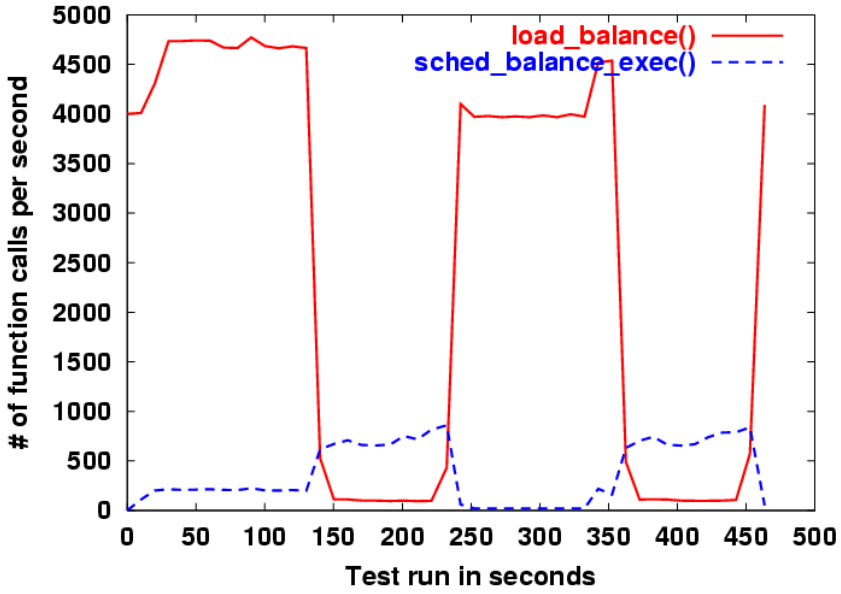


Figure 1: `load_balance()` and `sched_balance_exec()` counts

The two pieces of information that proved most useful were the number of calls per second (*cps*) for `load_balance()` and `sched_balance_exec()`. In Figure 1, you can see that the *cps* for `load_balance()` varies markedly between plateaus of around 4000-4500, and valleys of 100-200. When the system is idle, it calls `load_balance()` as often as once a millisecond to try to find work. When it is busy, it backs off to five times a second. The graph here is clearly indicating that this benchmark has at least two periods of about 100 seconds each out of about 450 seconds total where it is largely idle.

At about the same time that the *cps* for `load_balance()` is high, the *cps* for `sched_balance_exec()` is low. This function is called when tasks issue the `exec()` system call, and is used to do some opportunistic rebalancing. We observed that just as the system starts to get busy, `sched_balance_exec()` tails off.

The data suggested that this benchmark had a notable rampup and cooldown period. With this information in hand, simple observation of *top(1)* while running the benchmark confirmed what the scheduler statistics suggested. The benchmark had a fairly lengthy single-threaded setup: creating log files, making directories for results, and compiling short programs it would use. It then forked many tasks and set them all running to actually start the benchmark. When the test was over, there was again a single threaded task that collected the data created before several tasks organized the data.

5.2 Using latency and runslice information

In another situation, a disk-intensive benchmark was doing much worse with a different version of the scheduler. Figure 2 shows a measurement of the latency from the two runs.

In the “broken” run, the latencies were nearly twice that of the “working” run. Tasks were taking longer to reach the CPU in the broken case. Yet the runslice information shows comparable (and very short) times spent on the CPUs. If tasks were running very short periods of time, but waiting longer to run, what could

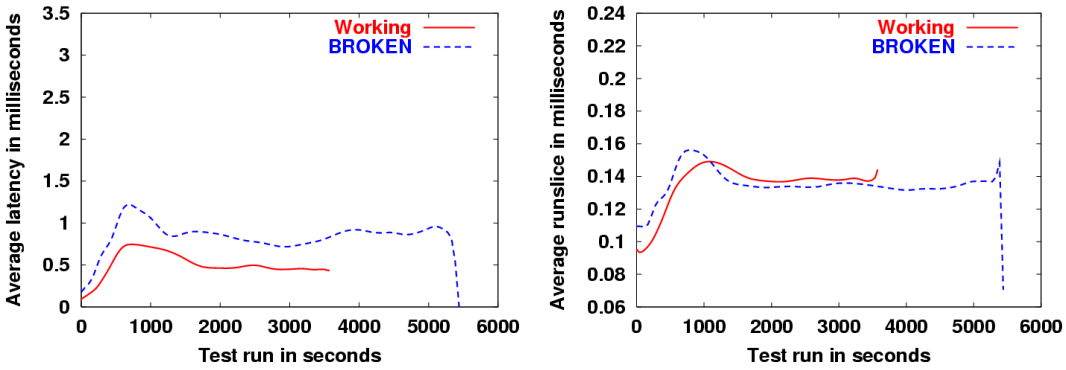


Figure 2: Latency and runslice duration

have been the cause?

Enlightenment was finally attained by viewing the average imbalance (Figure 3) during each of the runs. On the average, the imbalance was twice as great in the broken run as in the working run. Since the runslice was so small, this suggested that tasks were becoming runnable quickly but simply not being balanced often enough. Some queues were getting quite long while others (presumably) were staying short. Additional debugging showed that tasks were indeed awakening (probably by completed I/O) quite frequently but most of the balancing was happening only when one CPU fell idle and went looking for work. These longer queues in the broken run were persisting longer than those in the working run, and tasks stuck in them were waiting a fraction of a millisecond longer than before.

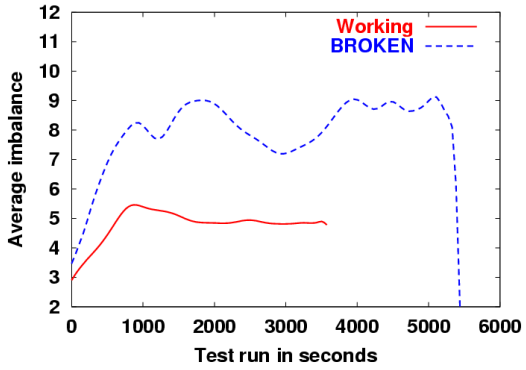


Figure 3: Average load imbalances

6 Conclusion

There is still work to do.

Recent scheduler changes present in Andrew Morton's -mm tree will dramatically change what is important to measure in the scheduler. Additionally, these same changes introduce some self-tuning characteristics which may benefit from statistics describing how often they are retuned.

There is also some evidence that NUMA machines may benefit from device, task, or memory affinitization strategies which try to keep data from crossing NUMA node boundaries. Scheduler statistics can be used to reliably demonstrate whether these strategies are being effective.

Lastly, the data provided by scheduler statistics probably ought to be moved out of /proc eventually, as there is an ongoing effort to return /proc to its original task of just listing processes.

Scheduler statistics provide a quantifiable means of measuring scheduler changes. Much as disk statistics can be used to a variety

of ends—measuring disk utilization, throughput rates, and transfer rates, for example—scheduler statistics can help with analysis of a variety of situations. The latest revisions go to lengths to avoid creating "Heisenbugs," or bugs which disappear when you try to examine them closely. Perhaps best of all, developers need not rely on mice and windowing systems to measure their test results. Latency numbers, in particular, provide a key way of measuring scheduler success, and runslice figures can help characterize the load that tests create so that the best set of tests can be chosen to test a particular feature or system. Cursor wiggles and audible skips can be set aside until they are needed again.

Disclaimer

This work represents the view of the author and does not necessarily represent the views of IBM.

IBM is registered trademark of International Business Machines Corporation in the United States and/or other countries worldwide.

Other company, product, and service names may be trademarks or service marks of others.

Appendix A

Table 1 is a sample of what `/proc/schedstat` might look like for a 2-proc machine. The actual format and number of counters will vary between different versions. For purposes of this example, the last three lines are artificially folded for readability, but in actual output, each would be one long line.

This is a brief description of each of the 23 counters for version 4 output. Applications can check the `version` field to make sure they look for and correctly interpret the counters. Note that all counters may wrap back to zero, and applications using these counters should be prepared to deal with that. Since all counters start at zero at boot time, the most useful way to use them is to get periodic snapshots of the counters, then subtract one set from a previously obtained one to obtain the delta. All counters are per-processor.

1. in `sched_yield()`, number of times both the active and the expired queue were empty
2. in `sched_yield()`, number of times just the active queue was empty
3. in `sched_yield()`, number of times just the expired queue was empty
4. in `sched_yield()`, number of times `sched_yield()` was called
5. in `schedule()`, number of times the active queue had at least one other task on it
6. in `schedule()`, number of times we switched to the expired queue and reused it
7. number of times `schedule()` was called
8. number of times `load_balance()` was called at an idle tick
9. number of times `load_balance()` was called at a busy tick
10. number of times `load_balance()` was called from `schedule()`
11. number of times `load_balance()` was called
12. sum of imbalances discovered (if any) with each call to `load_balance()`
13. number of times `load_balance()` was called when we did not find a "busiest" queue

```
version 4
timestamp 4295814751
cpu0 8909 9103 612 11869 264585 9821 392921 1065335 406 140662 1206403 62905
1192940 0 13440 13469 0 0 0 0 82278 1497607 264615
cpu1 5138 5328 577 8126 265205 6270 402877 943453 1005 149999 1094457 77670
1074828 0 13469 13440 0 0 0 0 200998 448842 265175
totals 14047 14431 1189 19995 529790 16091 795798 2008788 1411 290661 2300860
140575 2267768 0 26909 26909 0 0 0 0
```

Table 1: Sample output from `/proc/schedstat`

14. number of times `load_balance()` was called from `balance_node()`
15. number of times `pull_task()` moved a task to this cpu
16. number of times `pull_task()` stole a task from this cpu
17. number of times `pull_task()` moved a task to this cpu from another node (requires `CONFIG_NUMA`)
18. number of times `pull_task()` stole a task from this cpu for another node (requires `CONFIG_NUMA`)
19. number of times `balance_node()` was called
20. number of times `balance_node()` was called at an idle tick
21. sum of all time spent running by tasks (in ms)
22. sum of all time spent waiting by tasks (in ms)
23. number of tasks (not necessarily unique) given to the processor

The last three make it possible to find the average latency on a particular runqueue or, if taken from the `totals` fields, the overall system. Given two points in time, A and B, $(22B - 22A)/(23B - 23A)$ will give you the average time tasks had to wait after being scheduled to run but before actually running.

`/proc/<pid>/stat`

This version of the patch also changes the `stat` output of *individual tasks* to include the same latency and runslice information described above. Three new fields, corresponding to the last three fields described above, are added to the end of the per-task `stat` file, but apply only for that task rather than a whole processor.

On a Kernel Events Layer and User-space Message Bus System

Robert Love
Novell
rml@ximian.com

Abstract

Various Linux usage scenarios, particularly the widely accepted server and the rapidly growing desktop, require a lightweight, simple, asynchronous mechanism for kernel to user-space communication. Such a mechanism is crucial for the transmissions of events to user-space in a type-safe and clean manner. Further, a system-level messaging bus, which can deliver messages up the system stack on both a system-wide and per-user level, is required to further the integration of the Linux system.

This talk will discuss the design and implementation for two specific solutions, the Kernel Events Layer and D-BUS, to these two problems. Finally, useful solutions built on the sum of these technologies will be discussed—such as a fully integrated Linux desktop, from the kernel up through the GNOME desktop.

1 Introduction

Usually considered a plus of open source development, the Linux system is developed piece-meal, resulting in cleanly separated layers and properly defined interfaces. This separation, however, also results in a lack of integration among the various components comprising the system stack. In particular, the lack of integration is readily manifest between the lower levels of the stack—kernel and system-level components—and the upper levels of the system, such as the desktop environment on desktop machines.

A particularly important, but missing, component of the Linux system is an ubiquitous IPC mechanism and events system. Such a component would facilitate the dissemination of information up the system stack, better integrating the Linux system from the kernel up through the system layers, the desktop, and the end user applications and daemons. With well defined interfaces, such integration could occur while continuing the current separation and interoperability of Linux components.

What would such an IPC mechanism and event system allow? Quite a bit. Photo applications could start automatically in response to camera insertion. The volume of your music player could automatically lower in response to your phone ringing. System shutdown, reboot, and suspend messages could be trasmitted up the stack. HA applications could receive instant notifications from the kernel. No longer need components in the system live separate lives from the kernel, the layers below them, and themselves. Now, applications can communicate, listen, and evolve.

Such a system may be broken into three requirements:

- Kernel support implementing a kernel-to-user event mechanism

- A user-space message transport and IPC mechanism
- Applications sending and receiving such messages

This paper will discuss two specific implementations of these requirements:

- The Kernel Events Layer
- D-BUS

2 The Kernel Events Layer

2.1 Goals and Design

Current user-space grokking of the kernel typically requires some combination of periodic polling, parsing of unformatted text files from `/proc`, and luck. The Linux kernel currently lacks a mechanism for kernel to user-space communication.

The requirements for such a system include:

- simple and clean
- low overhead and scalable
- asynchronous transport accessible without polling
- type-safe
- generic enough for use in multiple usage scenarios
- support for formalized sender interfaces, allowing standardized messaging

Event systems have been proposed and even implemented, but they generally receive minimal community buyin, presumably due to a lack of one or moe of these requirements (more than likely, the “simple” bit).

2.2 Implementation

The Kernel Events Layer implements an event system satisfying these requirements.

Usage is simple:

```
send_event(int type, char
*interface, char *fmt, ...)
```

The `type` parameter specifies a constant value representing the type of message being sent. The `interface` value specifies the originator of the message. It is used to provide an interface object for object-based component and IPC systems such as CORBA and D-BUS. Finally, `fmt` and any following arguments provide the usual `va_list` of format and arguments.

Example:

```
send_event (DBUS_NORMAL,
"org.kernel.arch.cpu",
"overheating")
```

This specifies a message from the `org.kernel.arch.cpu` interface with a value of `overheating`.

The actual implementation of the Kernel Events Layer uses netlink. In fact, the Kernel Event Layer is simply specific netlink socket into user-space in which the event is formated and then reconstructed by user-space. Netlink is fast, simple, and already in the kernel. Thus it was a natural choice.

The Kernel Events Layer code uses `netlink_broadcast()` internally.

2.3 Real World Usage

The Kernel Events Layer is independent of any specific user-space transport mecha-

nism. The assumed use case is to create a new daemon (or modify an existing daemon, like the D-BUS system message bus, `dbus-system-1`). This daemon listens on the netlink socket, reading each event as it occurs. The events are parsed and reconstructed into the format native to the user-space transport mechanism.

In the case of D-BUS, the `dbus-system-1` daemon sends the kernel events out the system message bus. Components up the system stack may then receive the kernel events right off the D-BUS system bus, along with other system-wide messages.

3 D-BUS

D-BUS is a user-space IPC system.

D-BUS varies from other IPC mechanisms in that it provides a bus system (as opposed to point-to-point) over which messages (as opposed to byte streams) are transported. Messages include a header containing metadata about the message itself and a body containing the data. The bus system is created by forming a point-to-point connection between the D-BUS daemon and each listener. The daemon acts as the hub and the listeners as the spokes of a wheel.

D-BUS provides both a system-wide and a per-user session bus. The system-wide bus is used to dissemenate information on a machine-global scale. A single system daemon provides this service, allowing applications up the stack to receive messages from components down the stack. A security system implements access control.

The per-user session bus exists on a per-user basis, with one daemon created for each user session. The per-user daemon is used for general application IPC and is physically separate from the system-wide bus. The per-user daemon is generally used for traditional point-to-point IPC.

D-BUS is the name given to this system. It is composed of several architectural layers:

- The message bus daemon
- The D-BUS library, `libdbus`, which connects to applications together
- Wrapper libraries and bindings that wrap `libdbus` for direct use on various application frameworks, such as Glib or QT, and various languages, such as C# and Python. The wrapper libraries and bindings provide the API that most programmers should use as they both simplify the rather low-level `libdbus` API and provide an API more familiar and fit for that particular environment.

3.1 D-BUS Concepts

D-BUS introduces various concepts that comprise the IPC system.

- The **bus** is either the system-wide global bus or the per-user session bus.
- **Objects** represent an instance of a specific listener of a D-BUS message. Objects are contained within the applications that use D-BUS, and generally map to objects in object-oriented languages. Because D-BUS would not find using a pointer or reference to identify an object very friendly, it introduces a name for each object. The name resembles a UNIX filesystem path, such as /org/kernel/fs/filesystem.
- **Interfaces** represent methods or signals implemented on an object. Each object supports at least one interface.

- **Messages** are sent to and from a defined method or signal. D-BUS supports multiple message types: method invocation, method return, error message, and signal.

3.2 Use of D-BUS

D-BUS's simplicity, performance, and use of the message and bus paradigm set it up for use across the entire Linux system and make it a perfect replacement for CORBA, DCOP, and other IPC mechanisms.

Multiple projects are taking advantage of D-BUS. They include:

- Project Utopia uses D-BUS as the IPC mechanism to link the kernel, udev, HAL, and the GNOME desktop.
- A CUPS patch uses D-BUS to transmit information about the printer spool.
- Jamboree uses D-BUS to automatically mute the volume.
- A Gconf patch uses D-BUS as the Gconf transport mechanism.

4 The Kernel Events Layer, D-BUS, and Project Utopia

D-BUS is used as the backbone of Project Utopia, an umbrella project aiming to bring improved hardware management and system integration to the Linux system and GNOME desktop. Project Utopia uses D-BUS to link the kernel, up through hotplug, udev and HAL to the rest of the system. Libraries utilizing D-BUS and built on top of HAL provide enhanced hardware support. Applications at the desktop level can then reap the benefits.

4.1 Example: libinput

`libinput` is a simple library for managing input devices that sits on top of HAL and communicates to HAL beneath it and the applications above it via D-BUS. `libinput` is used to enumerate all input devices on the system. `libinput` also provides an interface for applications to register callbacks, and integrate these callbacks into its mainloop. The callbacks are invoked when input devices are added to or removed from the system.

Sample usage of enumerating all input devices on the system:

```
struct input *devices;

if (input_init ())
        /* error ... */

devices = input_devices_get ();
while (devices) {
        /* ... */
        devices = devices->next;
}
input_devices_put (devices);
```

Given a specific `struct input`, the library provides wrappers for opening and closing the device via `open`(2) and `close`(2). This is not strictly required, but furthers the abstracting of device nodes not only from the user but even from the application.

Example:

```
fd = input_device_open (device, 0);

/* ... */

input_device_close (device);
```

Registering of the callbacks is also easy:

```
void my_mainloop
```

```
(DBusConnection *dbus_connection)
{
  dbus_connection_setup_with_g_main
  (dbus_connection, NULL);
}

void my_added
(struct input *device)
{
  printf
  ("%s was just "
          "hotplugged!\n",
          device->product);
}

void my_removed
(struct input *device)
{
  printf
  ("%s was just "
         "hotunplugged!\n",
         device->product);
}

  /* ... */
  input_init_with_callbacks
        (&my_mainloop,
         &my_added,
         &my_removed);

  gtk_main ();
```

When an input device is added or removed from the system, `my_added` and `my_removed` are invoked as appropriate.

The goals behind such a library are twofold:

- Abstract away concepts of device nodes and low-level system-specific behavior and allow application developers to search for enumerate the devices on a system through simple interfaces.

- Allow asynchronous poll-free hack-free callbacks into the application to notify the program of changes in events, such as a new joystick on the system.

5 Conclusion

The Kernel Events Layer and D-BUS are two crucial components in better unifying and integrating the Linux system. They provide the infrastructure required for a future rich with information exchange. Where all levels of the desktop can communicate—talking, listening, evolving.

Linux-tiny And Directions For Small Systems

Matt Mackall
Digeo, Inc.
mpm@digeo.com

Abstract

Linux-tiny is a project to reduce the memory and storage footprint of the 2.6 Linux kernel for embedded, handheld, legacy, and other small systems. I describe strategies for kernel size reduction, some of the major areas already investigated and the results achieved, as well as some avenues for further exploration.

1 Introduction

Historically, Linux had a reputation for running on very modest systems. My first dedicated Linux box, running a 0.99 kernel circa 1994, provided mail, FTP, web, dial-in, and shell services on a 16MHz 386SX with a mere 4 megabytes of RAM. In the 10 years since then, Linux has grown to the point where it runs on machines with over a thousand processors and a terabyte of RAM. Not surprisingly, a modern Linux distribution can have difficulty getting to a shell prompt on machines with less than 8 megabytes of RAM, let alone doing useful work.

1.1 What happened?

In the time between the 0.99 and 2.6 kernels, we've seen Linux become a serious commercial endeavor, we've seen kernel hackers get jobs (and get big machines on their desks), and we've seen a massive boom in Internet use and personal computing. Linux developers have been targeting high end computing and rising demand for hardware has seen prices drop tremendously.

But there are still small machines! Hand-helds and embedded systems are perennially pressed for space to match their desktop counterparts and many people throughout the world still rely on legacy machines to get their work done. What can be done to recapture the 'small is beautiful' utility of those early systems?

1.2 Where is the growth?

The process by which any large software project grows can aptly be described as *death by a thousand cuts*. The accumulation of bloat occurs change by change and creeps in from several different directions.

Perhaps the most visible is the addition of new **features**, which generally requires the introduction of wholly-new code. Frequently features are considered so small or so essential that no thought is given to making them optional. As the median system size grows, this new code tends to be more verbose and less concerned with space issues.

The next, more subtle culprit is **performance**. Given the fundamental importance of kernel performance to overall system performance, trade-offs of size for speed are easy to justify. Unfortunately the accumulation of many such trade-offs can leave us with a system that no longer boots. Ironically, the evolution of pro-

cessors has brought us to a point where cache footprint can be critical to performance so a lot of the choices that have been made in this area bear rethinking.

Next we have **compatibility** and **correctness**. Every time the system is extended to better support a slightly different piece of hardware or work around another corner case, more code is added. Occassionally cleanups and unifications make some of this code redundant, but this is the exception. A related phenomenon is the evolution of the kernel APIs and the accumulation of obsolete code for the sake of backward compatibility.

2 Linux-Tiny for the small system niche

There have been numerous efforts to address the above phenomena for various components of Linux systems, but most of the attention has been addressed at userspace (arguably the biggest offender). Experiments with pre-2.6.0 kernels however suggested it was time to pay some more attention to the kernel itself. So in December of 2003, I decided to create a new 2.6-based tree dedicated to small systems which I named Linux-Tiny [3] (someone had already borrowed my initials for their tree).

2.1 Methodology

With stated targets of embedded, hand-held, and legacy machines, the -tiny tree attempts to tailor the kernel to the needs of small systems. The tree is maintained as a series of small patches stacked on top of mainline kernel releases, managed with the quilt tool [1] (previously with Andrew Morton's patch scripts [4]).

Patches try to observe the following criteria:

- **configurable**: changes that are not clearly wins for all systems should be configurable so that users can make their own trade-offs
- **non-invasive**: patches should be small, self-contained, and largely independent so that integrators can cherrypick the patches they'd like to use
- **mergeable**: while not mandatory, patches should try to be acceptable to the mainline kernel in both style and approach; merging to mainline is a priority

In addition to patches focusing on reducing kernel footprint, I've also added a number of patches to do debugging and auditing including netconsole, kgdb, and kgdb-over-ethernet support.

2.2 Setting goals

Everyone has a different set of functionality requirements in mind for small systems. The features needed on a handheld are very different from those needed for a network appliance or a kiosk. Thus, choosing a subset of features to develop towards is tricky.

The approach I've taken is to choose a series of targets to optimize, and the first is a minimal x86 kernel with filesystem, console, and TCP/IP support. How small can we make this kernel? This puts a focus on the most of the common core functionality of Linux and provides a useful benchmark for progress.

3 Finding bloat

As mentioned above, there are many sources of bloat. There are also several forms it can take: as superfluous code, statically or dynamically allocated data, inline functions or macros, compiler mis-optimizations, or cut-n-paste coding.

Given that the kernel is on the order of several hundred kilobytes, tackling bloat is going to be a matter of trimming several kilobytes here and a couple kilobytes there. While one could simply pick any source file and read through it searching for cleanup opportunities, there are some more straightforward ways of finding the "low-hanging fruit".

3.1 Using nm(1) and size(1)

The easiest place to begin is by using the `nm` tool to find large functions and data structures. Comparing the (hexadecimal) numbers from `nm(1)` with `size(1)` gives us a good start at understanding the relative sizes of some of the major subsystems and their components compared to the kernel as a whole. For instance, we can see by comparing Table 1 and Table 2 that the static `ide_hwifs` data structure alone takes 15360 bytes, over 2% of the data portion of the default kernel.

3.2 Measuring function inlining

Function inlining and macro expansion present a special problem for our bloat detection efforts. In the early 1990s, inlining was a very popular performance technique to avoid function call branches. A great number of key functions are marked for inlining in the kernel and their usage and size impact is obscured because they become a seamless part of the functions that use them. Auditing their usage becomes a matter of convincing the compiler to tell us when inlines are being instantiated in a build and then estimating how large these functions are when expanded inline.

Rather than modifying the compiler itself, the first part of this puzzle was hacked around by redefining `inline` to include the GCC extension `__attribute__((deprecated))`. This causes a very useful warning like the following to be generated:

```
arch/i386/kernel/semaphore.c:58:
warning: 'get_current' is
deprecated (declared at
include/asm/current.h:16)
```

By post-processing these voluminous warning messages, we can determine which inline functions are instantiated directly in C files as well as which are called as parts of other inlines and finally calculate the total number of direct or indirect instantiations of each (see Table 3).

The second part of this puzzle was more challenging. While we know in which modules and how often inlines are instantiated, we cannot yet calculate their sizes. I made several attempts to generate approximate size data by looking at GCC's symbolic debugging output, but this tended to be easily confused by inlining and was too inaccurate for use.

Recently Denis Vlasenko took another stab at this and wrote a set of scripts called inline_hunter [5] to generate a set of dummy functions wrapping single calls to inlines. While these sizes won't directly reflect the size of inline instantiations due to function call overhead and lost optimization opportunities, for larger inline functions, it has proven fairly representative. Some of the larger inlines found with this approach are shown in Table 4.

3.3 Tracking dynamic allocations

Of course much of the kernel's memory footprint is from dynamic allocations. Memory used for page tables, tracking running processes, indexing hashes and so forth is allocated at runtime and can vary with the size of the load. A number of these are hash tables to increase look-up performance, which for small systems can be less important than simply fitting in memory.

There are several important allocators in the kernel. First, the bootmem allocator which

```
2.6.5\$ nm --size -r vmlinux | head -20
00008000 b __log_buf
00007000 D irq_desc
00004e78 d pci_vendor_list
00004000 b bh_wait_queue_heads
00003c00 B ide_hwifs
0000213a T vt_ioctl
00002000 D init_thread_union
00001880 D contig_page_data
0000163b T journal_commit_transaction
00001500 b irq_2_pin
000012f5 T tcp_sendmsg
00001162 t n_tty_receive_buf
00001080 d per_cpu__tvec_bases
00001000 t translation_table
00001000 b sd_index_bits
00001000 D init_tss
00001000 b doublefault_stack
00001000 B con_buf
00001000 b cache_defer_hash
00000fe0 T cdrom_ioctl
```

Table 1: nm output for 2.6.5 default config

handles a number of critical allocations at startup. As there are not terribly many of these, they can be audited very simply with `printk()` techniques.

Second, the SLAB allocator is used to quickly allocate sets of objects of the same size and type. The kernel provides a way to track these allocations with `/proc/slabinfo`.

The more general `kmalloc()` allocator has been rebuilt on top of the aforementioned SLAB allocator, translating kmalloc requests into requests from a set of ascending generic SLAB sizes. Thus all `kmalloc()` allocations are lumped together by size in the `/proc/slabinfo` output. That can be helpful if you know what you're looking for, but doesn't give many hints as to which parts of the kernel are using that memory.

To address this deficiency, I've created a small footprint tool for tracking allocations via `/proc/kmalloc` (see Table 5). This works by tracking the address of each allocation along with the address of the allocating function in a simple hash table. Also tracked are net and gross allocation sizes and counts per caller. When a `kfree()` call is made, it is matched up to its caller for accounting purposes and removed from the hash. Thus it is possible not only to determine how much dynamic memory is used by each function but also to easily identify memory leaks.

4 Some notable opportunities for code trimming

The above methods have revealed numerous opportunities for cutting back the kernel's

```
2.6.5\$ size vmlinux */built-in.o
   text    data     bss     dec     hex filename
3366220  673296  166824 4206340  402f04 vmlinux
1181276  250808   48000 1480084  169594 drivers/built-in.o
 735152   32593   30628  798373   c2ea5 fs/built-in.o
  18151    1120    1316   20587    506b init/built-in.o
  21841     172     204   22217    56c9 ipc/built-in.o
 159632   16115   42402  218149   35425 kernel/built-in.o
   2870       0       0    2870     b36 lib/built-in.o
 129669    9068    2884  141621   22935 mm/built-in.o
 580407   33816   18856  633079   9a8f7 net/built-in.o
   1869       0       0    1869     74d security/built-in.o
 325923   11114    3016  340053   53055 sound/built-in.o
    134       0       0     134      86 usr/built-in.o
```

Table 2: size output for 2.6.5 default config

memory footprint, many of which remain to be examined. What follows are some of the more notable areas that have been explored.

4.1 Debugging data

The kernel has numerous facilities for trapping and reporting problem conditions and other status information, including `printk()`, `bug()`, `warn()`, `panic()`, and friends. In ideal circumstances, these facilities go unexercised. And in the extreme, embedded boxes may have no means of reporting this data, due to lack of a display, writable storage, or the like. Unfortunately, not only do these facilities use a substantial amount of code, their users need extra space for error message strings, filenames, and line numbers.

Linux-tiny has a set of configuration options to compile out most of this code and remove the debugging strings and data from the kernel. Disabling support for `printk()` saves well over 100K. Independent options control the inclusion of the `bug()` infrastructure and support for trapping panics and doublefaults.

4.2 Optional interfaces

For systems with well-defined application requirements, many of the kernel's APIs are unnecessary. Cutting-edge, obsolete, or obscure features are obvious candidates for configurable removal.

- **sysfs**: The new sysfs filesystem makes substantial memory demands (which can be more than half a megabyte even on the smallest systems) but its features may well not be essential to current systems. The -tiny tree was a testbed for options to entirely remove sysfs or to use a lighter "backing store" version.
- **ptrace, aio, posix-timers**: These features are among those that are only used by a small set of applications. These and other Linux-tiny options are enabled under the CONFIG_EMBEDDED menu, which marks them as making the kernel non-standard.
- **uid16, vm86**: Some of the many legacy interfaces in the kernel. Modern applications and libraries use 32-bit user and

group IDs and vm86 support is used to run 16-bit code for emulators like DOSEMU and Wine and for some video drivers used by X.

- **ethtool, tcpdiag, igmp, rtnetlink**: One of the most complicated parts of the kernel is the networking layer, which has grown a variety of APIs to gain access to its many features. But for most users, the interfaces used by the classic `ifconfig(8)` and `route(8)` tools are sufficient.

4.3 4K stacks

During the 2.1 kernel series (circa 1998), the x86 kernel increased the size of the per-task kernel stacks from 4K to 8K to work around issues with stack depth. In addition to the obvious increase in overhead for every userspace process, several new kernel daemons have been added, all with their own stacks. Another issue is that finding pairs of contiguous pages to build an 8K stack can be very difficult on a machine with memory pressure and especially so on machines with a small number of total pages.

Many of the problems that made 4K stacks problematic have since been addressed and 4K stacks are now practical for most applications. Linux-tiny has served as an early testbed for reintroducing 4K stack support to the mainline 2.6 kernel and includes a developer tool called `checkstack` that will automatically disassemble a kernel to find the most extreme stack space users.

4.4 The SLOB allocator

Most memory in the kernel is managed either directly or indirectly through the SLAB allocator. SLAB maintains separate caches for objects of given sizes and types and can very quickly manage allocations for them. In some cases, it can even arrange for objects to be pre-initialized without any additional overhead. SLAB also has some resistance to troublesome memory fragmentation issues. While simple in principle, the SLAB code ends up being quite complex from its efforts to squeeze the maximum possible performance out of the allocator.

The primary downside to SLAB is that because it maintains a collection of independent caches which are all one or more pages, it ends up leaving quite a bit of unused space in each SLAB cache. In addition, as `kmalloc` is implemented on top of SLAB using a set of preset object size SLABs, there is quite a bit of extra space allocated for the average `kmalloc` call. Measurements with the previously described `/proc/kmalloc` tool report that extra overhead can amount to 25-30% of the total memory allocated by `kmalloc`.

Linux-tiny provides an optional replacement for SLAB that I've dubbed *SLOB* (simple list of blocks). SLOB trades performance for space efficiency by implementing a more traditional list-based allocator that also understands requests for objects with particular alignments. The APIs used by SLAB and `kmalloc()` are provided by a small emulation layer.

SLOB manages all objects at a granularity of 8 bytes so overhead for odd object sizes is minimized. It also does away with the numerous partly-used caches of the SLOB approach. Finally, the SLOB code is much simpler and takes up less than one tenth of the space of the standard SLAB allocator.

4.5 TinyVT

As you can see from Table 1, the largest single function in the default kernel is `vt_ioctl()`, which manages many of the special features of the Linux console. As most early Linux

users didn't have the memory for running a full-fledged X desktop, the native Linux text console is very powerful, with support for scrollback, selection, virtual console switching, Unicode translation and character sets, screen blanking, and so on.

These features can be very handy for some users, but on a palmtop or kiosk running a GUI, or for a minimal rescue disk, they're dead weight. Linux-tiny includes a heavily trimmed down replacement for the standard console code which drops many of these features and can trim a couple percent off the size of the kernel image.

5 Results

Recent releases of Linux-tiny contain the above options and numerous others. My test configuration, with support for a text console, IDE disks, the Ext2 filesystem, TCP/IP, and a PCI-based network card results in a 363K compressed kernel image. Other users of Linux-tiny have reported kernel configurations resulting in images as small as 191K.

Booting the test configuration with `mem=2M`, which gives a total of of 1664K after accounting for BIOS memory holes, still leaves ample room for a lightweight userspace (see Table 6). A similarly configured mainline kernel without the -tiny patches compiles to a kernel image of over 500K and has difficulty booting with `mem=4M`.

For comparison, the earliest Linux distribution kernel I've been able to locate, a 0.99pl15 kernel from Slackware 1.1.2 circa 1994, is a mere 301K. Modern `highly-modularized` 2.6 kernels from Fedora Core 2 and SuSE 9.1 weigh in at 1.2M and 1.5M respectively while the default 2.6.5 kernel config builds a 1.9M compressed kernel.

6 Further directions

There are many further avenues to pursue and subsystems to trim. Some of the more aggressive ideas on the to-do list include:

- A lightweight replacement network stack: Minimal TCP stacks like uIP [2] have sufficient functionality for simple network applications and have extremely small footprints.
- Replacements for fixed-sized hash tables: Existing kernel hash tables have difficulty scaling with workloads and memory sizes. Other approaches like radix trees might be better in some areas and avoid wasted memory when the indexes are empty.
- Support for bunzip2: Linux-tiny now has a simplified interface to the boot-time decompressor and allows for replacements to be easily dropped in. While bzip2 compression won't save any memory at runtime, it will save valuable storage space on embedded systems.
- Pageable kernel memory: Following an approach similar to the `__init` approach in current kernels, it should be possible to mark specific functions and data in the kernel core as pageable, provided they meet some specific requirements.
- Tracking kernel growth: Using automated tools to track the size of kernel functions and subsystems from release to release will help catch new bloat when it appears.

Of course, as most of the bloat in the kernel has been introduced in small increments, most of the improvements will be of the same variety. Contributions are encouraged!

References

[1] patchwork quilt patch management tools. `http://savannah.nongnu.org/projects/quilt`.

[2] Adam Dunkels. The uip tcp/ip stack for embedded microcontrollers. `http://www.sics.se/~adam/uip/index.html`.

[3] Matt Mackall. The linux-tiny homepage. `http://www.selenic.com/tiny-about/`.

[4] Andrew Morton. Patch-scripts. `http://www.zip.com.au/~akpm/linux/patches/`.

[5] Denis Vlasenko. inline_hunter 2.0 and its results, 2004. `http://lkml.org/lkml/2004/4/16/191`.

```
1560  get_current (1294 in *.c)
calls:
callers: <other>(336) capable(122) unlock_kernel(44) lock_kernel(33)
flush_tlb_page(11) flush_tlb_mm(10) find_process_by_pid(6)
flush_tlb_range(4) current_is_kswapd(4) current_is_pdflush(3)
rwsem_down_failed_common(2) on_sig_stack(2) do_mmap2(2) __exit_mm(2)
walk_init_root(1) scm_check_creds(1) save_i387_fsave(1)
sas_ss_flags(1) restore_i387_fsave(1) read_zero_pagealigned(1)
handle_group_stop(1) get_close_on_exec(1) fork_traceflag(1)
ext2_init_acl(1) exec_permission_lite(1) dup_mmap(1) do_tty_write(1)
de_thread(1) copy_signal(1) copy_sighand(1) copy_fs(1) check_sticky(1)
cap_set_all(1) cap_emulate_setxuid(1) arch_get_unmapped_area(1)

546  current_thread_info (286 in *.c)
calls:
callers: <other>(207) copy_to_user(95) copy_from_user(86)
tcp_set_state(22) test_thread_flag(20) verify_area(13)
tcp_enter_memory_pressure(6) sock_orphan(3) icmp_xmit_lock(2)
csum_and_copy_to_user(2) tcp_v4_lookup(1) sock_graft(1)
set_thread_flag(1) neigh_update_hhs(1) ip_finish_output2(1) gfp_any(1)
fn_flush_list(1) do_getname(1) clear_thread_flag(1) alloc_buf(1)
activate_task(1)

413  atomic_dec_and_test (55 in *.c)
calls:
callers: put_page(103) kfree_skb(101) <other>(47) mntput(34)
in_dev_put(23) neigh_release(19) tcp_tw_put(18) fib_info_put(17)
sock_put(15) put_namespace(6) mmdrop(6) __put_fs_struct(4)
tcp_listen_unlock(3) ipq_put(3) finish_task_switch(2) __detach_pid(2)
task_state(1) de_thread(1)

255  tcp_sk (134 in *.c)
calls:
callers: <other>(117) tcp_reset_xmit_timer(30) tcp_set_state(22)
tcp_current_mss(13) tcp_initialize_rcv_mss(6) tcp_free_skb(6)
tcp_check_space(6) tcp_data_snd_check(5) tcp_clear_xmit_timer(5)
tcp_synq_removed(3) tcp_select_window(3) westwood_update_rttmin(2)
westwood_acked(2) tcp_synq_len(2) tcp_synq_drop(2)
tcp_ack_snd_check(2) __tcp_inherit_port(2) tcp_use_frto(1)
tcp_synq_young(1) tcp_synq_is_full(1) tcp_synq_added(1)
tcp_prequeue(1) tcp_listen_poll(1) tcp_event_ack_sent(1)
tcp_connect_init(1) tcp_acceptq_queue(1) do_pmtu_discovery(1)
```

Table 3: Some large inline counts and users for 2.6.5-tiny1

Size	Uses	Wasted	Name and definition	
56	461	16560	copy_from_user	include/asm/uaccess.h
122	119	12036	skb_dequeue	include/linux/skbuff.h
164	78	11088	skb_queue_purge	include/linux/skbuff.h
97	141	10780	netif_wake_queue	include/linux/netdevice.h
43	468	10741	copy_to_user	include/asm/uaccess.h
43	461	10580	copy_from_user	include/asm/uaccess.h
145	77	9500	put_page	include/linux/mm.h
49	313	9048	skb_put	include/linux/skbuff.h
109	101	8900	skb_queue_tail	include/linux/skbuff.h
381	21	7220	sock_queue_rcv_skb	include/net/sock.h
55	191	6650	init_MUTEX	include/asm/semaphore.h
61	163	6642	unlock_kernel	include/linux/smp_lock.h
59	165	6396	lock_kernel	include/linux/smp_lock.h
127	59	6206	dev_kfree_skb_any	include/linux/netdevice.h
41	289	6048	list_del	include/linux/list.h
73	83	4346	dev_kfree_skb_irq	include/linux/netdevice.h
131	39	4218	netif_device_attach	include/linux/netdevice.h
110	44	3870	skb_queue_head	include/linux/skbuff.h
84	59	3712	seq_puts	include/linux/seq_file.h
57	75	2738	skb_trim	include/linux/skbuff.h
45	96	2375	skb_queue_head_init	include/linux/skbuff.h
41	111	2310	list_del_init	include/linux/list.h
102	23	1804	__nlmsg_put	include/linux/netlink.h

Table 4: Size estimates found by inline_hunter

```
# cat /proc/kmalloc
total bytes allocated:   266848
slack bytes allocated:    37774
net bytes allocated:     145568
number of allocs:           732
number of frees:            282
number of callers:           71
lost callers:                 0
lost allocs:                  0
unknown frees:                0

   total    slack      net alloc/free  caller
     256      203      256     8/0     alloc_vfsmnt+0x73
    8192     3648     4096     2/1     atkbd_connect+0x1b
     192       48       64     3/2     seq_open+0x10
   12288        0     4096     3/2     seq_read+0x53
    8192        0        0     2/2     alloc_skb+0x3b
     960        0        0    10/10    load_elf_interp+0xa1
    1920      288        0    10/10    load_elf_binary+0x100
     320      130        0    10/10    load_elf_binary+0x1d8
     192       48       96     6/3     request_irq+0x22
    7200     1254     7200    75/0     proc_create+0x74
      64       43       64     2/0     proc_symlink+0x40
    4096      984        0     1/1     check_partition+0x1b
   69632        0    45056    17/6     dup_task_struct+0x38
     128       48      128     2/0     netlink_create+0x84
     128       20      128     1/0     ext2_fill_super+0x2f
      32       28       32     1/0     ext2_fill_super+0x385
      32       31       32     1/0     ext2_fill_super+0x3b6
     608       76      384    19/7     __request_region+0x18
      64       32       64     2/0     rand_initialize_disk+0xd
    8192     2016     8192     2/0     alloc_tty_struct+0x10
     128       56      128     2/0     init_dev+0xba
     128       56      128     2/0     init_dev+0xf3
     128        0      128     2/0     create_workqueue+0x28
    8960     1680     8960    70/0     tty_add_class_device+0x20
    2048      960     2048     4/0     alloc_tty_driver+0x10
    9280     2332     9280     4/0     tty_register_driver+0x2d
     288        0      288     9/0     mempool_create+0x16
    1280      196     1280     9/0     mempool_create+0x41
    1536      384     1536     8/0     mempool_create+0x8f
      64       28       64     1/0     kbd_connect+0x3e
     928      348        0    29/29    kmem_cache_create+0x235
   28288     1448    28288    81/0     do_tune_cpucache+0x2c
...
```

Table 5: Tracking usage of kmalloc/kfree in -tiny

```
Uncompressing Linux... Ok, booting the kernel.
# mount /proc
# cat /proc/meminfo
MemTotal:          980 kB
MemFree:           312 kB
Buffers:            32 kB
Cached:            296 kB
SwapCached:          0 kB
Active:            400 kB
Inactive:           48 kB
HighTotal:           0 kB
HighFree:            0 kB
LowTotal:          980 kB
LowFree:           312 kB
SwapTotal:           0 kB
SwapFree:            0 kB
Dirty:               0 kB
Writeback:           0 kB
Mapped:            380 kB
Slab:                0 kB
Committed_AS:      132 kB
PageTables:         24 kB
VmallocTotal:  1032172 kB
VmallocUsed:         0 kB
VmallocChunk:  1032172 kB
#
```

Table 6: Boot log for a 2.6.5-tiny1 test configuration with mem=2m

Xen and the Art of Open Source Virtualization

Keir Fraser, Steven Hand, Christian Limpach, Ian Pratt
University of Cambridge Computer Laboratory
{first.last}@cl.cam.ac.uk

Dan Magenheimer
Hewlett-Packard Laboratories
{first.last}@hp.com

Abstract

Virtual machine (VM) technology has been around for 40 years and has been experiencing a resurgence with commodity machines. VMs have been shown to improve system and network flexibility, availability, and security in a variety of novel ways. This paper introduces Xen, an efficient secure open source VM monitor, to the Linux community.

Key features of Xen are:

1. supports different OSes (e.g. Linux 2.4, 2.6, NetBSD, FreeBSD, etc.)
2. provides secure protection between VMs
3. allows flexible partitioning of resources between VMs (CPU, memory, network bandwidth, disk space, and bandwidth)
4. very low overhead, even for demanding server applications
5. support for seamless, low-latency migration of running VMs within a cluster

We discuss the interface that Xen/x86 exports to guest operating systems, and the kernel changes that were required to Linux to port it to Xen. We compare Xen/Linux to User Mode Linux as well as existing commercial VM products.

1 Introduction

Modern computers are sufficiently powerful to use virtualization to present the illusion of many smaller virtual machines (VMs), each running a separate operating system instance. This has led to a resurgence of interest in VM technology. In this paper we present Xen, a high performance resource-managed virtual machine monitor (VMM) which enables applications such as server consolidation, co-located hosting facilities, distributed web services, secure computing platforms, and application mobility.

Successful partitioning of a machine to support the concurrent execution of multiple operating systems poses several challenges. Firstly, virtual machines must be isolated from one another: it is not acceptable for the execution of one to adversely affect the performance of another. This is particularly true when virtual machines are owned by mutually untrusting users. Secondly, it is necessary to support a variety of different operating systems to accommodate the heterogeneity of popular applications. Thirdly, the performance overhead introduced by virtualization should be small.

Xen hosts commodity operating systems, albeit with some source modifications. The prototype described and evaluated in this paper can support multiple concurrent instances of our XenLinux guest operating system; each instance exports an application binary interface identical to a non-virtualized Linux 2.6. Xen ports of NetBSD and FreeBSD have been completed, along with a proof of concept port of Windows XP.[1]

There are a number of ways to build a system to host multiple applications and servers on a shared machine. Perhaps the simplest is to deploy one or more hosts running a standard operating system such as Linux or Windows, and then to allow users to install files and start processes—protection between applications being provided by conventional OS techniques. Experience shows that system administration can quickly become a time-consuming task due to complex configuration interactions between supposedly disjoint applications.

More importantly, such systems do not adequately support performance isolation; the scheduling priority, memory demand, network traffic and disk accesses of one process impact the performance of others. This may be acceptable when there is adequate provisioning and a closed user group (such as in the case of computational grids, or the experimental PlanetLab platform [11]), but not when resources are oversubscribed, or users uncooperative.

One way to address this problem is to retrofit support for performance isolation to the operating system, but a difficulty with such approaches is ensuring that *all* resource usage is accounted to the correct process—consider, for example, the complex interactions between applications due to buffer cache or page replacement algorithms. Performing multiplexing at a low level can mitigate this problem; unintentional or undesired interactions between tasks are minimized. Xen multiplexes physical resources at the granularity of an entire operating system and is able to provide performance isolation between them. This allows a range of guest operating systems to gracefully coexist rather than mandating a specific application binary interface. There is a price to pay for this flexibility—running a full OS is more heavyweight than running a process, both in terms of initialization (e.g. booting or resuming an OS instance versus `fork`/`exec`), and in terms of resource consumption.

For our target of 10-100 hosted OS instances, we believe this price is worth paying: It allows individual users to run unmodified binaries, or collections of binaries, in a resource controlled fashion (for instance an Apache server along with a PostgreSQL backend). Furthermore it provides an extremely high level of flexibility since the user can dynamically create the precise execution environment their software requires. Unfortunate configuration interactions between various services and applications are avoided (for example, each Windows instance maintains its own registry).

Experience with deployed Xen systems suggests that the initialization overheads and additional resource requirements are in practice quite low: An operating system image may be resumed from an on-disk snapshot in typically just over a second (depending on image memory size), and although multiple copies of the operating system code and data are stored in memory, the memory requirements are typically small compared to those of the applications that will run on them. As we shall show later in the paper, the performance overhead of the virtualization provided by Xen is low, typically just a few percent, even for the most demanding applications.

[1] The Windows XP port required access to Microsoft source code, and hence distribution is currently restricted, even in binary form.

2 XEN: Approach & Overview

In a traditional VMM the virtual hardware exposed is functionally identical to the underlying machine [14]. Although *full virtualization* has the obvious benefit of allowing unmodified operating systems to be hosted, it also has a number of drawbacks. This is particularly true for the prevalent Intel *x86* architecture.

Support for full virtualization was never part of the x86 architectural design. Certain supervisor instructions must be handled by the VMM for correct virtualization, but executing these with insufficient privilege fails silently rather than causing a convenient trap [13]. Efficiently virtualizing the x86 MMU is also difficult. These problems can be solved, but only at the cost of increased complexity and reduced performance. VMware's ESX Server [3] dynamically rewrites portions of the hosted machine code to insert traps wherever VMM intervention might be required. This translation is applied to the entire guest OS kernel (with associated translation, execution, and caching costs) since all non-trapping privileged instructions must be caught and handled. ESX Server implements shadow versions of system structures such as page tables and maintains consistency with the virtual tables by trapping every update attempt—this approach has a high cost for update-intensive operations such as creating a new application process.

Notwithstanding the intricacies of the x86, there are other arguments against full virtualization. In particular, there are situations in which it is desirable for the hosted operating systems to see real as well as virtual resources: providing both real and virtual time allows a guest OS to better support time-sensitive tasks, and to correctly handle TCP timeouts and RTT estimates, while exposing real machine addresses allows a guest OS to improve performance by using superpages [10] or page coloring [7].

We avoid the drawbacks of full virtualization by presenting a virtual machine abstraction that is similar but not identical to the underlying hardware—an approach which has been dubbed *paravirtualization* [17]. This promises improved performance, although it does require modifications to the guest operating system. It is important to note, however, that we do not require changes to the application binary interface (ABI), and hence no modifications are required to guest *applications*.

We distill the discussion so far into a set of design principles:

1. Support for unmodified application binaries is essential, or users will not transition to Xen. Hence we must virtualize all architectural features required by existing standard ABIs.
2. Supporting full multi-application operating systems is important, as this allows complex server configurations to be virtualized within a single guest OS instance.
3. Paravirtualization is necessary to obtain high performance and strong resource isolation on uncooperative machine architectures such as x86.
4. Even on cooperative machine architectures, completely hiding the effects of resource virtualization from guest OSes risks both correctness and performance.

In the following section we describe the virtual machine abstraction exported by Xen and discuss how a guest OS must be modified to conform to this. Note that in this paper we reserve the term *guest operating system* to refer to one of the OSes that Xen can host and we use the term *domain* to refer to a running virtual machine within which a guest OS executes; the

distinction is analogous to that between a *program* and a *process* in a conventional system. We call Xen itself the *hypervisor* since it operates at a higher privilege level than the supervisor code of the guest operating systems that it hosts.

2.1 The Virtual Machine Interface

The paravirtualized x86 interface can be factored into three broad aspects of the system: memory management, the CPU, and device I/O. In the following we address each machine subsystem in turn, and discuss how each is presented in our paravirtualized architecture. Note that although certain parts of our implementation, such as memory management, are specific to the x86, many aspects (such as our virtual CPU and I/O devices) can be readily applied to other machine architectures. Furthermore, x86 represents a *worst case* in the areas where it differs significantly from RISC-style processors—for example, efficiently virtualizing hardware page tables is more difficult than virtualizing a software-managed TLB.

2.1.1 Memory management

Virtualizing memory is undoubtedly the most difficult part of paravirtualizing an architecture, both in terms of the mechanisms required in the hypervisor and modifications required to port each guest OS. The task is easier if the architecture provides a software-managed TLB as these can be efficiently virtualized in a simple manner [5]. A tagged TLB is another useful feature supported by most server-class RISC architectures, including Alpha, MIPS and SPARC. Associating an address-space identifier tag with each TLB entry allows the hypervisor and each guest OS to efficiently coexist in separate address spaces because there is no need to flush the entire TLB when transferring execution.

Unfortunately, x86 does not have a software-managed TLB; instead TLB misses are serviced automatically by the processor by walking the page table structure in hardware. Thus to achieve the best possible performance, all valid page translations for the current address space should be present in the hardware-accessible page table. Moreover, because the TLB is not tagged, address space switches typically require a complete TLB flush. Given these limitations, we made two decisions: (i) guest OSes are responsible for allocating and managing the hardware page tables, with minimal involvement from Xen to ensure safety and isolation; and (ii) Xen exists in a 64MB section at the top of every address space, thus avoiding a TLB flush when entering and leaving the hypervisor.

Each time a guest OS requires a new page table, perhaps because a new process is being created, it allocates and initializes a page from its own memory reservation and registers it with Xen. At this point the OS must relinquish direct write privileges to the page-table memory: all subsequent updates must be validated by Xen. This restricts updates in a number of ways, including only allowing an OS to map pages that it owns, and disallowing writable mappings of page tables. Guest OSes may *batch* update requests to amortize the overhead of entering the hypervisor. The top 64MB region of each address space, which is reserved for Xen, is not accessible or remappable by guest OSes. This address region is not used by any of the common x86 ABIs however, so this restriction does not break application compatibility.

Segmentation is virtualized in a similar way, by validating updates to hardware segment descriptor tables. The only restrictions on x86 segment descriptors are: (i) they must have

lower privilege than Xen, and (ii) they may not allow any access to the Xen-reserved portion of the address space.

2.1.2 CPU

Virtualizing the CPU has several implications for guest OSes. Principally, the insertion of a hypervisor below the operating system violates the usual assumption that the OS is the most privileged entity in the system. In order to protect the hypervisor from OS misbehavior (and domains from one another) guest OSes must be modified to run at a lower privilege level.

Efficient virtualizion of privilege levels is possible on x86 because it supports four distinct privilege levels in hardware. The x86 privilege levels are generally described as *rings*, and are numbered from zero (most privileged) to three (least privileged). OS code typically executes in ring 0 because no other ring can execute privileged instructions, while ring 3 is generally used for application code. To our knowledge, rings 1 and 2 have not been used by any well-known x86 OS since OS/2. Any OS which follows this common arrangement can be ported to Xen by modifying it to execute in ring 1. This prevents the guest OS from directly executing privileged instructions, yet it remains safely isolated from applications running in ring 3.

Privileged instructions are paravirtualized by requiring them to be validated and executed within Xen—this applies to operations such as installing a new page table, or yielding the processor when idle (rather than attempting to `hlt` it). Any guest OS attempt to directly execute a privileged instruction is failed by the processor, either silently or by taking a fault, since only Xen executes at a sufficiently privileged level.

Exceptions, including memory faults and software traps, are virtualized on x86 very straightforwardly. A table describing the handler for each type of exception is registered with Xen for validation. The handlers specified in this table are generally identical to those for real x86 hardware; this is possible because the exception stack frames are unmodified in our paravirtualized architecture. The sole modification is to the page fault handler, which would normally read the faulting address from a privileged processor register (`CR2`); since this is not possible, we write it into an extended stack frame[2]. When an exception occurs while executing outside ring 0, Xen's handler creates a copy of the exception stack frame on the guest OS stack and returns control to the appropriate registered handler.

Typically only two types of exception occur frequently enough to affect system performance: system calls (which are usually implemented via a software exception), and page faults. We improve the performance of system calls by allowing each guest OS to register a 'fast' exception handler which is accessed directly by the processor without indirecting via ring 0; this handler is validated before installing it in the hardware exception table. Unfortunately it is not possible to apply the same technique to the page fault handler because only code executing in ring 0 can read the faulting address from register `CR2`; page faults must therefore always be delivered via Xen so that this register value can be saved for access in ring 1.

Safety is ensured by validating exception handlers when they are presented to Xen. The only required check is that the handler's code segment does not specify execution in ring 0. Since no guest OS can create such a segment,

[2] In hindsight, writing the value into a pre-agreed shared memory location rather than modifying the stack frame would have simplified the XP port.

it suffices to compare the specified segment selector to a small number of static values which are reserved by Xen. Apart from this, any other handler problems are fixed up during exception propagation—for example, if the handler's code segment is not present or if the handler is not paged into memory then an appropriate fault will be taken when Xen executes the `iret` instruction which returns to the handler. Xen detects these "double faults" by checking the faulting program counter value: if the address resides within the exception-virtualizing code then the offending guest OS is terminated.

Note that this "lazy" checking is safe even for the direct system-call handler: access faults will occur when the CPU attempts to directly jump to the guest OS handler. In this case the faulting address will be outside Xen (since Xen will never execute a guest OS system call) and so the fault is virtualized in the normal way. If propagation of the fault causes a further "double fault" then the guest OS is terminated as described above.

2.1.3 Device I/O

Rather than emulating existing hardware devices, as is typically done in fully-virtualized environments, Xen exposes a set of clean and simple device abstractions. This allows us to design an interface that is both efficient and satisfies our requirements for protection and isolation. To this end, I/O data is transferred to and from each domain via Xen, using shared-memory, asynchronous buffer-descriptor rings. These provide a high-performance communication mechanism for passing buffer information vertically through the system, while allowing Xen to efficiently perform validation checks (for example, checking that buffers are contained within a domain's memory reservation).

Linux subsection	# lines
Architecture-independent	78
Virtual network driver	484
Virtual block-device driver	1070
Xen-specific (non-driver)	1363
Total	**2995**
Portion of total x86 code base	**1.36%**

Table 1: The simplicity of porting commodity OSes to Xen.

Similar to hardware interrupts, Xen supports a lightweight event-delivery mechanism which is used for sending asynchronous notifications to a domain. These notifications are made by updating a bitmap of pending event types and, optionally, by calling an event handler specified by the guest OS. These callbacks can be 'held off' at the discretion of the guest OS—to avoid extra costs incurred by frequent wake-up notifications, for example.

2.2 The Cost of Porting an OS to Xen

Table 1 demonstrates the cost, in lines of code, of porting commodity operating systems to Xen's paravirtualized x86 environment.

The architecture-specific sections are effectively a port of the x86 code to our paravirtualized architecture. This involved rewriting routines which used privileged instructions, and removing a large amount of low-level system initialization code.

2.3 Control and Management

Throughout the design and implementation of Xen, a goal has been to separate policy from mechanism wherever possible. Although the hypervisor must be involved in data-path aspects (for example, scheduling the CPU between domains, filtering network packets before transmission, or enforcing access control

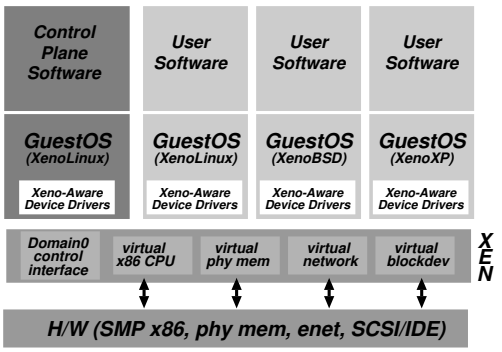


Figure 1: The structure of a machine running the Xen hypervisor, hosting a number of different guest operating systems, including *Domain0* running control software in a XenLinux environment.

when reading data blocks), there is no need for it to be involved in, or even aware of, higher level issues such as how the CPU is to be shared, or which kinds of packet each domain may transmit.

The resulting architecture is one in which the hypervisor itself provides only basic control operations. These are exported through an interface accessible from authorized domains; potentially complex policy decisions, such as admission control, are best performed by management software running over a guest OS rather than in privileged hypervisor code.

The overall system structure is illustrated in Figure 1. Note that a domain is created at boot time which is permitted to use the *control interface*. This initial domain, termed *Domain0*, is responsible for hosting the application-level management software. The control interface provides the ability to create and terminate other domains and to control their associated scheduling parameters, physical memory allocations and the access they are given to the machine's physical disks and network devices.

In addition to processor and memory resources, the control interface supports the creation and deletion of virtual network interfaces (VIFs) and block devices (VBDs). These virtual I/O devices have associated access-control information which determines which domains can access them, and with what restrictions (for example, a read-only VBD may be created, or a VIF may filter IP packets to prevent source-address spoofing or apply traffic shaping).

This control interface, together with profiling statistics on the current state of the system, is exported to a suite of application-level management software running in *Domain0*. This complement of administrative tools allows convenient management of the entire server: current tools can create and destroy domains, set network filters and routing rules, monitor per-domain network activity at packet and flow granularity, and create and delete virtual network interfaces and virtual block devices.

Snapshots of a domains' state may be captured and saved to disk, enabling rapid deployment of applications by bypassing the normal boot delay. Further, Xen supports *live migration* which enables running VMs to be moved dynamically between different Xen servers, with execution interrupted only for a few milliseconds. We are in the process of developing higher-level tools to further automate the application of administrative policy, for example, load balancing VMs among a cluster of Xen servers.

3 Detailed Design

In this section we introduce the design of the major subsystems that make up a Xen-based server. In each case we present both Xen and guest OS functionality for clarity of exposition. In this paper, we focus on the XenLinux guest OS; the *BSD and Windows XP ports use the Xen interface in a similar manner.

3.1 Control Transfer: Hypercalls and Events

Two mechanisms exist for control interactions between Xen and an overlying domain: synchronous calls from a domain to Xen may be made using a *hypercall*, while notifications are delivered to domains from Xen using an asynchronous event mechanism.

The hypercall interface allows domains to perform a synchronous software trap into the hypervisor to perform a privileged operation, analogous to the use of system calls in conventional operating systems. An example use of a hypercall is to request a set of page-table updates, in which Xen validates and applies a list of updates, returning control to the calling domain when this is completed.

Communication from Xen to a domain is provided through an asynchronous event mechanism, which replaces the usual delivery mechanisms for device interrupts and allows lightweight notification of important events such as domain-termination requests. Akin to traditional Unix signals, there are only a small number of events, each acting to flag a particular type of occurrence. For instance, events are used to indicate that new data has been received over the network, or that a virtual disk request has completed.

Pending events are stored in a per-domain bitmask which is updated by Xen before invoking an event-callback handler specified by the guest OS. The callback handler is responsible for resetting the set of pending events, and responding to the notifications in an appropriate manner. A domain may explicitly defer event handling by setting a Xen-readable software flag: this is analogous to disabling interrupts on a real processor.

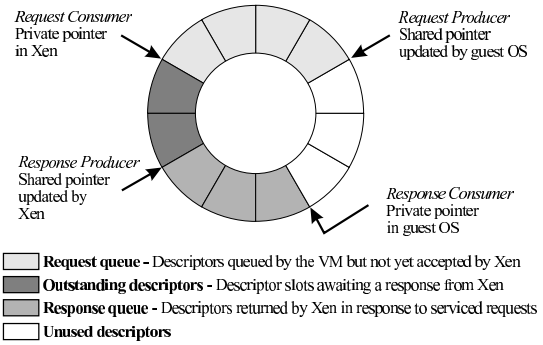


Figure 2: The structure of asynchronous I/O rings, which are used for data transfer between Xen and guest OSes.

3.2 Data Transfer: I/O Rings

The presence of a hypervisor means there is an additional protection domain between guest OSes and I/O devices, so it is crucial that a data transfer mechanism be provided that allows data to move vertically through the system with as little overhead as possible.

Two main factors have shaped the design of our I/O-transfer mechanism: resource management and event notification. For resource accountability, we attempt to minimize the work required to demultiplex data to a specific domain when an interrupt is received from a device—the overhead of managing buffers is carried out later where computation may be accounted to the appropriate domain. Similarly, memory committed to device I/O is provided by the relevant domains wherever possible to prevent the crosstalk inherent in shared buffer pools; I/O buffers are protected during data transfer by pinning the underlying page frames within Xen.

Figure 2 shows the structure of our I/O descriptor rings. A ring is a circular queue of descriptors allocated by a domain but accessible from within Xen. Descriptors do not directly contain I/O data; instead, I/O data buffers are al-

located out-of-band by the guest OS and indirectly referenced by I/O descriptors. Access to each ring is based around two pairs of producer-consumer pointers: domains place requests on a ring, advancing a request producer pointer, and Xen removes these requests for handling, advancing an associated request consumer pointer. Responses are placed back on the ring similarly, save with Xen as the producer and the guest OS as the consumer. There is no requirement that requests be processed in order: the guest OS associates a unique identifier with each request which is reproduced in the associated response. This allows Xen to unambiguously reorder I/O operations due to scheduling or priority considerations.

This structure is sufficiently generic to support a number of different device paradigms. For example, a set of 'requests' can provide buffers for network packet reception; subsequent 'responses' then signal the arrival of packets into these buffers. Reordering is useful when dealing with disk requests as it allows them to be scheduled within Xen for efficiency, and the use of descriptors with out-of-band buffers makes implementing zero-copy transfer easy.

We decouple the production of requests or responses from the notification of the other party: in the case of requests, a domain may enqueue multiple entries before invoking a hypercall to alert Xen; in the case of responses, a domain can defer delivery of a notification event by specifying a threshold number of responses. This allows each domain to trade-off latency and throughput requirements, similarly to the flow-aware interrupt dispatch in the ArseNIC Gigabit Ethernet interface [12].

3.3 Subsystem Virtualization

The control and data transfer mechanisms described are used in our virtualization of the various subsystems. In the following, we discuss how this virtualization is achieved for CPU, timers, memory, network and disk.

3.3.1 CPU scheduling

Xen currently schedules domains according to the Borrowed Virtual Time (BVT) scheduling algorithm [4]. We chose this particular algorithms since it is both work-conserving and has a special mechanism for low-latency wake-up (or *dispatch*) of a domain when it receives an event. Fast dispatch is particularly important to minimize the effect of virtualization on OS subsystems that are designed to run in a timely fashion; for example, TCP relies on the timely delivery of acknowledgments to correctly estimate network round-trip times. BVT provides low-latency dispatch by using virtual-time warping, a mechanism which temporarily violates 'ideal' fair sharing to favor recently-woken domains. However, other scheduling algorithms could be trivially implemented over our generic scheduler abstraction. Per-domain scheduling parameters can be adjusted by management software running in *Domain0*.

3.3.2 Time and timers

Xen provides guest OSes with notions of real time, virtual time and wall-clock time. Real time is expressed in nanoseconds passed since machine boot and is maintained to the accuracy of the processor's cycle counter and can be frequency-locked to an external time source (for example, via NTP). A domain's virtual time only advances while it is executing: this is typically used by the guest OS scheduler to ensure correct sharing of its timeslice between application processes. Finally, wall-clock time is specified as an offset to be added to the current real time. This allows the wall-clock time to be adjusted without affecting the forward

progress of real time.

Each guest OS can program a pair of alarm timers, one for real time and the other for virtual time. Guest OSes are expected to maintain internal timer queues and use the Xen-provided alarm timers to trigger the earliest timeout. Timeouts are delivered using Xen's event mechanism.

3.3.3 Virtual address translation

As with other subsystems, Xen attempts to virtualize memory access with as little overhead as possible. As discussed in Section 2.1.1, this goal is made somewhat more difficult by the x86 architecture's use of hardware page tables. The approach taken by VMware is to provide each guest OS with a virtual page table, not visible to the memory-management unit (MMU) [3]. The hypervisor is then responsible for trapping accesses to the virtual page table, validating updates, and propagating changes back and forth between it and the MMU-visible 'shadow' page table. This greatly increases the cost of certain guest OS operations, such as creating new virtual address spaces, and requires explicit propagation of hardware updates to 'accessed' and 'dirty' bits.

Although full virtualization forces the use of shadow page tables, to give the illusion of contiguous physical memory, Xen is not so constrained. Indeed, Xen need only be involved in page table *updates*, to prevent guest OSes from making unacceptable changes. Thus we avoid the overhead and additional complexity associated with the use of shadow page tables—the approach in Xen is to register guest OS page tables directly with the MMU, and restrict guest OSes to read-only access. Page table updates are passed to Xen via a hypercall; to ensure safety, requests are *validated* before being applied.

To aid validation, we associate a type and reference count with each machine page frame. A frame may have any one of the following mutually-exclusive types at any point in time: page directory (PD), page table (PT), local descriptor table (LDT), global descriptor table (GDT), or writable (RW). Note that a guest OS may always create readable mappings to its own page frames, regardless of their current types. A frame may only safely be retasked when its reference count is zero. This mechanism is used to maintain the invariants required for safety; for example, a domain cannot have a writable mapping to any part of a page table as this would require the frame concerned to simultaneously be of types PT and RW.

The type system is also used to track which frames have already been validated for use in page tables. To this end, guest OSes indicate when a frame is allocated for page-table use—this requires a one-off validation of every entry in the frame by Xen, after which its type is pinned to PD or PT as appropriate, until a subsequent unpin request from the guest OS. This is particularly useful when changing the page table base pointer, as it obviates the need to validate the new page table on every context switch. Note that a frame cannot be retasked until it is both unpinned and its reference count has reduced to zero – this prevents guest OSes from using unpin requests to circumvent the reference-counting mechanism.

3.3.4 Physical memory

The initial memory allocation, or *reservation*, for each domain is specified at the time of its creation; memory is thus statically partitioned between domains, providing strong isolation. A maximum-allowable reservation may also be specified: if memory pressure within a domain increases, it may then attempt to claim additional memory pages from Xen, up

to this reservation limit. Conversely, if a domain wishes to save resources, perhaps to avoid incurring unnecessary costs, it can reduce its memory reservation by releasing memory pages back to Xen.

XenLinux implements a *balloon driver* [16], which adjusts a domain's memory usage by passing memory pages back and forth between Xen and XenLinux's page allocator. Although we could modify Linux's memory-management routines directly, the balloon driver makes adjustments by using existing OS functions, thus simplifying the Linux porting effort. However, paravirtualization can be used to extend the capabilities of the balloon driver; for example, the out-of-memory handling mechanism in the guest OS can be modified to automatically alleviate memory pressure by requesting more memory from Xen.

Most operating systems assume that memory comprises at most a few large contiguous extents. Because Xen does not guarantee to allocate contiguous regions of memory, guest OSes will typically create for themselves the illusion of contiguous *physical memory*, even though their underlying allocation of *hardware memory* is sparse. Mapping from physical to hardware addresses is entirely the responsibility of the guest OS, which can simply maintain an array indexed by physical page frame number. Xen supports efficient hardware-to-physical mapping by providing a shared translation array that is directly readable by all domains – updates to this array are validated by Xen to ensure that the OS concerned owns the relevant hardware page frames.

Note that even if a guest OS chooses to ignore hardware addresses in most cases, it must use the translation tables when accessing its page tables (which necessarily use hardware addresses). Hardware addresses may also be exposed to limited parts of the OS's memory-management system to optimize memory access. For example, a guest OS might allocate particular hardware pages so as to optimize placement within a physically indexed cache [7], or map naturally aligned contiguous portions of hardware memory using superpages [10].

3.3.5 Network

Xen provides the abstraction of a virtual firewall-router (VFR), where each domain has one or more network interfaces (VIFs) logically attached to the VFR. A VIF looks somewhat like a modern network interface card: there are two I/O rings of buffer descriptors, one for transmit and one for receive. Each direction also has a list of associated rules of the form (<*pattern*>, <*action*>)—if the *pattern* matches then the associated *action* is applied.

Domain0 is responsible for inserting and removing rules. In typical cases, rules will be installed to prevent IP source address spoofing, and to ensure correct demultiplexing based on destination IP address and port. Rules may also be associated with hardware interfaces on the VFR. In particular, we may install rules to perform traditional firewalling functions such as preventing incoming connection attempts on insecure ports.

To transmit a packet, the guest OS simply enqueues a buffer descriptor onto the transmit ring. Xen copies the descriptor and, to ensure safety, then copies the packet header and executes any matching filter rules. The packet payload is not copied since we use scatter-gather DMA; however note that the relevant page frames must be pinned until transmission is complete. To ensure fairness, Xen implements a simple round-robin packet scheduler.

To efficiently implement packet reception, we

require the guest OS to exchange an unused page frame for each packet it receives; this avoids the need to copy the packet between Xen and the guest OS, although it requires that page-aligned receive buffers be queued at the network interface. When a packet is received, Xen immediately checks the set of receive rules to determine the destination VIF, and exchanges the packet buffer for a page frame on the relevant receive ring. If no frame is available, the packet is dropped.

3.3.6 Disk

Only *Domain0* has direct unchecked access to physical (IDE and SCSI) disks. All other domains access persistent storage through the abstraction of virtual block devices (VBDs), which are created and configured by management software running within *Domain0*. Allowing *Domain0* to manage the VBDs keeps the mechanisms within Xen very simple and avoids more intricate solutions such as the UDFs used by the Exokernel [6].

A VBD comprises a list of extents with associated ownership and access control information, and is accessed via the I/O ring mechanism. A typical guest OS disk scheduling algorithm will reorder requests prior to enqueuing them on the ring in an attempt to reduce response time, and to apply differentiated service (for example, it may choose to aggressively schedule synchronous metadata requests at the expense of speculative readahead requests). However, because Xen has more complete knowledge of the actual disk layout, we also support reordering within Xen, and so responses may be returned out of order. A VBD thus appears to the guest OS somewhat like a SCSI disk.

A translation table is maintained within the hypervisor for each VBD; the entries within this table are installed and managed by *Domain0* via a privileged control interface. On receiving a disk request, Xen inspects the VBD identifier and offset and produces the corresponding sector address and physical device. Permission checks also take place at this time. Zero-copy data transfer takes place using DMA between the disk and pinned memory pages in the requesting domain.

Xen services *batches* of requests from competing domains in a simple round-robin fashion; these are then passed to a standard elevator scheduler before reaching the disk hardware. Domains may explicitly pass down *reorder barriers* to prevent reordering when this is necessary to maintain higher level semantics (e.g. when using a write-ahead log). The low-level scheduling gives us good throughput, while the batching of requests provides reasonably fair access. Future work will investigate providing more predictable isolation and differentiated service, perhaps using existing techniques and schedulers [15].

4 Evaluation

In this section we present a subset of our evaluation of Xen against a number of alternative virtualization techniques. A more complete evaluation, as well as detailed configuration and benchmark specs, can be found in [1] For these measurements, we used our 2.4.21-based XenLinux port as, at the time of this writing, the 2.6-port was not stable enough for a full battery of tests.

There are a number of preexisting solutions for running multiple copies of Linux on the same machine. VMware offers several commercial products that provide virtual x86 machines on which unmodified copies of Linux may be booted. The most commonly used version is VMware Workstation, which consists

of a set of privileged kernel extensions to a 'host' operating system. Both Windows and Linux hosts are supported. VMware also offer an enhanced product called ESX Server which replaces the host OS with a dedicated kernel. By doing so, it gains some performance benefit over the workstation product. We have subjected ESX Server to the benchmark suites described below, but sadly are prevented from reporting quantitative results due to the terms of the product's End User License Agreement. Instead we present results from VMware Workstation 3.2, running on top of a Linux host OS, as it is the most recent VMware product without that benchmark publication restriction. ESX Server takes advantage of its native architecture to equal or outperform VMware Workstation and its hosted architecture. While Xen of course requires guest OSes to be ported, it takes advantage of paravirtualization to noticeably outperform ESX Server.

We also present results for User-mode Linux (UML), an increasingly popular platform for virtual hosting. UML is a port of Linux to run as a user-space process on a Linux host. Like XenLinux, the changes required are restricted to the architecture dependent code base. However, the UML code bears little similarity to the native x86 port due to the very different nature of the execution environments. Although UML can run on an unmodified Linux host, we present results for the 'Single Kernel Address Space' (skas3) variant that exploits patches to the host OS to improve performance.

We also investigated three other virtualization techniques for running ported versions of Linux on the same x86 machine. Connectix's Virtual PC and forthcoming Virtual Server products (now acquired by Microsoft) are similar in design to VMware's, providing full x86 virtualization. Since all versions of Virtual PC have benchmarking restrictions in their license agreements we did not subject them to closer analysis. UMLinux is similar in concept to UML but is a different code base and has yet to achieve the same level of performance, so we omit the results. Work to improve the performance of UMLinux through host OS modifications is ongoing [8]. Although Plex86 was originally a general purpose x86 VMM, it has now been retargeted to support just Linux guest OSes. The guest OS must be specially compiled to run on Plex86, but the source changes from native x86 are trivial. The performance of Plex86 is currently well below the other techniques.

4.1 Relative Performance

The first cluster of bars in Figure 3 represents a relatively easy scenario for the VMMs. The SPEC CPU suite contains a series of long-running computationally-intensive applications intended to measure the performance of a system's processor, memory system, and compiler quality. The suite performs little I/O and has little interaction with the OS. With almost all CPU time spent executing in user-space code, all three VMMs exhibit low overhead.

The next set of bars show the total elapsed time taken to build a default configuration of the Linux 2.4.21 kernel on a local ext3 file system with gcc 2.96. Native Linux spends about 7% of the CPU time in the OS, mainly performing file I/O, scheduling and memory management. In the case of the VMMs, this 'system time' is expanded to a greater or lesser degree: whereas Xen incurs a mere 3% overhead, the other VMMs experience a more significant slowdown.

Two experiments were performed using the PostgreSQL 7.1.3 database, exercised by the Open Source Database Benchmark suite (OSDB) in its default configuration. We present results for the multi-user Information

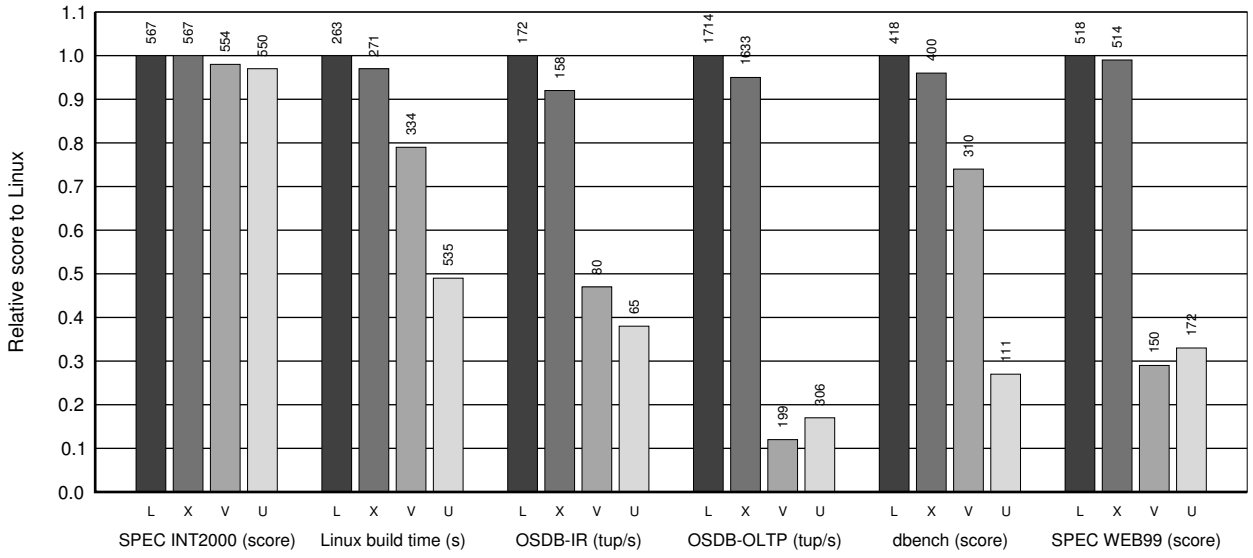


Figure 3: Relative performance of native Linux (L), XenLinux (X), VMware workstation 3.2 (V) and User-Mode Linux (U).

Retrieval (IR) and On-Line Transaction Processing (OLTP) workloads, both measured in tuples per second. PostgreSQL places considerable load on the operating system, and this is reflected in the substantial virtualization overheads experienced by VMware and UML. In particular, the OLTP benchmark requires many synchronous disk operations, resulting in many protection domain transitions.

The `dbench` program is a file system benchmark derived from the industry-standard 'NetBench'. It emulates the load placed on a file server by Windows 95 clients. Here, we examine the throughput experienced by a single client performing around 90,000 file system operations.

SPEC WEB99 is a complex application-level benchmark for evaluating web servers and the systems that host them. The benchmark is CPU-bound, and a significant proportion of the time is spent within the guest OS kernel, performing network stack processing, file system operations, and scheduling between the many `httpd` processes that Apache needs to handle the offered load. XenLinux fares well, achieving within 1% of native Linux performance. VMware and UML both struggle, supporting less than a third of the number of clients of the native Linux system.

4.2 Operating System Benchmarks

To more precisely measure the areas of overhead within Xen and the other VMMs, we performed a number of smaller experiments targeting particular subsystems. We examined the overhead of virtualization as measured by McVoy's *lmbench* program [9]. The OS performance subset of the lmbench suite consist of 37 microbenchmarks.

In 24 of the 37 microbenchmarks, XenLinux performs similarly to native Linux, tracking the Linux kernel performance closely. In Tables 2 to 4 we show results which exhibit interesting performance variations among the test systems; particularly large penalties for Xen are shown in bold face.

In the process microbenchmarks (Table 2), Xen

Config	null call	null I/O	open close	slct TCP	sig inst	sig hndl	fork proc	exec proc	sh proc
Linux	0.45	0.50	1.92	5.70	0.68	2.49	110	530	4k0
Xen	0.46	0.50	1.88	5.69	0.69	1.75	**198**	**768**	**4k8**
VMW	0.73	0.83	2.99	11.1	1.02	4.63	874	2k3	10k
UML	24.7	25.1	62.8	39.9	26.0	46.0	21k	33k	58k

Table 2: `lmbench`: Processes - times in μs

Config	2p 0K	2p 16K	2p 64K	8p 16K	8p 64K	16p 16K	16p 64K
Linux	0.77	0.91	1.06	1.03	24.3	3.61	37.6
Xen	**1.97**	**2.22**	**2.67**	**3.07**	**28.7**	**7.08**	39.4
VMW	18.1	17.6	21.3	22.4	51.6	41.7	72.2
UML	15.5	14.6	14.4	16.3	36.8	23.6	52.0

Table 3: `lmbench`: Context switching times in μs

Config	0K File create	0K File delete	10K File create	10K File delete	Mmap lat	Prot fault	Page fault
Linux	32.1	6.08	66.0	12.5	68.0	1.06	1.42
Xen	32.5	5.86	68.2	13.6	**139**	1.40	**2.73**
VMW	35.3	9.3	85.6	21.4	620	7.53	12.4
UML	130	65.7	250	113	1k4	21.8	26.3

Table 4: `lmbench`: File & VM system latencies in μs

exhibits slower *fork*, *exec*, and *sh* performance than native Linux. This is expected, since these operations require large numbers of page table updates which must all be verified by Xen. However, the paravirtualization approach allows XenLinux to batch update requests. Creating new page tables presents an ideal case: because there is no reason to commit pending updates sooner, XenLinux can amortize each hypercall across 2048 updates (the maximum size of its batch buffer). Hence each update hypercall constructs 8MB of address space.

Table 3 shows context switch times between different numbers of processes with different working set sizes. Xen incurs an extra overhead between 1μs and 3μs, as it executes a hypercall to change the page table base. However, context switch results for larger working set sizes (perhaps more representative of real applications) show that the overhead is small compared with cache effects. Unusually, VMware Workstation is inferior to UML on these microbenchmarks; however, this is one area where enhancements in ESX Server are able to reduce the overhead.

The *mmap latency* and *page fault latency* results shown in Table 4 are interesting since they require two transitions into Xen per page: one to take the hardware fault and pass the details to the guest OS, and a second to install the updated page table entry on the guest OS's behalf. Despite this, the overhead is relatively modest.

One small anomaly in Table 2 is that XenLinux has lower signal-handling latency than native Linux. This benchmark does not require any calls into Xen at all, and the 0.75μs (30%) speedup is presumably due to a fortuitous cache alignment in XenLinux, hence underlining the dangers of taking microbenchmarks too seriously.

4.3 Additional Benchmarks

We have also conducted comprehensive experiments that: evaluate the overhead of virtualizing the network; compare the performance of running multiple applications in their own guest OS against running them on the same native operating system; demonstrate performance isolation provided by Xen; and examine Xen's ability to scale to its target of 100 domains. All of the experiments showed promising results and details have been separately published [1].

5 Conclusion

We have presented the Xen hypervisor which partitions the resources of a computer between domains running guest operating systems. Our

paravirtualizing design places a particular emphasis on protection, performance and resource management. We have also described and evaluated XenLinux, a fully-featured port of the Linux kernel that runs over Xen.

Xen and the 2.4-based XenLinux are sufficiently stable to be useful to a wide audience. Indeed, some web hosting providers are already selling Xen-based virtual servers. Sources, documentation, and a demo ISO can be found on our project page[3].

Although the 2.4-based XenLinux was the basis of our performance evaluation, a 2.6-based port is well underway. In this port, much care is been given to minimizing and isolating the necessary changes to the Linux kernel and measuring the changes against benchmark results. As paravirtualization techniques become more prevalent, kernel changes would ideally be part of the main tree. We have experimented with various source structures including a separate architecture, *a la* UML, a subarchitecture, and a CONFIG option. We eagerly solicit input and discussion from the kernel developers to guide our approach. We also have considered transparent paravirtualization [2] techniques to allow a single distro image to adapt dynamically between a VMM-based configuration and bare metal.

As well as further guest OS ports, Xen itself is being ported to other architectures. An x86_64 port is well underway, and we are keen to see Xen ported to RISC-style architectures (such as PPC) where virtual memory virtualization will be much easier due to the software-managed TLB.

Much new functionality has been added since the first public availability of Xen last October. Of particular note are a completely revamped I/O subsystem capable of directly utilizing Linux driver source, suspend/resume and live migration features, much improved console access, etc. Though final implementation, testing, and documentation was not complete at the deadline for this paper, we hope to describe these in more detail at the symposium and in future publications.

As always, there are more tasks to do than there are resources to do them. We would like to grow Xen into the premier open source virtualization solution, with breadth and features that rival proprietary commercial products.

We enthusiastically welcome the help and contributions of the Linux community.

[3] `http://www.cl.cam.ac.uk/netos/xen`

Acknowledgments

Xen has been a big team effort. In particular, we would like to thank Zachary Amsden, Andrew Chung, Richard Coles, Boris Dragovich, Evangelos Kotsovinos, Tim Harris, Alex Ho, Kip Macy, Rolf Neugebauer, Bin Ren, Russ Ross, James Scott, Steven Smith, Andrew Warfield, Mark Williamson, and Mike Wray. Work on Xen has been supported by a UK EPSRC grant, Intel Research, Microsoft Research, and Hewlett-Packard Labs.

References

[1] P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt, and A. Warfield. Xen and the Art of Virtualization. In *Proceedings of the 19th ACM SIGOPS Symposium on Operating Systems Principles*, volume 37(5) of *ACM Operating Systems Review*, pages 164–177, Bolton Landing, NY, USA, Dec. 2003.

[2] D. Magenheimer and T. Christian. vBlades: Optimized Paravirtualization for the Itanium Processor Family. In *Proceedings of the*

USENIX 3rd Virtual Machine Research and Technology Symposium, May 2004.

[3] S. Devine, E. Bugnion, and M. Rosenblum. Virtualization system including a virtual machine monitor for a computer with a segmented architecture. *US Patent*, 6397242, Oct. 1998.

[4] K. J. Duda and D. R. Cheriton. Borrowed-Virtual-Time (BVT) scheduling: supporting latency-sensitive threads in a general-purpose scheduler. In *Proceedings of the 17th ACM SIGOPS Symposium on Operating Systems Principles*, volume 33(5) of *ACM Operating Systems Review*, pages 261–276, Kiawah Island Resort, SC, USA, Dec. 1999.

[5] D. Engler, S. K. Gupta, and F. Kaashoek. AVM: Application-level virtual memory. In *Proceedings of the 5th Workshop on Hot Topics in Operating Systems*, pages 72–77, May 1995.

[6] M. F. Kaashoek, D. R. Engler, G. R. Granger, H. M. Briceño, R. Hunt, D. Mazières, T. Pinckney, R. Grimm, J. Jannotti, and K. Mackenzie. Application performance and flexibility on Exokernel systems. In *Proceedings of the 16th ACM SIGOPS Symposium on Operating Systems Principles*, volume 31(5) of *ACM Operating Systems Review*, pages 52–65, Oct. 1997.

[7] R. Kessler and M. Hill. Page placement algorithms for large real-indexed caches. *ACM Transaction on Computer Systems*, 10(4):338–359, Nov. 1992.

[8] S. T. King, G. W. Dunlap, and P. M. Chen. Operating System Support for Virtual Machines. In *Proceedings of the 2003 Annual USENIX Technical Conference*, Jun 2003.

[9] L. McVoy and C. Staelin. lmbench: Portable tools for performance analysis. In *Proceedings of the USENIX Annual Technical Conference*, pages 279–294, Berkeley, Jan. 1996. Usenix Association.

[10] J. Navarro, S. Iyer, P. Druschel, and A. Cox. Practical, transparent operating system support for superpages. In *Proceedings of the 5th Symposium on Operating Systems Design and Implementation (OSDI 2002)*, ACM Operating Systems Review, Winter 2002 Special Issue, pages 89–104, Boston, MA, USA, Dec. 2002.

[11] L. Peterson, D. Culler, T. Anderson, and T. Roscoe. A blueprint for introducing disruptive technology into the internet. In *Proceedings of the 1st Workshop on Hot Topics in Networks (HotNets-I)*, Princeton, NJ, USA, Oct. 2002.

[12] I. Pratt and K. Fraser. Arsenic: A user-accessible gigabit ethernet interface. In *Proceedings of the Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM-01)*, pages 67–76, Los Alamitos, CA, USA, Apr. 22–26 2001. IEEE Computer Society.

[13] J. S. Robin and C. E. Irvine. Analysis of the Intel Pentium's ability to support a secure virtual machine monitor. In *Proceedings of the 9th USENIX Security Symposium, Denver, CO, USA*, pages 129–144, Aug. 2000.

[14] L. Seawright and R. MacKinnon. VM/370 – a study of multiplicity and usefulness. *IBM Systems Journal*, pages 4–17, 1979.

[15] P. Shenoy and H. Vin. Cello: A Disk Scheduling Framework for Next-generation Operating Systems. In *Proceedings of ACM SIGMETRICS'98, the International Conference on Measurement and Modeling of Computer Systems*, pages 44–55, June 1998.

[16] C. A. Waldspurger. Memory resource management in VMware ESX server. In *Proceedings of the 5th Symposium on Operating Systems Design and Implementation (OSDI 2002)*, ACM Operating Systems Review, Winter 2002 Special Issue, pages 181–194, Boston, MA, USA, Dec. 2002.

[17] A. Whitaker, M. Shaw, and S. D. Gribble. Denali: Lightweight Virtual Machines for Distributed and Networked Applications. Technical Report 02-02-01, University of Washington, 2002.

TIPC: Providing Communication for Linux Clusters

Jon Paul Maloy
Ericsson Research, Montreal
jon.maloy@ericsson.com

Abstract

Transparent Inter Process Communication (TIPC) is a protocol specially designed for efficient intra cluster communication, leveraging the particular conditions present within clusters of loosely coupled nodes.

TIPC provides a powerful infrastructure for designing distributed, site-independent, scalable, highly- available and high-performing applications, as well as a good support for cluster, network and software management functionality. In this paper, we will discuss the motives for developing TIPC and describe its architecture. Then, we will present the most important features of TIPC, such as its functional, location transparent, addressing scheme, lightweight reactive connections, reliable multicast, signalling link protocol, topology subscription services and more. We will also discuss the various design decisions that influenced the implementation of these features. We conclude by describing the current implementation status and our planned roadmap for TIPC.

1 Introduction

For the last six years, telecom equipment vendor Ericsson has been developing and deploying a tailor-made reliable communication protocol, TIPC,for their cluster-based products. This protocol has recently undergone a significant redesign, and is now available as a portable source code package of about 12,500 lines of C code. The code implements a Linux kernel driver, a design that has made it possible to improve performance (35% faster than TCP) and minimize code footprint.

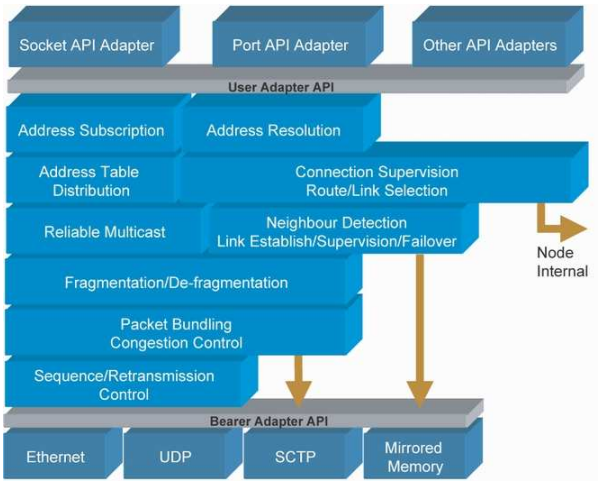


Figure 1: *Functional View of TIPC*

The current version is available under a dual BSD/GPL license from [1]. TIPC is supported on Linux 2.4 and 2.6; and several proprietary portations to other OS'es (OSE, True64, VxWorks) also exist.

TIPC offers an interesting combination of features, some of them quite unique, to achieve the overall goal: to make the cluster act as one single computer from a communication viewpoint, while helping applications to keep track of and adapt to topology changes. Figure 1 illustrates a functional view of TIPC.

2 Motivation

There are no standard protocols available today that fully satisfy the special needs of application programs working within highly available, dynamic cluster environments. Clusters may grow or shrink by orders of magnitude; member nodes may crash and restart, routers may fail and be replaced, services may be moved around due to load balancing considerations, etc. All this must be possibe to handle without significant disturbances of the service offered by the cluster. In order to minimize the effort by the application programmers to deal with such situations, and to maximize the chance that they are handled in a correct and optimal way, the cluster internal communication service should provide special support, helping the applications to adapt to changes in the cluster. It should also, when possible, leverage the special conditions present within cluster environments to present a more efficient and fault-tolerant communication service than more general protocols are capable of.

2.1 Existing Protocols

TCP [2] has the advantage of being ubiquitous, proven, and wellknown by most programmers. Its most significant shortcomings in a real-time cluster environment are the following:

- TCP lacks any notion of functional addressing and addressing transparency. Mechanisms exist (DNS, CORBA Naming Service) for transparent and dynamic lookup of the correct IP-adress of a destination, but those are in general too static and too inefficient to be useful in a dynamic, real-time environment.

- Performance is not as good as it could be, especially for intra-node communication and for short messages in general. For intra-node communication, other more efficient mechanisms are available, at least on Unix, but then the location of the destination process has to be assumed, and can not be changed. It is desirable to have a protocol working efficiently for both intra-node and inter-node messaging, without forcing the user to distinguish between these cases in his code.

- The heavy connection setup/shutdown scheme of TCP is a disadvantage in a dynamic environment. The minimum number of packets exchanged for even the shortest TCP transaction is nine (SYN, SYNACK, etc.), while with TIPC this can be reduced to two, or even to one if connectionless mode is used.

- The connection-oriented nature of TCP makes it impossible to support true multicast.

Stream Control Transmission Protocol (SCTP) [3] is message oriented; it provides some level of user connection supervision, message bundling, loss-free changeover, and a few more features that may make it more suitable than TCP as an intra-cluster protocol. Otherwise, it has all the drawbacks of TCP already listed above.

Apart from these weaknesses, neither TCP nor SCTP provide any topology information/subscription service, something that has proven very useful both for applications and for management functionality operating within cluster environments.

Both TCP and SCTP are general purpose protocols, in the sense that they can be used safely over the Internet as well as within a closed cluster. This virtual advantage is also their major weakness: they require funtionality and header space to deal with situations that will

never happen, or only infrequently, within clusters.

2.2 Assumptions

The TIPC design is based on the following assumptions, empirically known to be valid within most clusters.

- Most messages cross only one direct hop.
- Transfer time for most messages is short.
- Most messages are passed over intra-cluster connections.
- Packet loss rate is normally low; retransmission is infrequent.
- Available bandwidth and memory volume is normally high.
- For all relevant bearers packets are checksummed by hardware.
- The number of inter-communicating nodes is relatively static and limited at any moment in time.
- Security is a less crucial issue in closed clusters than on the Internet.

Because of the above one can use a simple, traffic-driven, fixed-size sliding window protocol located at the signalling link level, rather than a timer-driven transport level protocol. This in turn gives a lot of other advantages, such as earlier release of transmission buffers, earlier packet loss detection and retransmission, and earlier detection of node unavailability, only to mention some. Of course, situations with long transfer delays, high loss rates, long messages, security issues, etc. must be dealt with as well, but rather from the viewpoint of being exceptions than as the general rule.

3 Five-Layer Network Topology

From a TIPC viewpoint the network is organized in a five-layer structure (Figure 2).

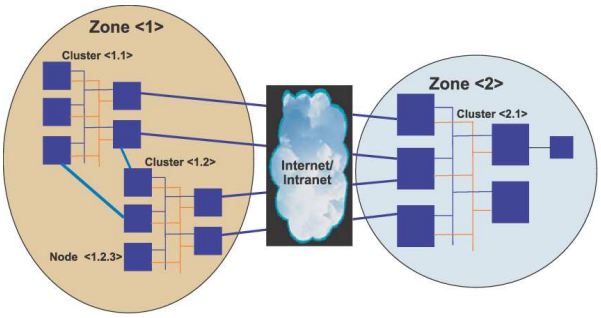


Figure 2: *TIPC Network Topology*

The top level is the *TIPC network*. This is the ensemble of all computers (nodes) interconnected via TIPC, i.e., the domain where any node can reach any other node by using a TIPC network address. A TIPC network is distringuished from other networks by its *network identity*, a 32-bit value that is known by all nodes.

The next level in the hierarchy is an entity called *zone*. This "cluster of clusters" is the maximum scope of location transparency within a network, i.e., the domain where any process can reach any other process by using a functional address rather than a network addresses.

The third level is what we call the *cluster*. This is a group of nodes interconnected all-to-all via one or two TIPC links.

The fourth level is the individual *system node*, or just *node*. There may be up to 2047 system nodes in a cluster.

The lowest level is the *slave node*. Slave nodes provide the same properties regarding location transparency and availability as system nodes, but they don't need full physical connectivity

to the rest of the cluster. One link to one system node is sufficient, although there may be more for redundancy reasons.

All entities within a TIPC network are accessed using a TIPC network address, a 32-bit value subdivided into a zone, cluster, and node field. This address is internally mapped to the address type for the communication media actually used, e.g., an Ethernet address or an IP-address/port number tuplet.

4 Location-Transparent Functional Addressing

To present a cluster as one computer, the addressing scheme used must hide the physical location of a requested service to its users. To achieve this, TIPC provides a functional address type, called *port name*, to be used both for connectionless messaging and connection setup calls. Binding a socket to a port name corresponds to binding it to a port number in other protocols, except that the port name is unique and has validity for the whole cluster, not only the local node. A caller wanting to set up a connection needs only to specify this address, and the TIPC internal translation service ensures that the request ends up in the right socket, on the right node.

A port name consists of two 32-bit fields. The first field is called the *name type* and typically identifies a certain service type or function. The second field is the *name instance* and is used as a key for accessing a certain instance of the requested service. This address structure gives excellent support for both service partitioning and service load sharing.

Further support for service partitioning is provided by an address type called *port name sequence*. This is a three-integer structure defining a range of port names, i.e., a name type plus the lower and upper boundary of the instance range. By allowing a socket to bind to a sequence, instead of just an individual port name, it is possible to partition a service's scope of responsibility into sub-ranges, without having to create a vast number of sockets to do so.

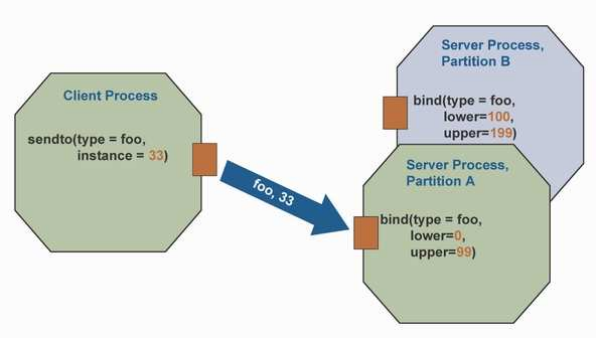


Figure 3: *Functional Addressing*

This addressing scheme is illustrated by the example in Figure 3. Two processes, partition A and partition B of the service *foo*, bind their sockets to the port name sequences *[foo,0,99]* and *[foo,100,199]* respectively (*foo* represents the name type part of the sequence). A process wanting to send a message to instance number 33 of that service, uses the port name *[foo,33]* as destination address. The TIPC name translation function will find that the indicated instance is within the range bound to by partition A, and directs the message to A's socket.

There are very few limitations on how name sequences may be bound to sockets. One may bind many different sequences, or many instances of the same sequence, to the same socket, to different sockets on the same node, or to different sockets anywhere in the cluster.

4.1 Binding Scope

Although complete location transparency is desirable and sufficient for most applications,

there must be ways to control this property for those who may need to do so. Hence, when binding a name sequence to a socket, it's possible to qualify it with a *binding scope* parameter, indicating how far the knowledge of the binding should be distributed in the network. The typical behavior is to spread it to the nodes in the binder's cluster, but it is possible to extend the scope to the whole zone, or limit it to the local node.

4.2 Lookup Domain

Similarly, a client may indicate a *lookup domain* for a message or connection setup request. This is a TIPC network address not only indicating where the lookup, i.e., the translation from a port name to socket address, should first be done, but implicitly even the lookup algorithm to be used.

Two such algorithms are available: 1) *round-robin* lookup is used when the lookup domain is non-zero and there is more than one matching server. Internally TIPC selects the server from a circular list; which root entry is stepped between each lookup. 2) *Closest-first* lookup is used when the lookup domain is zero. Here, the translation is always performed at the client's node and will first look for a matching socket on the local node. If none such is found, the algorithm will successively look for matches elsewhere in the cluster and finally in the whole zone.

5 Reliable Functional Multicast

Functional addressing is also used to provide a *reliable multicast* service. If the sender of a message indicates a port name sequence instead of a port name as destination, a replica of the message is sent to all sockets bound to a name sequence fully or partially overlapping with that sequence (Figure 4).

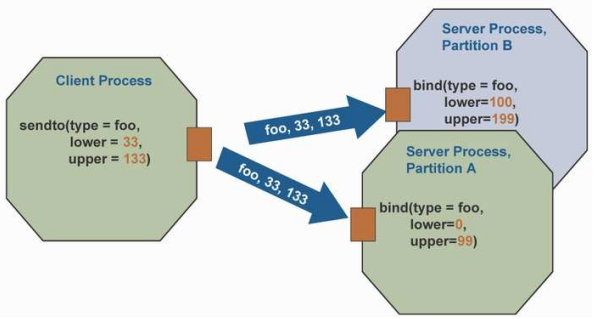


Figure 4: *Reliable Functional Multicast*

Only one replica of the message is sent to each identified target port, even if it is bound to more than one matching sequence. Whenever possible, this function will make use of the multicast/broadcast properties of the carrying media. In such cases, reliability is ensured by a special *reliable cluster broadcast* [4][5] protocol implemented internally in TIPC.

6 Name Translation Table

Translation from port name to socket addresses is performed transparently and on-the-fly via an internal translation table, replicated on each node. When a socket is bound to a port name sequence, a corresponding table entry is distributed to all nodes within the binding scope, i.e., the local cluster in most cases.

7 Topology Services

TIPC also provides a mechanism for inquiring or subscribing for the availability of port names or ranges of port names.

7.1 Functional Topology Service

This *functional topology service* is built on and uses the contents of the local instance of the name translation table.

To access this service, a user makes a blocking or nonblocking request to TIPC, asking it to indicate when a name sequence within the requested range is bound to or unbound. The request is associated with a timer, giving the duration of the subscription. A timer value of zero causes the call to return or issue a subscription event immediately, making it a pure inquiry, while a value of -1 makes it stay forever, indicating every change pertaining to the requested name sequence.

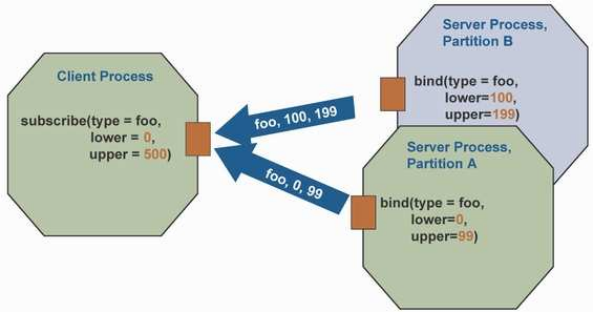


Figure 5: *Functional Topology Subscription*

Figure 5 illustrates this service: If the client process (see also example in Figure 3) wants to syncronize itself with the servers before starting any communication he issues a *subscribe()* call to TIPC, telling it to indicate when a server overlapping with the subscribed range becomes available. Since both ranges of partition A and B are within the given range *[foo,0,500]*, the client will receive two such indications, informing about the exact range of the new bindings. If there is only a partial overlap, e.g., if the client should subscribe for *[foo,0,150]* instead, he will only be informed about the actual overlap, i.e., *[foo,100,150]* for partition B.

7.2 Physical Topology Service

The physical network topology may be considered a special case of the functional topology, and can be kept track of in the same way. Hence, to subscribe for the availability/disappearance of a specific node, a group of nodes, or a whole cluster, the user specifies a dedicated port name sequence, representing this function and the range of nodes he wants to subscribe for. A special name type (zero) is used for this purpose, while the lower and upper boundaries are represented by TIPC network addresses—as described earlier, those are in reality 32-bit numbers.

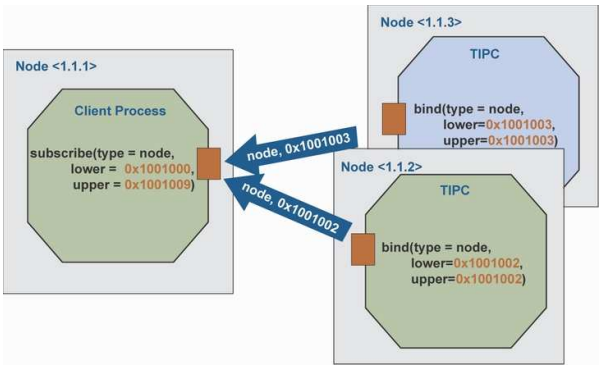


Figure 6: *Physical Topology Subscription*

In the example in Figure 6, the client process subscribes for the node range [0,9] within zone number 1, cluster number 1. Hence, when node <1.1.3> (i.e., zone 1, cluster 1, node 3) establishes a link to the client's node, the client will immediately be informed about this. For this particular service, TIPC will by itself bind/unbind the corresponding port name as soon as it discovers or loses contact with a node.

8 Lightweight Connections

The number of active user connections within a big cluster may be extremely large, and each cluster node must be able to establish and terminate thousands of such connections per second.

8.1 Simple Setup/Shutdown

To deal with this dynamism, TIPC connections are made very lighweight, in reality leaving the user to decide the setup/shutdown sequence. The protocol as such does not specify how connections are established and shut down, so an application caring about performance is free to use its own scheme, e.g., only exchanging payload-carrying messages.

For convenience an alternative, TCP-style connection type is also provided on Linux, with exchange of hidden protocol messages and stream-oriented data exchange.

8.2 Reactive Connections

TIPC connections are highly reactive and give the users almost immediate failure indication if anything should happen at the endpoints, or to the media between them. This is due to a connection supervision and abortion mechanism, which takes advantage of the properties of the local operating system to detect process crashes, or the status of the concerned links to detect node crashes or carrier failure. When any of this happens, a special *connection shutdown* message is spontaneously generated by TIPC and sent to the affected endpoint or endpoints, along with an appropriate error code. This error code delivered up to the user in the failure indication. In some cases, when the failure is detected due to inability to deliver a message, the original message is returned to the sender along with the error code, to further enable him to analyze the situation and take proper action. This *message rejection* mechanism is also used when connection-less messages are undeliverable.

9 Link-Level Protocol

Assuming that most clusters are relatively static in size, some of the tasks normally performed at the transport protocol level have been moved down to the signalling link level.

9.1 Link-level Retransmission

Implementing the retransmission protocol at this level has several advantages. First, it gives better resource utilization since all packets, connectionless and connection oriented, are funneled into one single packet sequence per node pair. Each packet can hence carry the acknowledge of many received packets, regardless of their origin, and we need not keep transmission buffers longer than strictly necessary. Second, packet losses can be detected and restransmission performed earlier than would otherwise be the case. Third, packet delivery and sequentiality guaranteed at the link level eliminates any need for per packet timers at the transport level—a background timer per link is sufficient to ensure those properties. As a result, we obtain a packet flow that is both smoother and more “traffic driven” than with corresponding transport level protocols, which often rely on timers to keep traffic running.

9.2 Link-level Node Supervision

Internode connectivity is also ensured at the link level. First, a background timer for each link endpoint supervises the traffic flow on the link and initiates a probing procedure if the peer is silent too long. Second, if a link is found to have failed after probing, there is a mechanism to steer its traffic over to the remaining link to the same node, if there is one.

9.3 Link-level Redundancy and Load Sharing

In fact, having two links and two carriers between each node pair is considered the normal configuration when using TIPC, as it eliminates any single point of failure in the communication service. The failover procedure used on such occations is completely transparent to the users, and complies to the same QOS as is guaranteed by each individual link: no message losses, no duplicates, and in-sequence delivery. The relationship between dual links is configurable; while full load sharing is the default behavior, an active-standby scheme is also supported.

Detection time for a failed link, and consequently for a crashed node, is configurable and is by default set to 1500 ms in the current implementation.

10 Automatic Neighbour Detection

Signalling links may be configured manually, but this is a tedious task if the size of a cluster runs up to dozens or even hundreds of nodes. Therefore, TIPC uses a designated neighbour detection protocol to establish links between nodes. Within a cluster this protocol is very simple. Each starting node uses the multicast or broadcast capability of the carrying media to tell about its existence, and expects a corresponding unicast response from all nodes recognizing it as part of the cluster.

Between clusters, both multicast and a unicast "pilot" link may be used, and results in a link pattern where each node in one cluster has links to a configurable (default two) number of nodes in the other cluster.

11 Performance

The performance figures we have are from the Linux-2.4 version of TIPC. We have not yet been able to do code optimizations and corresponding measurements on the Linux-2.6 version.

Performance was measured by letting a set of 16 process pairs on two nodes exchange messages in a ping-pong like manner at full speed. This ensures that the CPUs always runs at 100% load, and we can assume that almost all execution time is spent on transferring TIPC messages. We measured the time it took to exchange a message of a certain size 16 X 10 000 000 times, and divided the obtained value with number of messages. The result gives pure CPU execution time per message, automatically excluding latency times on the network and in the OS's sceduling queues, which is anyway the same for all protocols. For comparison, a similar measurement sequence was done for TCP, on the same OS and hardware.

Table 1 shows measured execution time for transferring a message process-to-process between two 750 Mhz Pentium III based nodes. The communication media used was two parallel 100 Mb Fast Ethernet switches.

Msg Size [bytes]	TIPC [μs]	TCP [μs]
64	25	38
256	29	42
1024	44	52
4096	176	178
16384	704	716
65408	3200	2800

Table 1: *Inter Node Execution time (send + receive) for TIPC and TCP messages*

The overall result shows that TIPC is around 35% faster than TCP for inter-node messages

smaller than Ethernet MTU, while performance is about the same for larger messages. A similar measurement, where all processes were kept on the same node, showed that TIPC is about four times faster (6 μs vs 25 μs) than TCP for 64 byte intra-node messages; the difference decreasing linearly with message size. At 64 Kbyte messages performance was even here almost the same.

12 Implementation

12.1 Source Code

The latest implementation on Linux is available as a source code package of 12,500 lines of C-code from [1]. It compiles into a loadable module of 167 Kbyte for the Linux-2.6 kernel, and it requires no kernel patches to be installed. This version, just as an earlier one for Linux-2.4, is stable, but still has some limitations. Most notably, only single-cluster communication is supported for now; it is not possible to set up links between nodes in different clusters or different zones.

12.2 Standardization

Open Source Development Lab (OSDL) has defined TIPC as a cornerstone in their Carrier Grade Linux (CGL) strategy, and people from OSDL are contributing actively to the code. TIPC meet several Priority 1 requirements and many Priority 2 requirements in the clustering specifications of Carrier Grade Linux version 2.0 [7]. Within IETF, the ForCES Work Group is considering TIPC to be used as transport protocol between forwarding and control elements in distributed routers. An IETF-draft [4] with a complete specification was presented for the WG at IETF-59 for this purpose.

12.3 Roadmap

The goal is to have TIPC accepted as an integrated part of the Linux kernel in future releases (2.7/2.8). Before the end of 2004, we also want to have it accepted as the preferred protocol for intra cluster transport of the ForCES protocol. Also, before the end of this year, we plan to have developed full support for inter-cluster and inter-zone communication, as well as a redesigned slave node communication framework.

13 Conclusion

Within Ericsson, TIPC has proven to be a very useful toolbox for design of high-availability clusters. It is our hope that this experience will be repeated by others now as the potential of advanced clustering is becoming more widely recognized.

References

[1] TIPC code and documentation
`http://tipc.sourceforge.net`

[2] Postel, J., Transmission Control Protocol, RFC 793
`http://www.ietf.org/rfc/rfc0793.txt`

[3] Stream Control Transmission Protocol, RFC 2960
`http://www.ietf.org/rfc/rfc2960.txt`

[4] Maloy, J., Transparent Inter Process Communication, IETF Draft
`http://www.ietf.org/internet-drafts/draft-maloy-tipc-00.txt`

[5] Guo Min: TIPC Reliable Multicast Design

http://www.linux-ericsson.
ca/papers/tipc-multicast/
tipc_multicast.pdf

[6] Maloy, J., Make Clustering Easy With TIPC, Linux World Journal April 2004
http:
//www.linux.ericsson.ca/
papers/tipc_lwm/index.shtml

[7] OSDL CGL Requirement Definition 2.0
http://www.osdl.org/lab_
activities/carrier_grade_
linux/documents.html

Object-based Reverse Mapping

Dave McCracken
IBM
dmccr@us.ibm.com

Abstract

Physical to virtual translation of user addresses (reverse mapping) has long been sought after to improve the pageout algorithms of the VM. An implementation was added to 2.6 that uses back pointers from each page to its mapping (pte chains). While pte chains do work, they add significant spaceoverhead and significant time overhead during page mapping/unmapping and fork/exit.

I will describe an alternative method of reverse mapping based on the object each page belongs to. I will discuss the partial implementation I did last year as well as the work done by Hugh Dickins and Andrea Arcangelli to complete it. I will describe the current implementations, their relative strengths and weaknesses, and what plans if any there are for solutions to the remaining issues.

1 Introduction

Up through version 2.4, the Linux® kernel had no mechanism for translating physical addresses to user virtual addresses, commonly called reverse mapping, or rmap. This meant it was not possible for the memory management subsystem to point to a physical page and remove all its mappings. There was a mechanism that walked through each process's mappings and selected pages to unmap. Only after all a page's mappings were removed could it be selected for pageout.

Many in the memory management community considered this very inefficient. Page aging and removal could be made much more efficient if the page could be directly unmapped when it was ready to be removed. Some form of rmap was clearly needed for this to work.

2 PTE Chains

Rik van Riel implemented an rmap mechanism that added a chain of pointers to each page back to all its mappings, commonly called `pte_chains`. It works by adding a linked list to the control structure for each physical page (struct page) which points to all the page table entries that map that page. His code was accepted into mainline early in the 2.5 development cycle.

Once this rmap implementation was in place the page aging and removal algorithm was changed to use it, streamlining the code and allowing better tuning.

One negative to the `pte_chain` implementation was a significant performance cost to `fork`, `exec`, and `exit`. The cost to these functions was related to the amount of memory mapped to the process, but was close to an order of magnitude worse.

A second cost was space. In its original form `pte_chains` cost two pointers per mapping. An optimization eliminated the extra structure for singly-mapped pages and another optimiza-

tion added multiple pointers per list entry, but the space taken by the `pte_chain` structures was still significant.

3 A Brief History of Object-based Rmap

Processes do not really map memory one page at a time. They map a range of data from an offset within some object (usually a file) to a range of addresses. The virtual addresses of all pages within that range can be calculated from their offset in that object and the base mapping address of the range.

The kernel has the information to do object-based reverse mapping for files. Each `struct page` for a file has an offset and a pointer to a `struct address_space`, which is the base anchor for all memory associated with a file. Every time a range of data from that file is mapped to a process, a `vm_area_struct` or `vma` is created. The `vma` contains the virtual address of the mapping and the base offset within the file. It is then added to a linked list of all `vma`s in the `address_space` for that file.

The remaining problem in the kernel is anonymous memory. Blocks of anonymous memory have `vma`s but these `vma`s are not connected to any common object that can be used for reverse mapping.

3.1 Partial Object-based Rmap

Given this information, last year I did a sample implementation of object-based rmap for files, but left the `pte_chain` implementation in place for anonymous memory. It works by following the pointer in the `struct page` to the `struct address_space`, then walking the linked list of `vma`s to find all that contain the page. A simple calculation then determines the virtual address of that page and a page table walk finds the page table entry.

This implementation recovers the performance of `fork`, `exec`, and `exit` and eliminates the space penalty used by `pte_chain` structures. It introduces a performance penalty when it walks the linked list of `vma`s, but this is incurred by the page aging code instead of the application code. It could still be significant, however, since it rises linearly with the number of times any part of the file is mapped while with `pte_chains` the cost rises linearly with the number of times that page is mapped.

3.2 First Cut at Full Object-based Rmap

Hugh Dickins took my implemenation and extended it to handle anonymous mappings, eliminating `pte_chains` entirely. He did this by creating an `anonmm` object for each process that all anonymous pages belong to. All `anonmm` structures are linked together by `fork`. A new `anonmm` structure is allocated on `exec`. The offset stored in `struct page` is the virtual address of the page, while the object pointer points to an `anonmm` that the page is mapped in.

Finding all mappings of a page is simple. The pointer in `struct page` is followed to the `anonmm` chain, which is then walked looking for mappings of that page at the virtual address specified in the offset.

Hugh's initial patch ignored the problem of shared anonymous pages that were remapped by an `mremap` call. The problem with `mremap` is that it allows an anonymous page to be at different virtual addresses in different processes, but there is only one offset for the page.

After some initial discussion among the community, both Hugh and I moved on to other things.

3.3 A Second Cut at Full Object-based Rmap

In February of this year Andrea Arcangeli began to investigate what could be done about the problems of `pte_chains`. He took my partial object-based rmap patch and implemented his own solution for anonymous memory, called `anon_vma`.

The basic mechanism of `anon_vma` is the addition of an `anon_vma` structure linked to each `vma` that has anonymous pages. The `anon_vma` structure has a linked list of all `vma`s that map that anonymous range. The pointer in `struct page` points to the `anon_vma` and the index is the offset into the current mapping.

An advantage of Andrea's `anon_vma` structure is that it solves the mremap problem that the `anonmm` structure did not. Since the offset stored in each page is relative to the base of the `vma` that maps it, the region can be remapped without changing the offset. However, since `vma`s can be merged, it is not an an absolutely painless solution.

4 Advancements All Around

In response to Andrea's patch, Hugh resumed work on his `anonmm` patch. Prompted by a discussion among the community and an approach suggested by Linus, Hugh implemented a simple scheme for handling the remap case. For each page, if there is only one reference, that page can simply have its offset changed. If the page is shared, a copy is forced and the new unshared page is mapped at the new address. Since all anonymous pages are already copy-on-write, it is likely that the page would be written to eventually and the copy taken. It is possible that some read-only pages might be duplicated, but to date there is no evidence that any code actually remaps shared read-only anonymous pages.

5 The `vma` List Problem

All these implementations still include the original implementation for file pages, including the need to walk the linked list of `vma`s attached to the `address_space` structure. This has been identified as a possible performance issue for massively mapped files, though few if any real-life examples have been found. A few optimizations have been tried, including sorting the list by start address and making a two level list based on start and end address. Both these solutions share the problem that adding or modifying a `vma` is fairly expensive and holds the associated lock for a long time.

A recent contribution by Rajesh Venkatasubramanian is the use of a `prio_tree`, which is similar to a radix tree but supports sorting objects by both start and end addresses. It adds some complexity to the `vma` list but greatly reduces the potential performance impact of a large number of mappings.

6 The `remap_file_pages` Problem

While object-based rmap appears relatively simple, there is one new feature that greatly complicates the problem. This feature is `remap_file_pages`.

The `remap_file_pages` system call was introduced during the 2.5 development cycle. It works on a range of shared memory mapped from a file, and allows an application to change the memory range to map a different offset within that file. This is done without modifying the `vma` describing the mapping. This means the offsets specified within the `vma` are now

wrong. Since the `address_space` pointer and offset within the `page` structure are intact, the page can still be mapped back to its place in the file, but it is no longer possible to use this information to find its virtual mappings. The `vma` is called a `nonlinear vma` and is put on a special list within the `address_space`.

Andrea and Hugh have provided two different solutions to the problem of what to do when a nonlinear page is called to be unmapped. Andrea's solution is the more draconian in that it walks the list of nonlinear `vma`s and unmaps all pages in them until the page in question has no more mappings. Hugh's solution only unmaps a fixed number of nonlinear pages and makes no attempt to unmap the actual page passed in.

7 Release Status

As of the date this was written, Hugh has been submitting incremental rmap changes to Andrew Morton for the -mm tree over the past couple of months. The early submissions were primarily cleanup, but later patches included first my partial object-based rmap implementation followed by his `anonmm` implementation, which completely removed the `pte_chain` code.

Hugh has just submitted a final set of patches to Andrew that removes his `anonmm` implementation and replaces it with Andrea's `anon_vma` implementation.

The general expectation among the VM developer community is that once this code has been adequately tested in the -mm tree that it will replace the existing `pte_chain` implementation in mainline 2.6.

Legal Statement

This paper represents the views of the author and does not necessarily represent the view of IBM.

The World of OpenOffice

Michael Meeks
Novell, Inc.
mmeeks@novell.com

Abstract

In this talk I will present some of the issues facing OpenOffice.org, particularly related to: performance, interoperability, buildability, ABI / engineering and release practice. We'll look at how to build the beast, the UNO component model, and iterate a quick hack before your eyes. We'll also show some of the flash new features including the Gnome desktop integration work.

1 A friendly giant

The OpenOffice.org source base is one of the largest monolithic Free software projects in existence, even with the pre-compiled mozilla binaries for several architectures stripped out:

Project	Source bz2 (MB)
Mozilla 1.4.1	31
Linux 2.6.7	33
GNOME 2.6.2	108
OO.o 1.1.2	160

OpenOffice.org (OO.o) represents one of the largest single contributions to Free software ever. Given this, it is somewhat incredible that Sun immediately settled on a licensing scheme in that is both liberal and substantially symmetric.

OpenOffice.org is licensed under two licenses:

- `LGPL` – the familiar, and best Lesser GPL.
- `SISSL` – essentially X11 with trip-wires for malicious UNO API, and XML file format compatibility breakage.

While it is necessary to share copyright with Sun by signing the Joint Copyright Assignment (JCA)[2], the use of OO.o code in StarOffice can be considered as being achieved under the SISSL[3] provisions.

Thus there is clearly huge potential for add-ins, integration with proprietary data-feeds, macros, etc.

2 Sun's dilemma

Sun's StarOffice product substantially consists of the OpenOffice.org core, as seen in public CVS, with the addition of a few extra proprietary modules. While this means that all the latest bug fixes are available in public CVS, it creates a number of frustrating artificial problems:

2.1 Release Engineering

- `minor release cycles` – there is a correct separation of commercial updates; of around once per quarter; thus this tends to be the frequency of minor OO.o releases regardless of bugginess.
- `release patch-size` – there is a fixed upper-bound on the size of a cus-

tomer patch download, thus ABI alterations in low-level libraries which would have a large knock-on effect, are forbidden.

- `ultra conservatism` – since customer updates are infrequent there is little incentive to back-port fixes to the stable branch; so many, trivial but high-impact fixes don't make it.
- `major release cycles` – for reasons unknown StarOffice works on an 18-month release cycle, so—at times (given freezes, etc.), it is possible to punt a feature / fix by nearly 2 years.

Clearly many of these problems make the OO.o development process somewhat cumbersome.

2.2 Portability Engineering

In contrast to many Free software project, StarOffice and hence OO.o, is designed to run on a broad spectrum of operating systems and versions. By contrast, e.g. GNOME applications, would typically require the latest version of GNOME to run.

This creates a number of interesting, hard-core engineering issues, and shows up the true state of Linux as a robust platform for ISVs.

For example, for font discovery much Linux software will link to the pleasant fontconfig library, and use purely client-side font rendering. OO.o in contrast has to run on older (or newer) platforms where there is either no fontconfig install, or it has a changed ABI, or it is badly configured. Thus the OO.o font discovery method uses the following heuristics:

- `fontconfig` – since this may not be available, we try to `dlopen` it, hook out various symbols, and extract a simple list of font filenames.
- `chkfontpath` – Red Hat, and others once shipped this tool which dumps a list of font paths; we try to `popen` and parse the output.
- `hard-coded paths` – various directories such as `/usr/X11R6/lib/X11/fonts/truetype` are known to be a good bet, and are scanned for fonts, including several language specific variants.
- `X server query` – the X server is queried to see what it can do, and a load of XLFDs are parsed.
- `internal fonts` – whatever internal fonts, and font-metric files we distribute are added to the mix.

Naturally, after doing all this work, we build a OO.o specific cache of much of the information, to accelerate subsequent startup.

This heavily engineered approach is not constrained to any one API-set, or technology—so, e.g., OO.o will attempt to use either lpr or cups for printing in a dynamic fashion.

Even glibc problems show up in Figure 1.

In addition, the cross-platform nature of OO.o and the unpredictability of the Linux feature-set (particularly the C++ ABI), leads to a large number of software packages being included inside the OO.o build itself. Thus, a stock OOo would include it's own compiles of (at least): `python`, `freetype`, `zlib`, `expat`, `libdb`, `NAS`, `neon`, `curl`, `sane`, `myspell`, `Xrender`.

As is probably obvious, this level of old platform support, and dependency aversion is hard to get enthusiastic about.

```
typedef struct {
    struct { long status; int spinlock; } sem_lock;
    int sem_value;
    void *sem_waiting;
} glibc_21_sem_t;
    /*
     * XXX this a hack of course. since sizeof(sem_t) changed
     * from glibc-2.0.7 to glibc-2.1.x, we have to allocate the
     * larger of both XXX
     */
#ifdef LINUX
    if (sizeof(glibc_21_sem_t) > sizeof(sem_t))
        Semaphore = malloc(sizeof(glibc_21_sem_t));
    else
#endif
        Semaphore = malloc(sizeof(sem_t));
}
```

Figure 1: compatibility with old `glibc` versions

3 Community Issues

In addition to these unusual constraints, the OO.o project is encumbered by acute tooling and collaboration inadequacies.

Perhaps the most serious problem, is that it appears CVS was not designed with 200+ MB of source / binaries in mind. Thus, even basic operations, such as a `cvs tag` can take up to a couple of hours, and are frequently blocked by robots slowly traversing the repository.

Secondarily, the collab.net SourceCast system adds a level of bureaucracy, and lack of responsiveness which when combined with being totally un-fixable makes for an unnecessarily painful experience. It seems likely that SourceCast is ideal for the use of existing, established Free software projects, or even newly formed projects—but it stumbles with OO.o. Furthermore, using closed software for Open Source collaboration is an intrinsically interesting decision.

4 The other side of the coin

4.1 `http://ooo.ximian.com/`

To make up for the existing inadequate web-tools, and documentation we provide several ‘external’ tools of interest.

- `hackers guide` – a Linux focused, hackers guide on how to build, iterate, and some basics of the OO.o code structure.
- `LXR/Bonsai` – basic web tools without which navigating the OO.o source is substantially more difficult.
- `bug filing` – a gateway that demangles the curious user-focused issue filing process, and allows bug filing directly against given code modules.
- `Planet OO.o` – the obligatory RSS aggregator.

4.2 ooo-build

The process of productising OO.o into a Linux package is filled with pain; so to amortise this

a collaboration has coalesced between various Linux vendors: Novell ne Ximian, Debian, Red Hat, SuSE, Ark, and PLD Linux around *ooo-build*.

ooo-build provides a growing set of useful patches many of which may arrive in OO.o in many months time; indeed all our work is intended to go up-stream into OO.o. We also provide a simple patch sub-setting system, to allow vendors to select a suitable set of patches.

Many of the features associated with ooo-build are desktop integration, system integration, and GUI cleanup pieces; e.g.:

- attractive new icons
- native-widget rendering
- GNOME-VFS integration
- ergonomic & aesthetic fixes
- system library usage

The ooo-build wrapper is also intended to make OO.o substantially easier to compile with a familiar `./configure; ./download; make; make install` process.

5 Performance

Performance is an area ripe for substantial improvement in OO.o, however, poor performance is caused by many factors, and identifying the most important of these is not always easy.

5.1 Linking

The linker has a very hard time linking OO.o, and while this can be reduced by pre-linking, the architecture of OO.o—whereby the majority of the code is in shared libraries required not by the main binary—but by other shared component libraries, linked at run-time.

Ulrich's analysis of OO.o [1] shows that 20,000 relocations are performed during startup, which combined with lookups across multiple libraries gives 1,700,000 string comparisons to startup. The sheer size of the symbol tables and the lack of locality of reference in the linking process causes much of this work to fall outside the processors' cache—giving abnormally poor performance.

5.2 C++ issues

Some features of C++ exacerbate the problems of large symbol tables, and poor startup performance. The stripping / re-working of static initialisers has helped accelerate performance—these being replaced with a thread-safe late instantiation based on accessor method local static variables.

C++ is a very symbol-hungry language—particularly with respect to virtual functions, which create an unnecessary burden (Figure 2). Virtual functions, despite resolving to a simple function pointer export a symbol, which is referred to directly to chain to parent implementations. While of course this can often be resolved away at link time, in a cross-library situation it would perhaps be more efficient to dereference a parent vtable function pointer.

Similarly, since in theory at least, a single class can be implemented across multiple shared objects, even 'private:' methods export symbols.

In addition to these problems, a more proactive approach to pruning old, and redundant code has been adopted in the development branch, to reduce code footprint, and symbol count.

```
class Foo : public Baa {
    virtual void VFunc();
  private:
    void ExportsSymbol();
};
...
void VFunc()
{
    ...
    Baa::VFunc();
}
```

Figure 2: C++ virtual functions

5.3 Binary filter code

To shrink the OO.o footprint, a large chunk of creaking binary format code has been extracted, along with compatible chunks of the core. This code pre-dating the XML file formats scattered the process of serialisation across the code, and resulted in a complex, hard-to-maintain and increasingly irrelevant maintenance problem. In OO.o 2.0 it will be used only on the rare occasions it is necessary as a binary to XML filter.

5.4 system libraries

Shrinking the large number of internal libraries, on Linux systems, and increasing the number of libraries shared with the system is an important part of performance improvement in 2.0. It clearly makes little sense to have an internal gtk+ library when the system version is ABI compatible, and better maintained.

Using system libraries—e.g., neon—also reduces the pain of handling security updates in the built-in libraries.

5.5 mmap performance

Possibly the most significant speedup in the 1.0 to 1.1 transition was the process of forcing as much of the OO.o code into memory before attempting to run it. This gave a very noticeable win; this was implemented in a simple fashion with mmap, and a loop reading a byte from each page. Ideally of course, the underling operating system would be able to do better here.

6 Interoperability

In a world where a tiny fraction of people are using Free software, the ability to share documents in a loss-less fashion with other people is crucial to the adoption of OO.o.

Much work has been done in this area for 2.0, of particular note the row-limit in calc has been raised to that of Excel, and much work has been done on form controls.

There are also exciting developments in VBA interoperability. OO.o provides a VBA-like language: StarBasic, and by devious means it has been possible to extract VBA text from Office files for some while. Office for performance reasons however stores VBA in 3 forms: an SRP stream, a compiled form, and a compressed text form. Since these are authoritative in that order (the text providing only a final fallback), it was thought that effective export would entail reverse engineering at least the the compiled form.

However in recent time, yet more devious means have been discovered to export macros as text to Excel and have them run transparently to the user. This it turns out is the foundation of macro interoperability between Office versions 97 through XP. Thus work is ongoing to improve the macro support so crucial for effective Excel interoperability.

7 Desktop integration

Much of the work of ooo-build has been adopted in one form or another up-stream for 2.0, giving the prospect of a highly desktop integrated OO.o experience out-of-the-box.

To achieve this, the lowest levels of OO.o's cross-platform abstraction: the Visual Class Library (VCL) have been virtualised, and now the main-loop, and top-level windows on a GNOME system are handled by the gtk+ toolkit. In order to avoid a complete re-write of the widget system—we use a simplified theming system that virtualises only the rendering of widgets, allowing basic widgets to match the look of the rest of the desktop.

Similarly main-loop integration makes things such as integrating the gtk+ file-selector and other GNOME dialogs fairly simple. The main-loop integration was made substantially more painful by the mis-match between the recursive OO.o toolkit lock, and the non-recursive gtk+ lock. In order to reconcile these and provide a single, comprehensible locking pattern—after considerable thought we added hooks to gtk+ to allow a shared (recursive) lock to be used. This makes gtk+ use in OO.o virtually seamless.

8 UNO component model

OO.o provides a rich, and well documented component model, which is exported for the use of language bindings. The power of this, and its flexibility have resulted in active bindings for StarBasic, Java, and Python.

The UNO model is particularly interesting, since it consumes little overhead beyond a stock C++ virtual function call. In addition each class has associated, small compiled IDL type information. This can be used, to dynamically (at run time) construct bridges to other languages, and allow dynamic method invocation. While this adds a compiler version / ABI dependency to the OO.o core, it avoids the problem of creating stub / skeleton code which ended up consuming many MB before the dynamic approach was adopted.

9 Conclusions

OO.o provides an unusualm and particularly pathalogical case of a gigantic C++ project. This leads us to push the boundaries of the system, showing up several areas for potential improvement.

The ooo-build infrastructure provides a solid base for contributing work to OO.o in a familiar and accessible manner, and seeing the deployed results of your work quickly.

OpenOffice 2.0 will give substantial performance, code-cleanliness and interoperability improvements, in addition to many new features.

References

[1] Ulrich Drepper. How to write shared libraries, 2004. `http://people.redhat.com/drepper/dsohowto.pdf`.

[2] Sun Microsystems, Inc. Joint copyright assignment. `http://www.openoffice.org/licenses/jca.pdf`.

[3] Sun Microsystems, Inc. Sun industry standards source license. `http://www.openoffice.org/licenses/sissl_license.html`.

TCPfying the Poor Cousins

Arnaldo Carvalho de Melo
Conectiva S.A.
acme@conectiva.com.br
http://advogato.org/person/acme
http://www.conectiva.com.br

Abstract

In this paper I will describe the work I am doing on the Linux networking infrastructure, with emphasis on cleaning the code, but with important "side effects" like reduction of core structures already saving over 600 bytes on UDP sockets all over the net in 2.5/2.6 (tcp, etc.), elimination of data dependencies, reduction of the non-mainstream network families maintenance cost by making them use code that now is in `net/ipv4` but can be moved to `net/core`, leaving only the really ipv4-specific code and making LLC use it as a proof of concept (work done in my net-exp tree, pending submission).

TCP code becomes used by the poor cousins, they appreciate that!

1 How This Started

Making IPX uptodate with regards to advances in the core networking infrastructure, to kill `deliver_to_old_ones`, i.e., special cases in the core kernel for protocols that hasn't been converted to shared skbs and multithreading.

In the process I noticed several areas where code was replicated or used a different, older framework, due to the evolution of the core networking infrastructure.

Also de experience of porting the NetBEUI and LLC code released as GPL by Procom Inc. from the 2.0 Linux kernel networking infrastructure to 2.4 and then to 2.5/6, working on a BSD sockets API for `PF_LLC`, initially contributed by Jay Schullist was instrumental in realising the existing similarities in the infrastructure needs required by several protocol families.

2 TCP/IP Evolves Faster

Most of the attention is given, of course, to TCP/IP, and in the process new infrastructure is created, with TCP/IP using it at first and sometimes leaving things like the `deliver_to_old_ones` function to simulate the previously existing big networking lock and the `SOCKOPS_WRAPPED` macro, to allow the other protocol families to continue working, hoping that their maintainers do the necessary work, but this sometimes doesn't happen for a long time.

In other cases code is added to TCP/IP that, upon further inspection, could be moved to `net/core` and be useful for the other protocol families.

Doing this factorization will help make these improvements to TCP/IP be taken advantage of by the other protocol families and will help in realising the ultimate goal of keep the proto-

col families code with just what is completely specific.

3 Trimming `struct sock`

In 2.4, `struct sock` has a big fat union that has most of the private data for each protocol family, so when any change had to be done to a specific protocol family the layout of `struct sock` would change, generating unnecessary recompilation of most of the network related code in the kernel.

In 2.6 this has changed and `struct sock` nowadays is mostly free of details specific to network protocol families.

In the process two ways were devised to store the network protocol private area, one for protocols that have stringent performance requirements, like TCP/IP, using per-protocol slab caches and another one, simpler, that allows protocol families to just allocate a chunk of memory and store its pointer in the `struct sock` member `sk_protinfo`. As most stacks now use helper macros to access its private area, the eventual switch to the slab cache approach is easily done.

With this in place the footprint of the `struct sock`, that was of about 1280 bytes on a UP machine in 2.4 to 308 bytes for the generic sock slabcache in 2.6, with the `tcp_sock` slabcache using 1004 bytes, `udp_sock` slabcache using 484 bytes and finally the `unix_sock` (`PF_UNIX` sockets) using just 356 bytes.

This changes also resulted in a performance gain in the establishment of connections, as was verified with the `lmbench` tool.

Another related change was to diminish the data dependency among `struct sock` and `struct tcp_tw_bucket`, that is a "mini socket" used to represent TCP connections in the `TIME_WAIT` state. To accomplish this, `struct sock_common` was introduced, that is the minimal required set of members common to these structs. With this data layout we will certainly avoid bugs introduced when changing only one of the structs, like has happened at least once to my knowledge.

4 Using `list.h` in the Networking Code

With the advent of the hashed lists (`struct hlist_node`) it turned out to be useful to make the networking code follow the general kernel trend of using the `linux/list.h` macros, replacing the ad-hoc lists present in the networking code.

The work consisted of introducing a set of helper macros to handle `struct sock` list handling, namely `sk_add_node` and `sk_del_node_init`, and bind list variants.

These functions also bump the reference count for the socket, something that was not being done by some protocols, that have since been converted to use this new set of helper macros, thus fixing some bugs in the process.

It should also be noted that `hlist` started using prefetch as part of the process of convincing David Miller, the Linux Networking maintainer, to accept such changes. Performance gains are an important technique in getting code-cleaning patches accepted.

5 Socket Timers Manipulation Helpers

Another area that received attention was the socket timers manipulation routines, that in some protocols aren't always bumping the ref-

erence count as they should and do in the Bluetooth and TCP/IP code.

To abstract this handling the `sk_reset_timer` and `sk_stop_timer` functions were introduced recently to do the `timer_list` handling and deal with `struct sock` reference counting.

6 Factorization of net/ipv4 Code

In the past Alan Cox worked on having datagram code that could be shared among several network families shared at the `net/core/datagram.c` file, moving chunks of code out of the UDP implementation.

Now with this work I'm trying to do the same with the stream code, now moving chunks of TCP code to the core infrastructure.

Initial steps are just moving code around, like `tcp_eat_skb`, that became `sk_eat_skb`; `tcp_data_wait` became `sk_wait_data`; and here we see something interesting, namely the fact that this function correctly sets the `SOCK_ASYNC_WAITDATA` bits in the `struct socket` flags member, something that some protocols aren't doing now but will as soon as they start using `sk_wait_data`.

In my `net-experimental` tree I have introduced some new members to the `struct sock` member `sk_prot`, allowing both TCP and LLC to use common `stream_sendmsg` and `stream_sendpage` functions, that are generalizations of `tcp_sendmsg` and `tcp_sendpage`. Further work is needed to fully determine the performance implications of such changes, but no noticeable performance drop or stability problems have been verified in using this patched kernel in my main machine for over a month.

7 BSD Sockets Layer

There is some duplication of work at the BSD sockets level among the network protocol families implementation. Trying to reduce the code required to implement a protocol family is being investigated, with some proofs-of-concept already implemented, where the functions now used for TCP/IP are being shared with LLC.

The idea here is to to reduce the protocol-specific implementation to just that, i.e., what is absolutely specific to each protocol.

Perhaps this will make it easier to stack protocols, allowing combinations that are possible in other kernels but not on Linux right now.

The extra function pointers in `sk->sk_prot` probably won't be a problem because they will make it possible to eliminate `sock->proto_ops` by calling directly the `sk->sk_prot` functions.

8 Future Developments

With this newly common infrastructure, it may be possible to add features like network async I/O to all protocols. More sharing will be investigated, trying to avoid pitfalls that appeared in similar work done in other kernel subsystems.

9 Conclusion

Looking every other year at how core infrastructures evolve and how the implementations of subsystems attached to those infrastructure evolve is something that should be done, paying off in terms of code clarity, reduction of the cost of maintaining code that has come out of mainstream but are still used in lots of legacy setups.

Another eventual benefit gained is the performance one, as making the code clear and more general is not incompatible with having fast code.

Reuse the code, Luke.

10 Acknowledgments

I'd like to thank David S. Miller for all the support he gives me in continuing this work, reviewing my patches, and providing much valued words of wisdom. And my respect for Andi Kleen, for helping me out in my childhood as a Linux networking wannabe hacker, Alan Cox for throwing me the Procom NetBEUI stack, that was fun! And nasty as well. As well as all the fine kernel networking hackers that spare some of their time to comment on my ideas.

IPv6 IPsec and Mobile IPv6 implementation of Linux

Kazunori MIYAZAWA
USAGI Project/Yokogawa Electric Corporation
kazunori@miyazawa.org

Masahide NAKAMURA
USAGI Project/Hitachi Communication Technologies, Ltd
masahide_nakamura@hitachi-com.co.jp

Abstract

USAGI Project [8] has improved Linux IPv6 [1] stack. IPv6 IPsec is one of the products of our efforts. Linux IPsec [6] stack is implemented based on XFRM architecture which is introduced in linux-2.5. We design and implement Mobile IPv6 (MIPv6) [4] Stack on the architecture. MIPv6 uses IPsec for its secure signaling. Accordingly IPv6 IPsec and MIPv6 closely cooperate each other. In this paper we describe the architecture and how they work.

1 Introduction

IPv6 is the next version of an Internet Protocol. The protocol was developed against IPv4 address exhaustion. It was developed for not only spreading address space but improving some features such as plug and play, aggregatable routing architecture, IPsec native support and smooth transition.

IPsec provides security services which are integrity, authentication, anti-replay attacks and confidentiality. Because IPsec is mandatory in IPv6 specification, we must implement IPsec to conform to it.

MIPv6 provides all IPv6 nodes with mobility service which allows nodes to remain reachable while moving around IPv6 networks. To support mobility, We need some signaling architecture to notify movement and deliver mechanisms to assure reachability. Using MIPv6, we can keep routability to mobile node's home link address and deliver a packet to mobile node wherever it is on the network. Because IPv6 is able to process these extension headers natively, we no longer need to arrange foreign agents to all links where mobile node may move to as Mobile IPv4 does, so that IP mobility is easier to be introduce in IPv6 than IPv4.

Linux supported IPsec at version 2.5.47. However it supporting only IPv4 IPsec, we implemented IPsec stack for IPv6. Linux version 2.6 supports IPsec on both IPv6 and IPv4. XFRM architecture and stackable destination were introduced into the kernel for IPsec packet processing [7]. They can be not only for IPsec packet processing, but also general packet processing such as MIPv6. USAGI Project decided to expand the architecture to implement MIPv6.

To develop Linux MIPv6, we cooperate with GO/Core Project [2] which is proven in linux-

2.4.

2 XFRM and stackable destination

XFRM architecture is mainly consist of three structures which are xfrm_policy, xfrm_state and xfrm_tmpl. xfrm_policy corresponds to IPsec policy and xfrm_state to IPsec SA. xfrm_tmpl is intermediate structure between xfrm_policy and xfrm_state. Each IPsec policy and SA database are realized with list of the structures which are also contained hash database.

The kernel provides three interface to configure xfrm structures about IPsec. One is PF_KEY interface which is standard interface to manipulate IPsec database. another is netlink socket interface. The last is socket option interface.

Stackable destination is architecture for efficient outbound packet processing. It is a link list of dst_entry structure which is cached in xfrm_policy. To create stackable destination, the kernel linearly searches xfrm_policy with flow information for a sending packet after routing looking up. After finding xfrm_policy corresponding to the flow information, the kernel searches and gathers xfrm_state from xfrm_state database by xfrm_tmpl in the xfrm_policy. Gathering xfrm_states, the kernel builds up stackable destination and substitutes it into its own member “bundles” to cache it. Additionally xfrm_policy itself is cache in flow_cache. Therefore the kernel only needs to lookup xfrm_policy after second until xfrm_state expired.

3 IPsec

IPsec functionality is consist of packet processing and key exchanging for automatic keying. In the implementation of Linux packet processing runs in the kernel and key exchange is done by a key exchange daemon in user space.

3.1 IPsec database and packet processing

IPsec packet processing is realized with XFRM architecture and stackable destination. Outbound process is explained in previous section. With searching XFRM database and building stackable destination, the kernel gets list of dst_entry structure. To process each function which are ah6_output, esp6_output and ipcomp6_output, the kernel searches insertion point on a packet because a packet is created including IPv6 header and other extension headers before stackable destination process (Figure 1). The insertion point is before upper layer payload, fragmentable destination options header, IPsec header or fragment header. This is not efficient because the kernel searches the insertion point every time when processing one dst_entry.

Inbound process is simpler than outbound process. When packet containing AH or ESP, the kernel finds xfrm_state corresponding to received packet and keep pointers of used xfrm_state in sec_path of skb structure. After process of IP layer, the kernel checks the packet correctly processed with comparing sec_path and xfrm_policy which is searched with flow information of the packet (Figure 2).

3.2 Interface for user and IKEd

Current linux kernel provides users with PF_KEY interface, which however is specified only for IPsec SA interface and it needs some extension to configure IPsec policy. Because this extension is not standardized, there are some different extensions and it prevents compatibility of IKEd. Linux adopts the extension which is compatible with KAME [5] so that racoon is the IKEd for linux. Racoon is originally product of KAME project and its could not compile on Linux. Fortunately

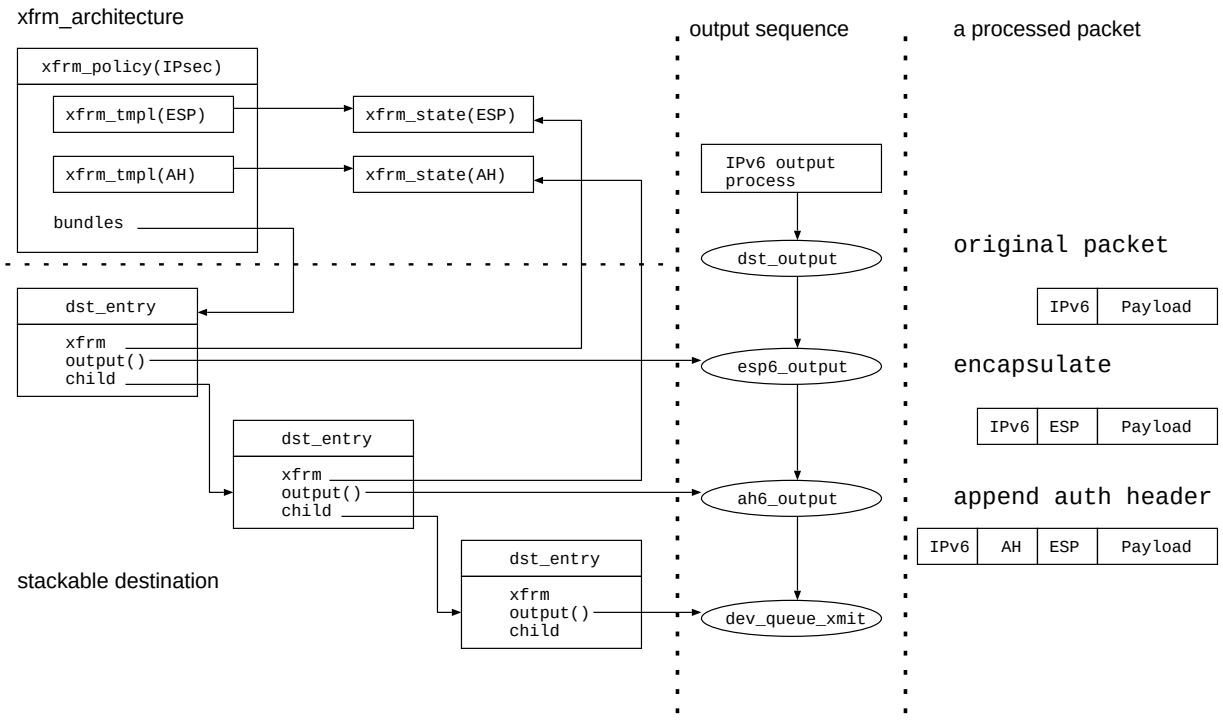


Figure 1: IPsec output process

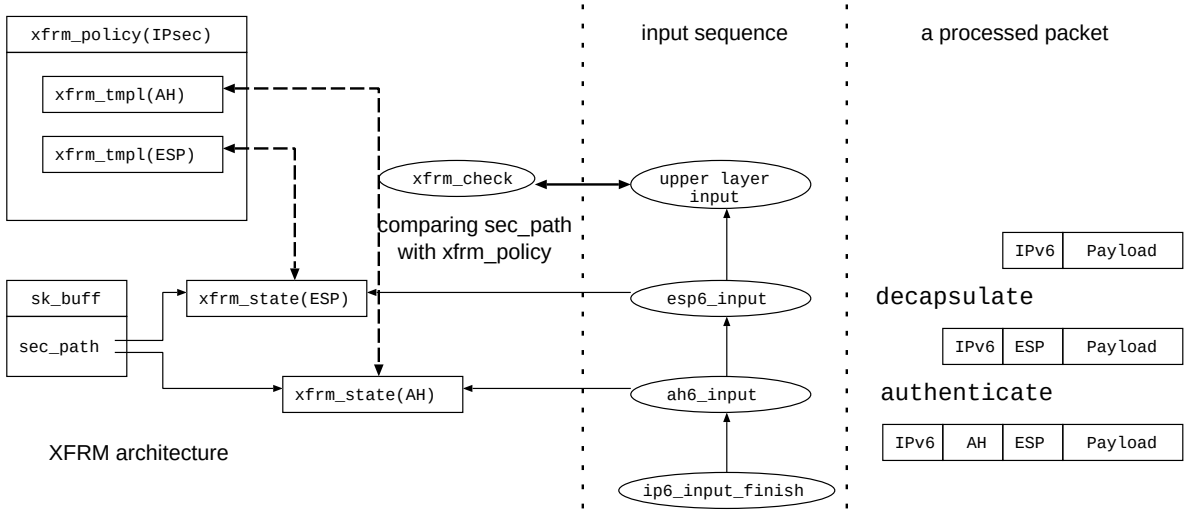


Figure 2: IPsec input process

ported racoon which is provided by ipsec-tools project [3] is available.

4 Mobile IPv6

4.1 Mobile IPv6

In MIPv6, nodes are classified into 3 types. One is a Mobile Node (MN) which moves in the IPv6 Internet bringing its home address (HoA) assigned in a home link which is a base of mobility and in which there is a home agent. Home agent (HA) is another type of node which is a router and manages MN's addresses and supports its signaling and ensures reachability. The other is a correspondent node (CN) which is a node communicating with a MN. CN may be either mobile or stationary.

When MN in a foreign link, it uses a care-of address (CoA) which is the address of a foreign link. MIPv6 accordingly needs to manage relationship between CoA and HoA. A MN sends a packet including HoA in an extension header from CoA.

MIPv6 appends two extension headers and one option for destination options header. Mobility Header (MH) is an extension header for signaling to manage binding cache which is a address list for optimized routing. Type2 routing header (RT2) which is different from routing header in RFC2460 effects destination address in IPv6 header and realizes direct routing according to binding cache. Home Address Option (HAO) is an option carried by destination options header to contain HoA which is an address of a MN in home link and swapped with CoA. HAO effects source address in IPv6 header.

We describe an outline of the procedure taking as an example that MN making binding cache on HA and communicating CN after MN moving to a foreign link (Figure 3). This procedure is divided two steps. First is making IPv6 over IPv6 tunnel between MN and HA (1-4). After this step, HoA of MN becomes routable and MN is able to communicate with all nodes by using HoA via HA through the tunnel. Second is route optimization between MN and CN because MN always communicating via HA (5-8), a packet goes through a superfluous route and communication uses more network resource.

1. MN sends a Binding Update (BU) to HA.
2. HA updates a binding cache and returns Binding Acknowledgment (BA) to MN.
3. MN updates a binding update list.
4. At this time, there is a tunnel between MN and HA.
5. MN sends HoTI to CN through the tunnel and CoTI to CN directly from CoA.
6. CN keeps contents of HoTI and CoTI. CN returns HoT via HA and CoT to CoA.
7. When MN receives HoT and CoT, MN sends BU to CN and updates its own binding list.
8. Then MN and CN have binding between HoA and CoA. They communicate directly with appending HAO and RT2 to packets. They have an optimized route.

4.2 Implementation

We design MIPv6 in Linux consisted with two part. One is packet processing for RT2 and HAO in the kernel and the other is MIPv6 daemon (MIPd) to handle the signaling and manage binding cache and binding update list. It is similar to separation of packet process and IKEd in IPsec.

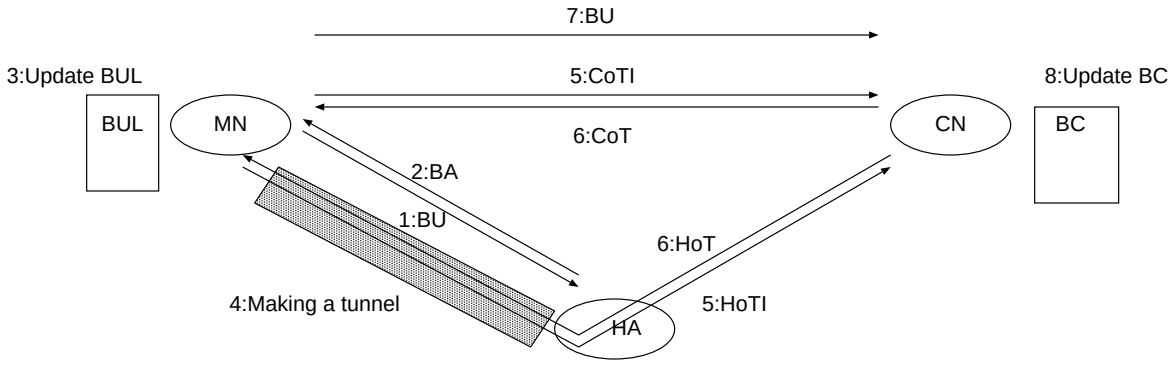


Figure 3: MIPv6 procedure outline

Packet processing for MIPv6 is realized with XFRM and stackable destination architecture, because they are general way to process a packet which matches some selector. Using XFRM, we can avoid to implement duplicate functionality in the kernel. MIPv6 needs to manage a binding cache which specifies an MN address on the network on CN and HA. It also needs to manage a binding update list which is list of sending binding update request for CN on MN. We have two choices to implement this functionality in the kernel or userland. Because we should implement functionalities in userland if it is possible, we consider to basically implement it in userland. Implementing in userland brings us advantages which are easier extension its functionality than implementing in the kernel and reducing the kernel size.

Our MIPd's roles are

- processing a signaling message including an error message
- managing xfrm_policy and xfrm_state of MIPv6 in the kernel through the netlink
- managing binding cache and binding update list
- moving detection and changing CoA when MIPd running on MN

4.3 XFRM operation

In this section, we describe MIPd XFRM operation relating each nodes state with an example which is a phase of binding update to HA and making tunnel for routability. It is called home registration. At first, we initialize MN and HA to send and receive binding message. On MN MIPd sets a xfrm_policy which allows an outbound packet from HoA to HA, proto MH, and type BU with appending HAO and a xfrm_state which appends HOA with CoA to a packet from HoA to HA and including MH of BU. It also set xfrm_policy to receive BA, the policy which allows an inbound packet from HA to HoA including MH of BA with appending RT2 and the inbound xfrm_state which processes RT2. Because MIPd on HA can not expect the source address of BU from MN, it sets a xfrm_policy which allows an inbound packet from Any to HA with MH of BU if it has HAO. It also set xfrm_state which processes HAO included in a packet from ANY to HA with MH of BU. See Figure 6:INITIALIZE.

MIPd on MN sends BU to HA, the packet matches with the xfrm_policy and process with the xfrm_state which appends HAO destination option and swap a source address in IPv6 header with a CoA. HA received the BU from MN. In the kernel the packet matching the xfrm_state, the kernel swaps addresses. Then

MIPd on HA receives BU and updates a binding cache. MIPd configures xfrm_policy and xfrm_state for route optimization with high priority. See Figure 6:Routing Optimization.

At this moment, route optimization is available for all packets between MN and HA. It also sets up a tunnel between MN and HA. After some xfrm_policy and xfrm_state configuration it returns BA with RT2. The kernel of MN receives BA with RT2 and processes it with the inbound xfrm_state and throws up BA packet to MIPd. MIPd on MN updates a binding update list and sets up the tunnel. Each nodes has totally 6 policies at the end of registration.

5 Cooperation of IPsec and MIPv6

MIPv6 uses IPsec for its secure signaling between MN and HA. Our design uses XFRM and stackable destination for both IPsec and MIPv6. MIPv6 needs two kind of IPsec SA one is a transport mode SA which is used for signaling. The other is a tunnel mode SA which is used instead of IPv6 over IPv6 tunnel. We consider two steps to implement MIPv6 with IPsec about IPesc policy and SA management. At first, we implement MIPd to not only manage xfrm_policy and xfrm_state of MIPv6 but also IPsec and a xfrm_policy for MIPv6 holds both MIPv6 and IPsec xfrm_tmpl. This implementation has a couple of issues. One is separation of management of xfrm_policy and xfrm_state of IPsec into MIPv6 and ordinary IPsec. Another issue is interaction between the kernel and IKE daemon. xfrm_policy including a xfrm_tmpls of Mobile IPv6 and IPsec sends a signal for only MIPd. The other is the order of xfrm_policy. When some situation such as configuration done with wrong order, a packet which would be originally applied MIPv6 and IPsec not be applied only IPsec.

For improvement, we will let the kernel hold two xfrm databases and mediate them because it is difficult to manage xfrm_tmpl in a xfrm_policy via userland interface by two management daemons and the xfrm_policies have probably different granularity (Figure 7). In current outbound process, the kernel looks up single xfrm_ policy database and gets a xfrm_policy which includes xfrm_tmpl for IPsec and xfrm_tmpl for MIPv6. However we will change the kernel to separately look up IPsec and MIPv6 xfrm databases and create temporary xfrm_policy which holds xfrm_tmpl gathered from each xfrm_policy. The list of xfrm_tmpl must be serialized as the order of packet processing. For instance, the kernel must put xfrm_state for AH at the end of the list. For inbound process, it is not so difficult, the kernel processes a packet by using xfrm_state which is searched and needs to check sec_path in skb against each xfrm_policy. To make it be efficient, the kernel should use flow_cache for inbound process.

If we could merge two policies correctly, we have another issue. MIPv6 needs two IPsec SA between NM and HA. One is a transport mode SA for signaling and the other is a tunnel mode SA for other packet. Taking outbound SA as an example, a transport mode SA is applied by the policy whose selector is from HoA to HA and protocol MH. On the other hand a tunnel mode SA is applied by the policy whose selector is from HoA to ANY and protocol ANY. The packet should be applied the transport mode SA has possibility to be applied the tunnel mode SA. We can avoid this mismatch by using priority in xfrm_policy.

racoon has a couple of issues as IKE daemon for MIPv6. One is that racoon can not handle multiple peers which have address ANY as peer's address in its configuration. When it behaves as responder on HA, the issue occurs because despite multiple peers being, each configuration has addresses from ANY to HA thus

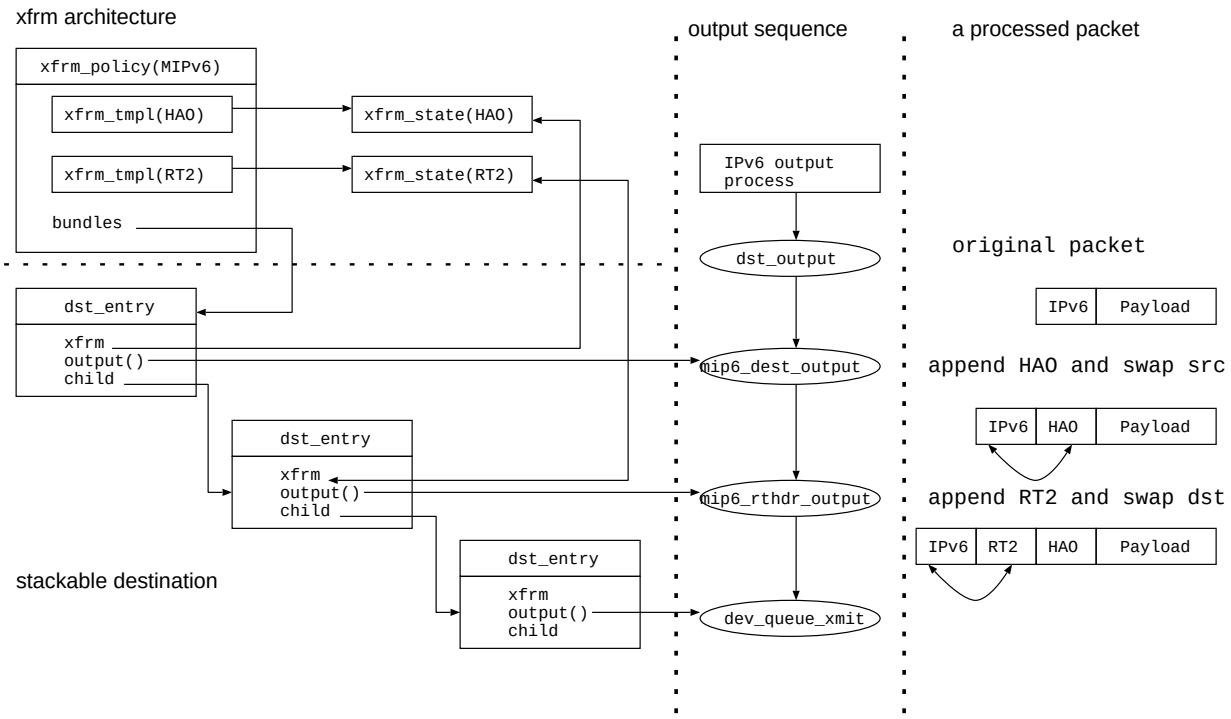


Figure 4: MIPv6 output process

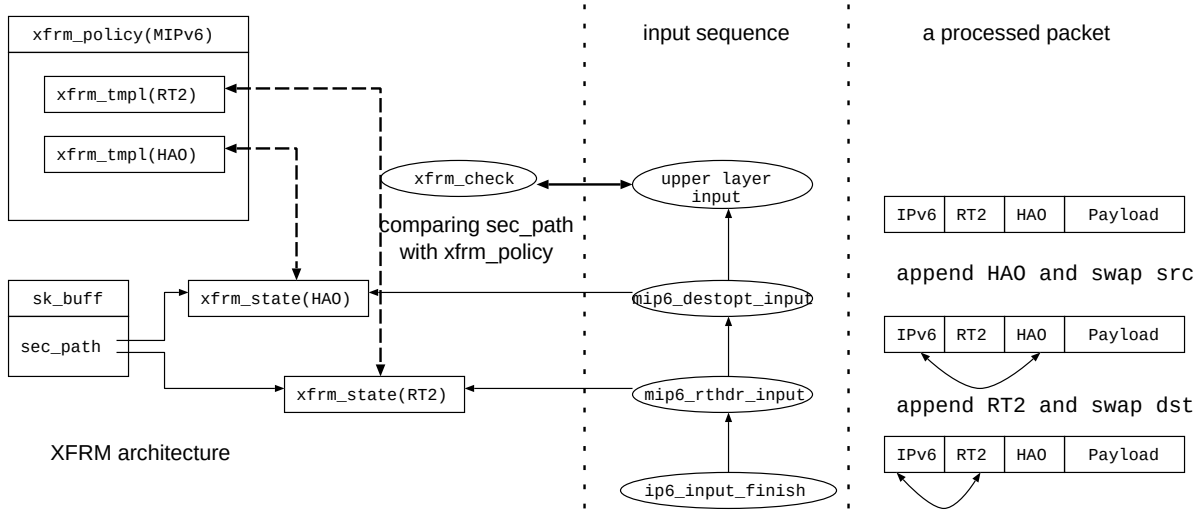


Figure 5: MIPv6 input process

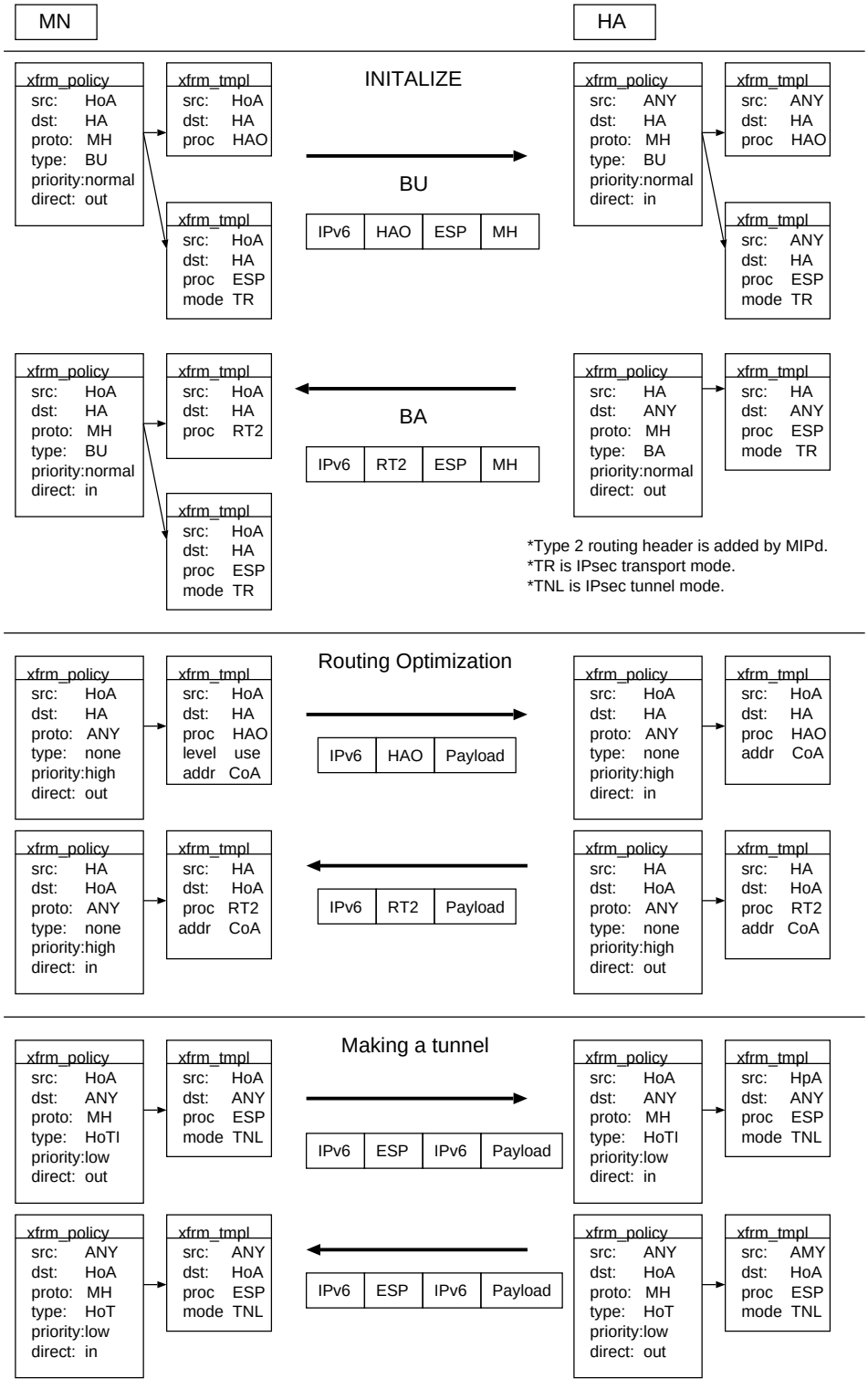


Figure 6: Binding update procedure to Home Agent

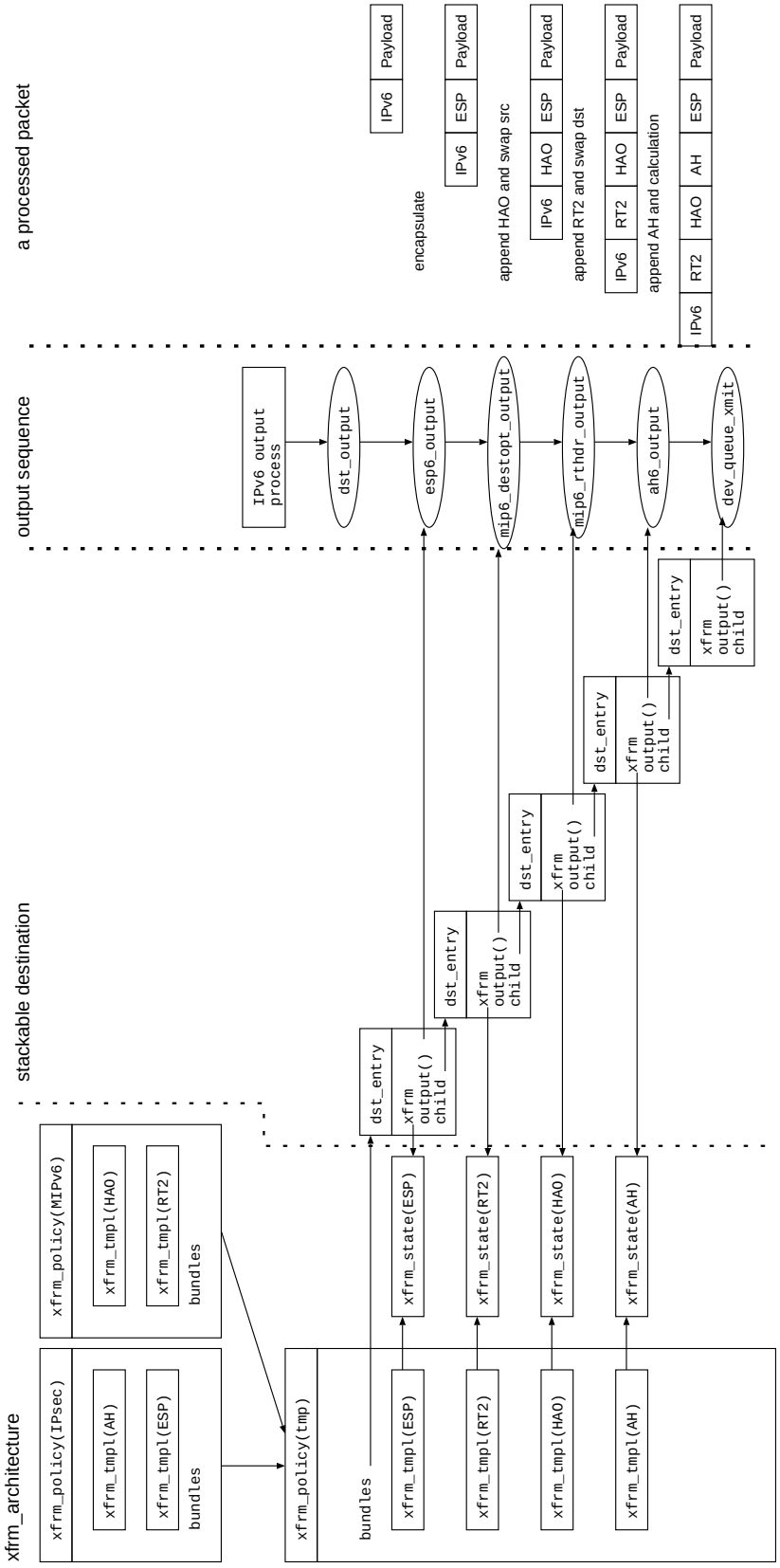


Figure 7: MIPv6 and IPsec output process

racoon can not distinct peer and fails to search proper key. The other issue is update ISAKMP SA end-point address. When MN moves, IKEs on MN and HA need to detect movement in some way and update its ISAKMP SAs because an address of those SAs is CoA. To solve these issues, we will make racoon handle the multiple peers listen netlink socket for the detection and make the kernel notify address changing via netlink socket.

6 Summary

USAGI Project implements IPv6 IPsec and MIPv6 by using XFRM and stackable destination architecture. In this paper we describe our design, implementation and issues. We also describe future design of IPv6 IPsec and MIPv6 which improves flexibility of xfrm configuration.

7 future work

Our future works about MIPv6 are

- implement our new design
- make racoon support MIPv6
- NEMO
- Multihome
- vertical hand-over

Additionally we consider that we should improve or change stackable destination itself because stackable destination runs after building a packet. Thus, IPv6 packet processing is not efficient itself because an IPv6 packet has some extension header and the order of headers is not always same as the order of process so that every process searches correct point on a packet from the head. We should improve its packet processing with keeping xfrm architecture and cache mechanism.

References

[1] S. Deering and R. Hinden. Internet Protocol, Version 6 Specification. RFC2460, December 1998.

[2] GO/Core Project. MIPL Mobile IPv6 for Linux. `http://www.mobile-ipv6.org`.

[3] IPsec Tools. IPsec Tools Web Page. `http://www.ipsec-tools.sourceforge.net/`.

[4] D. Johnson, C. Perkins, and J. Arkko. Mobility Support in IPv6. Work in Progress, June 2003.

[5] KAME Project. KAME Project Web Page. `http://www.kame.net`.

[6] S. Kent and R. Atkinson. Security Architecture for the Internet Protocol. RFC2401, November 1998.

[7] Kazunori Miyazawa, Hideaki Yoshifuji, and Yuji Sekiya. Linux IPv6 Networking—Past, Present, and Future. In *Proceedings of the Linux Symposium*, Ottawa, July 2003.

[8] USAGI Project. USAGI Project Web Page. `http://www.linux-ipv6.org`.

Getting X Off the Hardware

Keith Packard
Hewlett-Packard
Cambridge Research Laboratory
keithp@keithp.com

Abstract

The X window system is generally implemented by directly inserting hardware manipulation code into the X server. Mode selection and 2D acceleration code are often executed in user mode and directly communicate with the hardware. The current architecture provides for separate 2D and 3D acceleration code, with the 2D code executed within the X server and the 3D code directly executed by the application, partially in user space and partially in the kernel. Video mode selection remains within the X server, creating an artificial dependency for 3D graphics on the correct operation of the window system. This paper lays out an alternative structure for X within the Linux environment where the responsibility for acceleration lies entirely within the existing 3D user/kernel library, the mode selection is delegated to an external library and the X server becomes a simple application layered on top of both of these. Various technical issues related to this architecture along with a discussion of input device handling will be discussed.

1 History

The X11[SG92] server architecture was designed assuming significant operating assistance for supporting input and output devices. How that has changed over the years will inform the discussion of the design direction proposed in this paper.

1.1 Original Architecture

One of the first 2D accelerated targets for X11 was the Digital QDSS (Dragon) board. The Dragon included a 1024x768 frame buffer with 4 or 8 bits for each pixel. The frame buffer was not addressable by the CPU, rather every graphics operation was performed by the coprocessor. The Dragon board had only a single video mode supporting the monitor supplied with the machine. A primitive terminal emulator in the kernel provided the text mode necessary to boot the machine.

Graphics commands to the processor were queued to a shared DMA buffer. The X server would block in the kernel waiting for space in the buffer when full. This is similar to the architecture used by the DRI project for accelerated 3D graphics today.

Keyboard and mouse support were provided by another shared memory queue between the kernel and X server. Abstract event structures were constructed by the kernel from the raw device data, timestamped and placed in the shared queue. A file descriptor would be signalled when new data were inserted to awaken the X server, and the X server could also directly examine the queue indices which were stored in the shared segment. This low-overhead queue polling was used by the X server to check for new input after every X re-

quest was executed to reduce input latency.

The hardware sprite was handled in the kernel; its movement was directly connected with the mouse driver so that it could be moved at interrupt time, leading to a responsive pointer even in the face of high CPU load within the X server and other applications. The keyboard controller managed the transition from ASCII console mode to key-transition X mode internally; abnormal termination of the X server would leave the underlying console session working normally.

1.2 The Slippery Slope

Early Sun workstations had unaccelerated frame buffers. Like the QDSS above, they used fixed monitors and had no need to support multiple video modes. As the hardware advanced, they did actually gain programmable timing hardware, but that was not configurable from the user mode applications.

The X server simply mapped the frame buffer into its address space and manipulated the pixel values directly. Around 1990, Sun shipped the cgsix frame buffer which included an accelerator. Unlike the QDSS, the cgsix frame buffer could be mapped by the CPU, and the accelerator documentation was not published by Sun. X11R4 included support for this card as a simple dumb frame buffer. As CPU access to the frame buffer was slower than with Sun's earlier unaccelerated frame buffers, the result was a much slower display.

By disassembling the provided SunWindows driver, the author was able to construct an accelerated X driver for X11R5 entirely in user mode. This driver could not block waiting for the accelerator to finish, rather it would spin, polling the accelerator until it indicated it was idle.

Keyboard and mouse support were provided by the kernel as files from which events could be read. The lack of any shared memory mechanism to signal available input meant that the original driver would not notice input events until the X server polled the kernel, something which could take significant time. As there was no kernel support for the pointer sprite, the X server was responsible for updating it as well, leading to poor mouse tracking when the CPU was busy.

To ameliorate the poor mouse tracking, the X server was modified to receive a signal when input was present on the file descriptors and immediately process the input. When supported, the hardware sprite would also be moved at this time, leading to dramatically improved tracking performance. Still, the fact that the X server itself was responsible for connecting the mouse motion to the sprite location meant that under high CPU load, the sprite would noticeably lag the mouse.

Kernel support for the keyboard consisted of a special mode setting which would transform the keyboard from an ASCII input device to reporting raw key transition events. Because the kernel didn't track what state the keyboard was in, the X server had to carefully reset the keyboard on exit back to ASCII mode or the user would no longer be able to interact with the console.

Placing the entire graphics driver in user mode eliminated the need to write a kernel driver, but marginalized overall system performance by forcing the CPU to busy-wait for the graphics engine. Placing responsibility for manging the sprite led to poor tracking, while requiring the X server to always reset the keyboard mode frequently resulted in an unusable system when X terminated abnormally.

Fixing the kernel to address these problems was never even considered; the problems didn't prevent the system from functioning, they only

made it less than ideal.

1.3 The Dancing Bear

With widespread availability of commodity 386-based PC hardware, numerous vendors began shipping Unix (and Unix-like) operating systems for them. These originally did not include the X window system. A disparate group of users ported X to these systems without any support from the operating system vendors.

That these users managed to get X running on the early 386 hardware was an impressive feat, but that they had to do everything without any kernel support only increased the difficulty.

Early PC graphics cards were simple frame buffers as far as graphics operations went, but configuring them to generate correct video timings was far from simple. Because monitors varied greatly, each graphics card could be programmed to generate many different video timings. Incorrect timings could destroy the monitor.

Keyboard support in these early 386-based Unix systems was very much like the Sun operating system; the keyboard was essentially a serial device and could be placed in a mode which translated key transitions into ASCII or placed a mode which would report the raw bytes emitted from the keyboard.

The X server would read these raw bytes and convert them to X events. Again, there was latency here as the X server would not process them except when polling for input across all X clients and input devices. As with the Sun driver, if the X server terminated without switching the keyboard back to translated mode, it would not be usable by the console. This particular problem was eventually fixed in some kernels by adding special key sequences to reset the keyboard to translated mode.

Mouse support really was just a kernel serial driver—PS/2 mice didn't exist, and so bus and serial mice were used. The X server itself would open the device, configure the communication parameters and parse the stream of bytes. As there was no hardware sprite support, the X server would also have to draw the cursor on the screen; that operation had to be synchronized with rendering and so would be delayed until the server was idle.

Because the X server itself was managing video mode configuration, an abnormal X server termination would leave the video card misconfigured and unusable as the console. Similarly, the keyboard driver would be left in untranslated mode, so the user couldn't even operate the computer blind to reboot.

This caused the X server to assume the same reliability requirements as the operating system kernel itself; bugs in the X server would render the system just as unusable as bugs in the kernel.

1.4 The Pit of Despair

With the addition of graphics acceleration to the x86 environment, the X server extended its user-mode operations to include manipulation of the accelerator. As with the Sun GX driver described above, these drivers included no kernel support and were forced to busy-wait for the hardware.

However, unlike the GX hardware, PC graphics hardware would often tie down the PCI bus while transferring data between the CPU and the graphics card. Incorrect manipulation of the hardware would result in the PCI bus locking and the system not even responding to network or disk activity. Unlike the simple keyboard translation problem described above, this cannot be be fixed in the operating system.

Because the graphics devices had no kernel

driver support, there was no operating system management of their address space mappings. If the BIOS included with the system incorrectly mapped the graphics device, it fell to the X server to repair the PCI mapping spaces. Manipulating the PCI address configuration from a user-mode application would work only on systems without any dynamic management within the kernel.

If the machine included multiple graphics devices controlled through the standard VGA addresses, the X server would need to manipulate these PCI mappings on the fly to address the active card.

The overall goal was not to build the best system possible, but rather to make the code as portable as possible, even in the face of obviously incorrect system architecture.

1.5 A Glimmer of Hope

The Mesa project started as a software-only rasterizer for the OpenGL API. By providing a freely available implementation of this widely accepted API, people could run 3D applications on every machine, even those without custom 3D acceleration hardware. Of course, performance was a significant problem, especially as the 3D world moved from simple colored polygons to textures and complex lighting environments.

The Mesa developers started adding hardware support for the few cards for which documentation was available. At first, these were whole-screen drivers, but eventually the DRI project was started to support multiple 3D applications integrated into the X window system. Because of the desire to support secure direct rendering from multiple unprivileged applications, the DRI project had to include a kernel driver. That driver could manage device mappings, DMA and interrupt logic and even clean up the hardware when applications terminated abnormally.

The result is a system which is stable in the face of broken applications, and provides high performance and low CPU overhead.

However, the DRI environment remains reliant on the X server to manage video mode selection and basic device input.

2 Forward to the Past

Given the dramatic changes in system architecture and performance characteristics since the original user-mode X server architecture was promulgated, it makes sense to look at how the system should be constructed from the ground up. Questions about where support for each operation should live will be addressed in turn, first starting with graphics acceleration, then video mode selection and finally (and most briefly) input devices.

3 Graphics Acceleration

X has always directly accessed the lowest levels of the system to accelerate 2D graphics. Even on the QDSS, it constructed the register-level instructions within the X server itself. With the inclusion of OpenGL[SAe99] 3D graphics in some systems, the system requires two separate graphics drivers, one for the X server operating strictly in 2D mode and the other inside the GL library for 3D operations. Improvements to the 3D support have no effect on 2D performance.

As a demonstration of how effectively OpenGL can implement the existing X server graphics operations, Peter Nilsson and David Reveman implemented the Glitz library[NR04] which supports the Render[Pac01] API on top of the OpenGL API. In a few months, they managed

to provide dramatic acceleration for the Cairo graphics library[WP03] on any hardware with an OpenGL implementation. In contrast, the Render implementation within the X sample server using custom 2D drivers has never seen significant acceleration, even three and a half years after the extension was originally designed. Only a few drivers include even half-hearted attempts at acceleration.

The goal here is to have the X server use the OpenGL API for all graphics operations. Eliminating the custom 2D acceleration code will reduce the development burden. Using accelerated OpenGL drivers will provide dramatic performance improvements for important operations now ill-supported in existing X drivers. Work in this area will depend on the availability of stand-alone OpenGL drivers that work in the absence of an underlying window system. Fortunately, the Mesa project is busy developing the necessary infrastructure. Meanwhile, development can progress apace using the existing window-system dependent implementations, with the result that another X server is run just to configure the graphics hardware and set up the GL environment.

For cards without complete OpenGL acceleration, the desired goal is to provide DRI-like kernel functionality to support DMA and interrupts to enable efficient implementation of whatever useful operations the card does support. For 2D graphics, the operations needing acceleration are those limited by memory bandwidth—large area fills and copies. In particular acceleration of image composition results in dramatic performance improvements with minimal amounts of code. The spectacular amounts of code written in the past that provide modest acceleration for corner cases in the X protocol should be removed and those cases left to software to minimize driver implementation effort.

This architecture has been implemented by Eric Anholt in his kdrive-based Xati server[Anh04]. Using the existing DRI driver for the Radeon graphics card, he developed a 2D X driver with reasonable acceleration for common operations, including significant portions of the X render API. The driver uses only a small fraction of the Radeon DRI driver, a significantly smaller kernel driver would suffice for a ground-up implementation.

In summary, graphics cards should be supported in one of two ways:

1. With an OpenGL-based X server
2. With a 2D-only X server based on a simple loadable driver API.

3.1 Implications for Applications

None of the architectural decisions about the internal X server architecture change the nature of the existing X and Render APIs as the fundamental 2D interface for applications. Applications using the existing APIs will simply find them more efficient when the X server provides a better implementation for them. This means that applications needn't migrate to non-X APIs to gain access to reasonable acceleration.

However, applications that wish to use OpenGL should find a wider range of supported hardware as driver writers are given the choice of writing either an OpenGL or 2D driver, and aren't faced with the necessity of starting with a 2D driver just to support X.

In any case, use of the cairo graphics library provides insulation from this decision as it supports X and GL requiring only modest changes in initialization to select between them.

4 Video Mode Configuration

The area of video mode selection involves many different projects and interests; one significant goal of this discussion is to identify which areas are relevant to X and how those can be separated from the larger project.

4.1 Overview of the Problem

Back in 1984 when X was designed, graphics devices were fundamentally fixed in their relationship with the attached monitor. The hardware would be carefully designed to emit video timings compatible with the included monitor; there was no provision for adjusting video timings to adapt to different monitors, each video card had a single monitor connector.

Fast forward to 2004 when common video cards have two or more monitor connectors along with outputs for standard NTSC, SECAM, or PAL video formats. The desire to dynamically adjust the display environment to accommodate different use modes is well supported within the Macintosh and Microsoft environments, but the X window system has remained largely stuck with its 1984 legacy.

4.2 X Attempts to Fix Things

X servers for PC operating systems adapted to simple video mode selection by creating a 'virtual' desktop at least as large as the largest desired mode and making the current mode view a subset of that, panning the display around to keep the mouse on the screen. For users able to accept this metaphor, this provided usable, if less than ideal support. Most of the time, however, having content off of the screen which could only be reached by moving the mouse was confusing. To help address this, the X Resize and Rotate extension (RandR)[GP01] was designed to notify applications of changes in the pixel size of the screen and allow programmatic selection among available video modes.

The RandR extension solved the simple single monitor case well enough, even permitting the set of available modes to change on the fly as monitors were switched. However, it failed to address the wider problem of supporting multiple different video outputs and the dynamic manipulation of content between them.

Statically, the X server can address each video output correctly and even select between a large display spanning a collection of outputs or separate displays on each video screen. However, there is no capability to adjust these configurations dynamically, nor even to automatically adapt to detected changes in the environment.

4.3 X is Only Part of the Universe

With 2D performance no longer a significant marketing tool, graphics hardware vendors have been focusing instead on differentiating their products based on video output (and input) capabilities. This has dramatically extended the options available to the user, and increased the support necessary within the operating system.

As the suite of possible video configuration options continues to expand, it seems impossible to construct a fixed, standard X extension capable of addressing all present and future needs. Therefore, a fully capable mechanism must provide some "back door" through which display drivers and user agents can communicate information about the video environment which is not directly relevant to the window system or applications running within it.

One other problem with the current environment is that video mode selection is not a requirement unique to the X window system. Numerous other graphical systems exist which

are all dependent on this code. Currently, that is implemented separately for each video card supported by each system. The MxN combination of graphics systems and video cards means that only a few systems have support for a wide range of video cards. Support for systems aside from X is pretty sparse.

4.4 Who's in Charge Here, Anyway?

X itself places relatively modest demands on the system. The X server needs to be aware of what video cards are available, what video modes are available for each card and how to select the current mode. Within that mode there may be a wealth of information that is not relevant to the X server; it really only needs to know the pixel dimensions of each frame buffer, the physical dimension of pixels on each monitor and the geometric relationship among monitors. Details about which video port are in use, or how the various ports relate to the frame buffer are not important. Information about video input mechanisms are even less relevant.

As the X server need have no way of interpreting the complexity of the video mode environment, it should have no role in managing it. Rather, an external system should assume complete control and let the X server interact in its own simple way.

This external system could be implemented partially in the kernel and partially in usermode. Doing this would allow the kernel to share the same logic for video mode selection during boot time for systems which don't automatically configure the video card suitably on power-on. In addition, alternate graphics systems would be able to share the same API for their own video mode configuration.

5 Input Device Support

In days of yore, the X environment supported exactly one kind of mouse and one (perhaps of an internationalized family) keyboard. Sadly, this is no longer the case. The wealth of available input devices has caused no small trouble in X configuration and management. Add to that the relative failure of the X Input extension to gain widespread acceptance in applications and the current environment is relegated to emulating that available in 1984.

5.1 Uniform Device Access

The first problem to attack is that of the current hodgepodge device support where the X server itself is responsible for parsing the raw bytestreams coming from the disparate input devices. Fortunately, the kernel has already solved that problem—the new `/dev/input`-based drivers provide a uniform description of devices and standard interface to all. Converting the X server over to those interfaces is straightforward.

However, the `/dev/input/mice` interface has a significant advantage in todays world; it unifies all mouse devices into a single stream so that the X server doesn't have to deal with devices that come and go. So, to switch input mechanisms, the X server must first learn to deal with that.

5.2 Hotplug and HAL

Mice (and even keyboards) can be easily attached and detached from the machine. With USB, the system is even automatically notified about the coming and going of devices. What is missing here is a way of getting that notification delivered to the X server, having the X server connect to the new device (when appropriate), notifying X applications about the

availability of the new device and integrating the devices events into the core pointer or keyboard event stream.

The Hardware Abstraction Layer (HAL)[Zeu] project is designed to act as an intermediary between the Linux Hotplug system and applications interested in following the state of devices connected to the machine. By interposing this mechanism, the complexity of discovering and selecting input devices for the X server can be moved into a separate system, leaving the X server with only the code necessary to read events from the devices specified by the HAL. One open question is whether this should be done by a direct connection between the X server and the HAL daemon or whether an X client could listen to HAL and transmit device state changes through the X protocol to the X server.

One additional change needed is to extend the X Input Extension to include notification of new and departed devices. That extension already permits the list of available devices to change over time, all that it lacks is the mechanism to notify applications when that occurs. Inside the X server implementation, the extension is in for some significantly more challenging changes as the current codebase assumes that the set of available devices is fixed at server initialization time.

6 Migrating Devices

With X was developed, each display consisted of a single keyboard and mouse along with a fixed set of monitors. That collection was used for a single login session, and the input devices never moved. All of that has now changed; input devices come and go, computers get plugged into video projectors, multiple users login to the same display. The dynamic nature of the modern environment requires some changes to the X protocol in the form of new or modified extensions.

6.1 Whose Mouse Is This?

Input devices are generally located in physical proximity to the related output device. In a system with multiple output devices and multiple input devices, there is no existing mechanism to identify which device is where. Perhaps some future hardware advance will include geographic information along with the bus topology.

The best we can probably do for now is to provide a mechanism to encode in the HAL database the logical grouping of input and output devices. That way the X server would receive from the HAL the set of devices to use at startup time and then accept ongoing changes in that as the system was reconfigured.

One problem with this simplistic approach is that it doesn't permit the migration of input devices from one grouping to another; one can easily imagine the user holding a wireless pointing device to attempt to interact with the "wrong" display. Some mechanism for dynamically reconfiguring the association database will need to be included.

6.2 Hotplugging Video Hardware

While most systems have no ability to add or remove graphics cards, it's not unheard of—many handheld computers support CF video adapters. On the other hand, nearly all systems do support "hotplugging" of the actual display device or devices. Many can even detect the presence or absence of a monitor enabling true auto-detection and automatic reconfiguration.

When a new monitor is connected, the X server needs to adapt its configuration to include it. In the case where the set of physical screens are

gathered together as a single logical screen, the change can be reflected by resizing that single screen as supported by the RandR extension. However, if each physical screen is exposed to applications as a separate logical screen, then the X server must somehow adapt to the presence of a new screen and report that information to applications. This will require an extension.

In terms of the existing X server implementation, the changes are rather more dramatic. Again, it has some deep-seated assumptions that the set of hardware under its control will not change after startup. Fixing these will keep developers entertained for some time.

6.3 Virtual Terminal Switching

One capability Linux has had for a long time is the ability to rapidly switch among multiple sessions with "virtual terminals." The X server itself uses this to preserve a system console, running on a separate terminal ensures that the system console can be viewed by simply switching to the appropriate virtual terminal. Given this, multiple X servers can be started on the same hardware, each one on a different virtual terminal and rapidly switched among.

The virtual terminal mechanism manages only the primary graphics device and the system keyboard. Management of other graphics and input devices is purely by convention. The result is that multiple simultaneous X sessions are not easily supported by the standard build of the X server. The X server targeted at a non-primary graphics device needs to avoid configuring the virtual terminal. However, this also eliminates the ability for that device to support multiple sessions; there cannot be virtual terminal switching on a device which is not associated with any virtual terminals.

With the HAL providing some indication of which devices should be affiliated into a single session configuration, the X server can at least select them appropriately. Similarly, the X server should be able to detect which device is the console keyboard and manage virtual terminals from there. Whether the kernel needs to add support for virtual terminals on the other graphics/keyboard devices is not something X needs to answer.

The final problem is that of other input devices; when switching virtual terminals, the X server conventionally drops its connection to the other input devices, presuming that whatever other program is about to run will want to use the same ones. While that does work, it leaves open the possibility that an error in the X server will leave these devices connected and deny other applications access to them. Perhaps it would be better if the kernel was involved in the process and directing input among multiple consumers automatically as VT affiliation changed.

7 Conclusion

Adapting the X window system to work effectively and competently in the modern environment will take some significant changes in architecture, however throughout this process existing applications will continue to operate largely unaffected. If this were not true, the fundamental motivation for the ongoing existence of the window system would be in doubt.

Migrating responsibility for device management out of the X server and back where it belongs inside the kernel will allow for improvements in system stability, power management and correct operation in a dynamic environment. Performance of the resulting system should improve as the kernel can take better advantage of the hardware than is possible in user

mode.

Sharing graphics acceleration between 2D and 3D applications will reduce the effort needed to support new graphics hardware. Migrating the video mode selection will allow all graphics systems to take advantage of it. This should permit some interesting exploration in system architecture.

Significant work remains in defining the precise architecture of the kernel video drivers; these drivers need to support console operations, frame buffer device access and DRI (or other) 3D acceleration. Common memory allocation mechanism seem necessary, along with figuring out a reasonable division of labor between kernel and user mode for video mode selection.

Other work remains to resolve conflicts over sharing devices among multiple sessions and creating a mechanism for associating specific input and output devices together.

The resulting system regains much of the flavor of the original X11 server architecture. The overall picture of a system which provides hardware support at the right level in the architecture appears to have wide support among the relevant projects making the future prospects bright.

References

[Anh04] Eric Anholt. High Performance X Servers in the Kdrive Architecture. In *FREENIX Track, 2004 Usenix Annual Technical Conference*, Boston, MA, July 2004. USENIX.

[GP01] Jim Gettys and Keith Packard. The X Resize and Rotate Extension - RandR. In *FREENIX Track, 2001 Usenix Annual Technical Conference*, Boston, MA, June 2001. USENIX.

[NR04] Peter Nilsson and David Reveman. Glitz: Hardware Accelerated Image Compositing using OpenGL. In *FREENIX Track, 2004 Usenix Annual Technical Conference*, Boston, MA, July 2004. USENIX.

[Pac01] Keith Packard. Design and Implementation of the X Rendering Extension. In *FREENIX Track, 2001 Usenix Annual Technical Conference*, Boston, MA, June 2001. USENIX.

[SAe99] Mark Segal, Kurt Akeley, and Jon Leach (ed). *The OpenGL Graphics System: A Specification*. SGI, 1999.

[SG92] Robert W. Scheifler and James Gettys. *X Window System*. Digital Press, third edition, 1992.

[WP03] Carl Worth and Keith Packard. Xr: Cross-device Rendering for Vector Graphics. In *Proceedings of the Ottawa Linux Symposium*, Ottawa, ON, July 2003. OLS.

[Zeu] David Zeuthen. HAL Specification 0.2. `http://freedesktop.org/~david/hal-0.2/spec/hal-spec.html`.

Linux 2.6 performance improvement through readahead optimization

Ram Pai
IBM Corporation
linuxram@us.ibm.com

Badari Pulavarty
IBM Corporation
badari@us.ibm.com

Mingming Cao
IBM Corporation
mcao@us.ibm.com

Abstract

Readahead design is one of the crucial aspects of filesystem performance. In this paper, we analyze and identify the bottlenecks in the redesigned Linux 2.6 readahead code. Through various benchmarks we identify that 2.6 readahead design handles database workloads inefficiently. We discuss various improvements made to the 2.6 readahead design and their performance implications. These modifications resulted in impressive performance improvements ranging from 25%–100% with various benchmarks. We also take a closer look at our modified 2.6 readahead algorithm and discuss current issues and future improvements.

1 Introduction

Consider an application that reads data sequentially in some fixed-size chunks. The kernel reads data sufficiently enough to satisfy the request from the backing storage and hands it over to the application. In the meantime the application ends up waiting for the data to arrive from the backing store. The next request also takes the same amount of time. This is quite inefficient. What if the kernel anticipated the future requests and cached more data? If it could do so, the next read request could be satisfied much faster, decreasing the overall read latency.

Like all other operating systems, Linux uses this technique called *readahead* to improve read throughput. Although readahead is a great mechanism for improving sequential reads, it can hurt the system performance if used blindly for random reads.

We studied the performance of the readahead algorithm implemented in 2.6.0 and noticed the following behavior for large random read requests.

1. reads smaller chunks of data many times, instead of reading the required size chunk of data once.
2. reads more data than required and hence wasted resources.

In Section 2, we discuss the readahead algorithm implemented in 2.6 and identify and fix the inefficient behavior. We explain the performance benefits achieved through these fixes in Section 3. Finally, we list the limitations of our fixes in Section 4.

2 Readahead Algorithm in 2.6

2.1 Goal

Our initial investigation showed the performance on Linux 2.6 of the Decision Support System (DSS) benchmark on filesystem was

about 58% of the same benchmark run on raw devices. Note that the DSS workload is characterized by large-size random reads. In general, other micro-benchmarks like rawio-bench and aio-stress showed degraded performance with random workloads. The suboptimal readahead behavior contributed significantly toward degraded performance. With these inputs, we set the following goals.

1. Exceed the performance of 2.4 large random workloads.
2. DSS workload on filesystem performs at least 75% as well as the same on raw devices.
3. Maintain or exceed sequential read performance.

2.2 Introduction to the 2.6 readahead algorithm

Figure 1 presents the behavior of 2.6.0 readahead. The `current_window` holds pages that satisfy the current requests. The `readahead_window` holds pages that satisfy the anticipated future request. As more page requests are satisfied by the `current_window` the estimated size of the next `readahead_window` expands. And if page requests miss the `current_window` the estimated size of the `readahead_window` shrinks. As soon as the read request cross `current_window` boundary and steps into the first page of the `readahead_window`, the `readahead_window` becomes the `current_window` and the `readahead_window` is reset. However, if the requested page misses any page in the `current_window` and also the first page in the `readahead_window`, both the `current_window` and the `readahead_window` are reset and a new set of pages are read into the `current_window`. The number of pages read in the current window depends upon the estimated size of the `readahead_window`. If the estimated size of the `readahead_window` drop down to zero, the algorithm stops reading ahead, and enters the slow-read mode till page request pattern become sufficiently contiguous. Once the request pattern become sufficiently contiguous the algorithm re-enters into readahead-mode.

2.3 Optimization For Random Workload

We developed a user-level simulator program that mimicked the behavior of the above readahead algorithm. Using this program we studied the read patterns generated by the algorithm in response to the application's read request pattern.

In the next few subsections we identify the bottlenecks, provide fixes and then explain the results of the fix. As a running example we use a read sequence consisting of 100 random read-requests each of size 16 pages.

2.3.1 First Miss

Using the above read pattern, we noticed that the readahead algorithm generated 1600 requests of size one page. The algorithm penalized the application by shutting down readahead immediately, for not reading from the beginning of the file. It is sub-optimal to assume that application's read pattern is random, just because it did not read the file from the beginning. The offending code is at line 16 in Figure 1. Once shut down, the slow-read mode made readahead to not resume since the `current_window` never becomes large enough. For the ext2/ext3 filesystem, the `current_window` must become 32 pages large, for readahead to resume. Since the application's requests were all 16 pages large, the `current_window` never opened. We re-

```
1 for each page in the current request
2 do
3       if readahead is shutdown
4       then  // read one page at a time (SLOW-READ MODE)
5         if requested page is next to the previously requested page
6         then
7              open the current_window by one more page
8         else
9              close the current_window entirely
10        fi

11        if the current_window opens up by maximum readahead_size
12        then
13             activate readahead  // enter READAHEAD-MODE
14        fi
15        read in the requested page

        else  // read many pages at a time (READAHEAD MODE)
16        if this is the first read request and is for the first page
                          of this open file instance
17              set the estimated readahead_size to half the size of
18                      maximum readahead_size
19        fi

20        if the requested page is within the current_window
21              increase the estimated readahead_size by 2
22              ensure that this size does not exceed maximum
23                       readahead_size
24        else
25              decrease the estimated readahead_size by 2
26              if this estimate becomes zero, shutdown readahead
27        fi

28        if the requested page is the first page in the readahead_window
29        then
30               move the pages in the readahead_window to the
31                     current_window and reset the readahead_window
32               continue
33        fi
34
35        if the requested page is not in the current_window
36        then
37              delete all the page in current_window and readahead_window
38              read the estimated number of readahead pages starting
39                   from the requested page and place them into the current
40                   window.
41              if all these pages already reside in the page cache
42              then
43                   shrink the estimated readahead_size by 1 and
                     shutdown readahead if the estimate touches zero
44              fi
45            else if the readahead_window is reset
46                 then
47                       read the estimated number of readahead pages
48                       starting from the page adjacent to the last page
49                       in the current window and place them in the
50                       readahead_window.
51                       if all these pages already reside in the page cache
52                       then
53                           shrink the estimated readahead_size by 1 and
                             shutdown readahead if the estimate touches zero
54                       fi
55                 fi
56        fi
57      fi
58 done
```

Figure 1: *Readahead algorithm in 2.6.0*

moved the check at line 16 to not expect read access to start from the beginning.

For the same read pattern the simulator showed 99 32-page requests, 99 30-page requests, one 16-page request, and one 18-page request to the block layer. This was a significant improvement over 1600 1-page requests seen without these changes.

However, the DSS workload did not show any significant improvement.

2.3.2 First Hit

The reason why DSS workload did not show significant improvement was that readahead shut down because the accessed pages already resided in the page-cache. This behavior is partly correct by design, because there is no advantage in reading ahead if all the required pages are available in the cache. The corresponding code is at line 43. But shutting down readahead by just confirming that the initial few pages are in the page-cache and assuming that future pages will also be in the page cache, leads to worse performance. We fixed the behavior, to not close the `readahead_window` the first time, even if all the requested pages were in the page-cache. The combination of the above two changes ensured continuous large-size read activity.

The simulator showed the same results as the First-Miss fix.

However, the DSS workload showed 6% improvement.

2.3.3 Extremely Slow Slow-read Mode

We also observed that the slow-read mode of the algorithm expected 32 contiguous page access to resume large size reads. This is not a realistic expectation for random workload. Hence, we changed the behavior at line 9 to shrink the `current_window` by one page if it lost contiguity.

The simulator and DSS workload did not show any better results because the combination of First-Hit and First-Miss fixes ensured that the algorithm did not switch to the slow-read mode. However a request pattern comprising of 10 single page random requests followed by a continuous stream of 4-page random requests can certainly see the benefits of this optimization.

2.3.4 Upfront Readahead

Note that readahead is triggered as soon as some page is accessed in the `current_window`. For random workloads, this is not ideal because none of the pages in the `readahead_window` are accessed. We changed line 45, to ensure that the readahead is triggered only when the last page in the `current_window` is accessed. Essentially, the algorithm waits until the last page in the `current_window` is accessed. This increases the probability that the pages in the `readahead_window` if brought in, will get used.

With these changes, the simulator generated 99 30-page requests, one 32-page request, and one 16-page request.

There was a significant 16% increase in performance with the DSS workload.

2.3.5 Large `current_window`

Ideally, the readahead algorithm must generate around 100 16-page requests. Observe however that almost all the page re-

quests are of size 30 pages. When the algorithm observes that a page request has missed the `current_window`, it scraps both the `current_window` and the `readahead_window`, if one exists. It ends up reading in a new `current_window`, whose size is based on the estimated `readahead_size`. Since all of the pages in a given application's read request are contiguous, the estimated readahead size tends to reach the maximum `readahead_size`. Hence, the size of the new `current_window` is too large; most of the pages in the window tend to be wasted. We ensured that the new `current_window` is as large as the number of pages that were used in the present `current_window`.

With this change, the simulator generated 100 16-page requests, and 100 32-page requests. These results are awful because the last page of the application's request almost always coincides with the last page of the `current_window`. Hence, the readahead is triggered when the last page of the `current_window` is accessed, only to be scrapped.

We further modified the design to read the new `current_window` with one more page than the number of pages accessed in the present `current_window`.

With this change, the simulator for the same read pattern generated 99 17-page requests, one 32-page request, and one 16-page request to the block layer, which is close to ideal!

The DSS workload showed another 4% better performance.

The collective changes were:

1. Miss fix: Do not close readahead if the first access to the file-instance does not start from offset zero.

2. Hit fix: Do not close readahead if the first access to the requested pages are already found in the page cache.

3. Slow-read Fix: In the slow-read path, reduce one page from the `current_window` if the request is not contiguous.

4. Lazy-read: Defer reading the `readahead_window` until the last page in the `current_window` is accessed.

5. Large `current_window` fix: Read one page more than the number of pages accessed in the current window if the request misses the current window.

These collective changes resulted in an impressive 26% performance boost on DSS workload.

2.4 Sequential Workload

The previously described modifications were not without side effects! The sequential workload was badly effected. Trond Myklebust reported 10 times worse performance on sequential reads using the iozone benchmark on an NFS based filesystem. The lazy read optimization broke the pipeline effect designed for sequential workload. For sequential workload, readahead must be triggered as soon as some page in the current window is accessed. The application can crunch through pages in the `current_window` as the new pages get loaded in the `readahead_window`.

The key observation is that upfront readahead helps sequential workload and lazy readahead helps random workload. We developed logic that tracked the average size of the read requests. If the average size is larger than the maximum readahead size, we treat that workload as sequential and adapt the algorithm to do upfront readahead. However, if the average size is less than the maximum `readahead_`

```
1 for each page in the current request ; do
3       if readahead is shutdown
4       then // read one page at a time (SLOW-READ MODE)
5            if requested page is next to the previously requested page
6            then
7                 open the current_window by one more page
8            else
9                 shrink current_window by one page
10           fi
11           if the current_window opens up by maximum readahead_size
12           then
13                  activate readahead  // enter READAHEAD-MODE
14           fi
15           read in the requested page
       else // read many pages at a time (READAHEAD MODE)
16-17        if this is the first read request for this open file-instance ; then
18              set the estimated readahead_size to half the size of maximum readahead_size
19           fi
20           if the requested page is within the current_window
21              increase the estimated readahead_size by 2
22              ensure that this size does not exceed maximum readahead_size
23           else
24               decrease the estimated readahead_size by 2
25               if this estimate becomes zero, shutdown readahead
26           fi
27           if requested page is contiguous to the previously requested page
28           then
29              Increase the size of the present read request by one more page.
30           else
31               Update the average size of the reads with the size of the previous request.
32           fi
33           if the requested page is the first page in the readahead_window
34           then
35                move the pages in current_window to the  readahead_window
36                reset readahead_window
37                continue
38           fi
39-40        if the requested page is not in the current_window ; then
41              delete all pages in current_window and readahead_window
42              if this is not the first access to this file-instance
43              then
44                 set the estimated number of readahead pages to the
                   average size of the read requests.
45              fi
46              read the estimated number of readahead pages starting from
                    the requested page and place them into the current window.
47              if this not the first access to this file instance and
                      all these pages already reside in the page cache
48              then
49                    shrink the estimated readahead_size by 1 and
                            shutdown readahead if the estimate touches zero
50              fi
51           else if the readahead_window is reset and if the average
                    size of the reads is above the maximum readahead_size
52                then
53                     read the readahead_window with the estimated
54                       number of readahead pages starting from the
55                       page adjacent to the last page in the current window.
56                       if all these pages already reside in the page cache
57                       then
58                           shrink the estimated readahead_size by 1 and
                             shutdown readahead if the estimate touches zero
59                       fi
60                fi
61           fi
62-63    fi ; done
```

Figure 2: *Optimized Readahead algorithm*

`size`, we treat that workload as random and adapt the algorithm to do lazy readahead.

This adaptive-readahead fixed the regression seen with sequential workload while sustaining the performance gains of random workload.

Also we ran a sequential read pattern through the simulator and found that it generated large size upfront readahead. For large random workload it hardly read ahead.

2.4.1 Simplification

Andrew Morton rightly noted that reading an extra page in the `current_window` to avoid lazy-readahead was not elegant. Why have lazy-readahead and also try to avoid lazy-readahead by reading one extra page? The logic is convoluted. We simplified the logic through the following modifications.

1. Read ahead only when the average size of the read request exceeds the maximum `readahead_size`. This helped the sequential workload.
2. When the requested page is not in the `current_window`, replace the `current_window`, with a new `current_window` the size of which is equal to the average size of the application's read request.

This simplification produced another percent gain in DSS performance, by trimming down the `current_window` size by a page. More significantly the sequential performance returned back to initial levels. We ran the above modified algorithm on the simulator with various kinds of workload and got close to ideal request patterns submitted to the block layer.

To summarize, the new readahead algorithm has the following modifications.

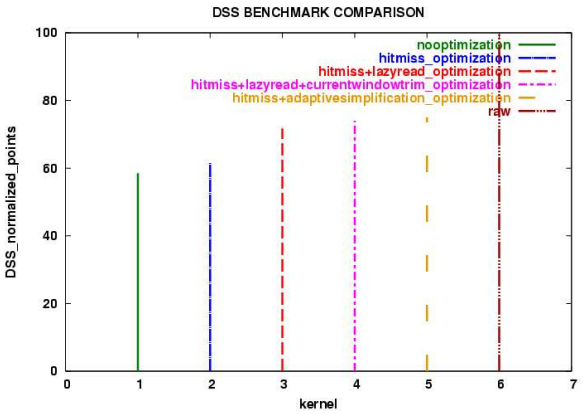


Figure 3: *Progressive improvement in DSS benchmark, normalized with respect to the performance of DSS on raw devices.*

1. Miss fix: Do not close readahead if the first access to the file-instance does not start from offset zero.
2. Hit fix: Do not close readahead if the first access to the requested pages are already found in the page cache.
3. Slow-read Fix: Decrement one page from the `current_window` if the request is not contiguous in the slow-read path.
4. Adaptive readahead: Keep a running count of the average size of the application's read requests. If the average size is above the maximum `readahead_size`, readahead up front. If the request misses the `current_window`, replace it with a new `current_window` whose size is the average size of the application's read requests.

Figure 2 shows the new algorithm with all the optimization incorporated.

Figure 3 illustrates the normalized steady increase in the DSS workload performance with each incremental optimization. The graph is normalized with respect to the performance of

DSS on raw devices. Column 1 is the base performance on filesystem. Column 2 is the performance on filesystem with the hit, miss and slow-read optimization. Column 3 is the performance on filesystem with first-hit, first-miss, slow-read and lazy-read optimization. Column 4 is the performance on filesystem with first-hit, first-miss, slow-read, and large `current_window` optimization. Column 5 is the performance on filesystem with first-hit, first-miss, slow-read, and adaptive read simplification. Column 6 is the performance on raw device.

3 Overall Performance Results

In this section we summarize the results collected through simulator, DSS workload, and iozone benchmark.

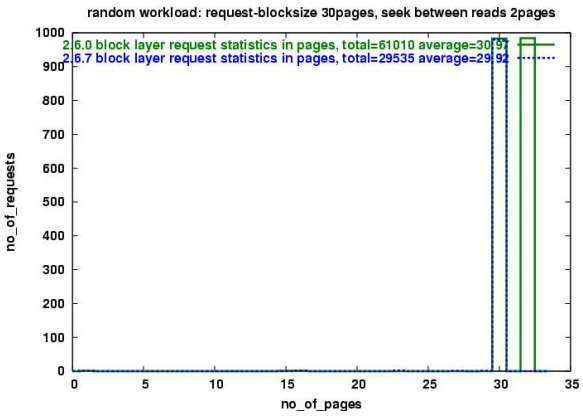
3.1 Results Seen Through Simulator

We generated different types of input read patterns. There is no particular reason behind these particular read pattern. However, we ensured that we get enough coverage. Overall the read requests generated by our optimized readahead algorithm outperformed the original algorithm. The graphs refer to our optimized algorithm as 2.6.7 because all these optimizations are merged in the 2.6.7 release candidate.

Figure 4 shows the output of readahead algorithm with and without optimization for 30-page read request followed by 2-page seek, repeated 984 times.

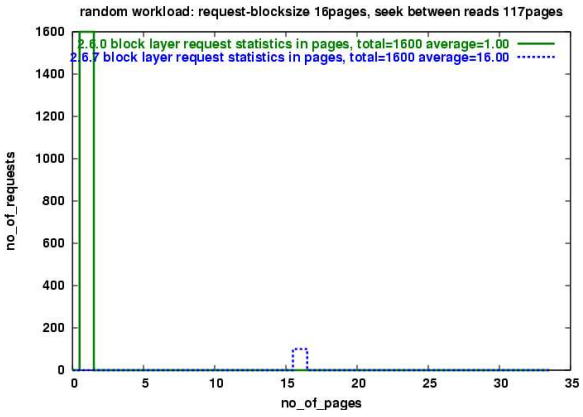
Figure 5 shows the output of readahead algorithm with and without optimization for 16-page read request followed by 117-page seek, repeated 100 times.

Figure 6 shows the output of readahead algorithm with and without optimization for 32-



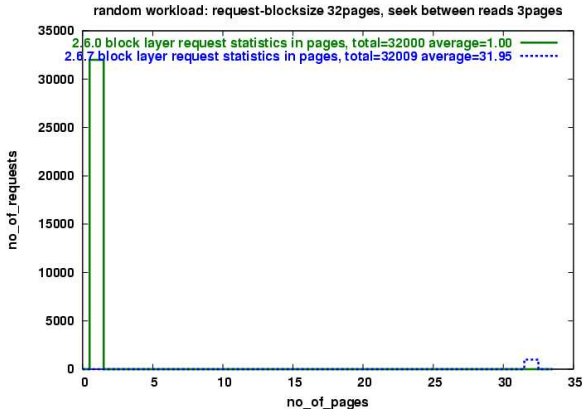
	2.6.0	2.6.7
Average Size	31	30
Pages Read	61010	29535
Wasted Pages	31490	15
No Of Read Requests	1970	987

Figure 4: *Application generates 30-page read request followed by 2-page seek, repeating 984 times. Totally 29520 pages requested.*



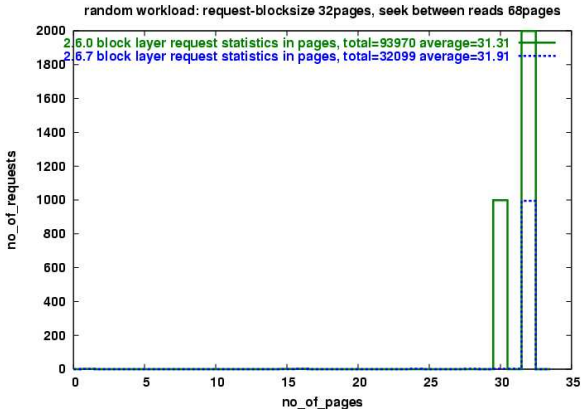
	2.6.0	2.6.7
Average Size	1	16
Pages Read	1600	1600
Wasted Pages	0	0
No Of Read Requests	1600	100

Figure 5: *Application generates 16-page read request followed by 117-page seek, repeating 100 times. Totally 1600 pages requested.*



	2.6.0	2.6.7
Average Size	1	31.95
Pages Read	32000	32009
Wasted Pages	0	9
No Of Read Requests	32000	1002

Figure 6: *Application generates 32-page read request followed by 3-page seek, repeating 1000 times. Totally 32000 pages requested.*



	2.6.0	2.6.7
Average Size	31.31	31.91
Pages Read	93970	32099
Wasted Pages	61970	99
No Of Read Requests	3001	1006

Figure 7: *Application generates 32-page read request followed by 68-page seek, repeating 1000 times. Totally 32000 pages requested.*

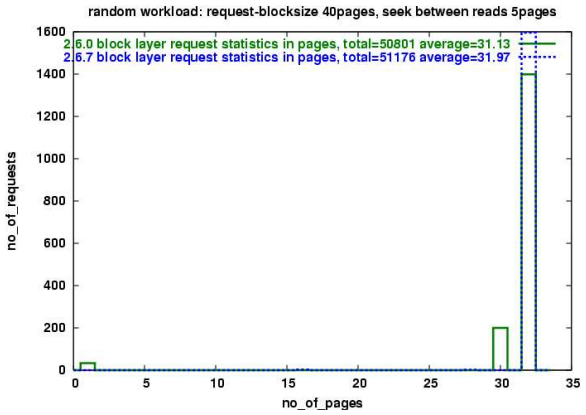
page read request followed by 3-page seek, repeated 1000 times.

Figure 7 shows the output of readahead algorithm with and without optimization for 32-page read request followed by 68-page seek, repeated 1000 times.

Figure 8 shows the output of readahead algorithm with and without optimization for 40-page read request followed by 5-page seek, repeated 1000 times.

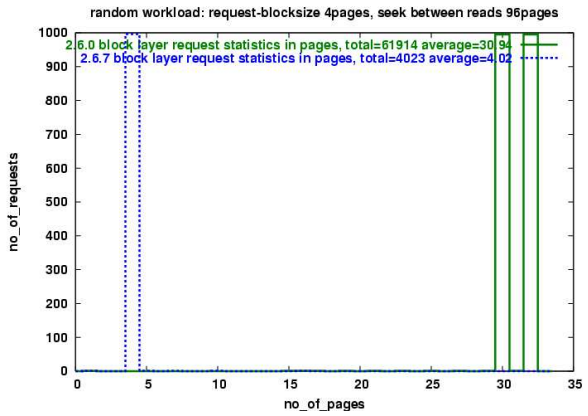
Figure 9 shows the output of readahead algorithm with and without optimization for 4-page read request followed by 96-page seek, repeated 1000 times.

Figure 10 shows the output of readahead algorithm with and without optimization for 16-page read request followed by 0-page seek, repeated 1000 times.



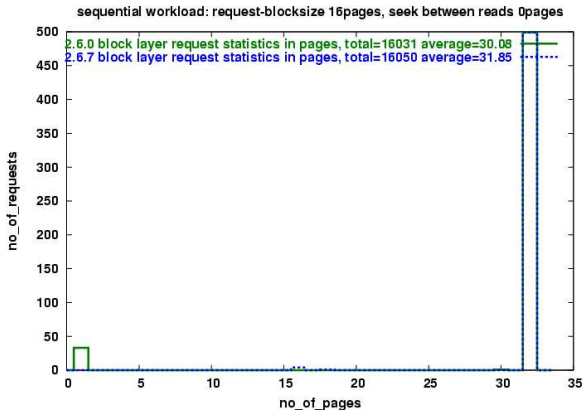
	2.6.0	2.6.7
Average Size	31.13	31.91
Pages Read	50810	51176
Wasted Pages	10801	11176
No Of Read Requests	1631	1601

Figure 8: *Application generates 40-page read request followed by 5-page seek, repeating 1000 times. Totally 40000 pages requested.*



	2.6.0	2.6.7
Average Size	30.94	4.02
Pages Read	61914	4023
Wasted Pages	57914	23
No Of Read Requests	2001	1001

Figure 9: *Application generates 4-page read request followed by 96-page seek, repeating 1000 times. Totally 4000 pages requested.*



	2.6.0	2.6.7
Average Size	30.08	31.85
Pages Read	16031	16050
Wasted Pages	31	50
No Of Read Requests	533	504

Figure 10: *Application generates 16-page read request with no seek, repeating 1000 times. Totally 16000 pages requested.*

3.2 DSS Workload

The configuration of our setup is as follows:

- 8-way Pentium III machine.
- 4GB RAM
- 5 fiber-channel controllers connected to 50 disks.
- 250 partitions in total each containing a ext2 filesystem.
- 30GB Database is striped across all these filesystems. No filesystem contains more than one table.
- Workload is mostly read intensive, generating mostly large 256KB random reads.

With this setup we saw an impressive 26% increase in performance. The DSS workload on filesystems is roughly about 75% to DSS workload on raw disks. There is more work to do, although the bottlenecks may not necessarily be in the readahead algorithm.

3.3 Iozone Results

The iozone benchmark was run a NFS based filesystem. The command used was `iozone -c -t1 -s 4096m -r 128k`. This command creates one thread that reads a file of size 4194304 KB, generating reads of size 128 KB. The results in Table 1 show an impressive 100% improvement on random read workloads. However we do see 0.5% degradation with sequential read workload.

4 Future Work

There are a couple of concerns with the above optimizations. Firstly, we see a small 0.5%

Read Pattern	2.4.20	2.6.0	2.6.0 + optimization
Sequential Read	10846.87	14464.20	13614.49
Sequential Re-read	10865.39	14591.19	13715.94
Reverse Read	10340.34	10125.13	20138.83
Stride Read	10193.87	7210.96	14461.63
Random Read	10839.57	10056.49	19968.79
Random Mix Read	10779.17	10053.37	21565.43
Pread	10863.56	11703.76	13668.21

Table 1: *Iozone benchmark Throughput in KB/sec for different workloads.*

degradation with the sequential workload using the iozone benchmark. The optimized code assumes the given workload to be random to begin with, and then adapts to the workload depending on the read patterns. This behavior can slightly affect the sequential workload, since it takes a few initial sequential reads before the algorithm adapts and does upfront readahead.

The optimizations introduce a subtle change in behavior. The modified algorithm does not correctly handle inherently-sequential clustered read patterns. It wrongly thinks that such read patterns seek after every page-read. The original 2.6 algorithm did accommodate such patterns to some extent. Assume an application with 16 threads reading 16 contiguous pages in parallel, one per thread. Based on how the threads are scheduled, the read patterns could be some combination of those 16 pages. An example pattern could be 1,15,8,12,9,6,2,14,10,7,5,3,4,11,12,13. The original 2.6.0 readahead algorithm did not care which order the page requests came in as long as the pages were in the current-window. With the adaptive readahead, we expect the pages to be read exactly in sequential order.

Issues have been raised regularly that the readahead algorithm should consider the size of the current read request to make intelligent decisions. Currently, the readahead logic bases its readahead decision on the read patterns seen in the past, including the request for the current page without considering the size of the current request. This idea has merit and needs investigation. We probably can ensure that we at least read the requested number of pages if readahead has been shutdown because of page-misses.

5 Conclusion

This work has significantly improved random workloads, but we have not yet reached our goal. We believe we have squeezed as much as possible performance from the readahead algorithm, though there is some work to be done to improve some special case workloads, as mentioned in Section 4. There may be other subsystems that need to be profiled to identify bottlenecks. There is a lot more to do!

6 Legal Statement

I would hate user space locking if it weren't that sexy...

(fusyn+RTNPTL: The making of a real-time synchronization infrastructure for Linux)

`http://developer.osdl.org/dev/robustmutexes/`

Iñaky Pérez-González
inaky.perez-gonzalez@intel.com

Boris Hu
boris.hu@intel.com

Salwan Searty
salwan.searty@intel.com

Adam Li
adam.li@intel.com

David P. Howell
david.p.howell@intel.com

Linux OS & Technology Team, Intel Corporation

Abstract

Linux has seen a lot of new features and developments in the last years in order to accommodate better scalability, interactivity, response time and POSIX compliance. With these changes, Telecom developers began to get serious about using Linux and started porting their systems to it. By doing that they brought new usage models and needs to the community; and among those needs was support for threads, mutual exclusion, priority inversion protection and robust synchronization for mission critical and fault-proof systems on both timesharing and soft real-time environments. This paper describes our experiences trying to meet this need, the current state and where are we headed. We will detail how originally we tried to modify the futex code, but later found we had to abandon that in favor of a similar design based on a layered implementation. This implementation accommodates a kernel and user space locking and synchronization infrastructure that will meet the requirements of those applications needing to use and port complex multithreaded real-time code.

1 A look at the requirements

The Carrier Grade Working group, or CGL, was created under the auspices of the OSDL; it provides a meeting point for all parties who share an interest on Linux use for Telecom: network equipment vendors, Linux distributors and developers, carriers, etc.

It was in this forum where missing features were identified. Carrier Grade Linux needed good soft real-time[1] features, specially with multi-threaded programs. As well, it needed a common feature provided by Solaris' mutexes

[1]For short, we'll use real-time to refer to *soft* real-time.

that was not present in Linux: *robustness*.

This project was started to provide a kernel synchronization infrastructure (*fusyn*) with the indicated characteristics, as well as the proper modifications to the NPTL user space library (*RTNPTL*) for it to use the new infrastructure and provide the new features.

The basic immediate requirements could be summarized in:

- The infrastructure should provide the primitives needed by NPTL to support the following POSIX tags:
  - TPS: thread priority scheduling
  - TPI: priority inheritance in mutexes
  - TPP: priority protection in mutexes

  Or simply: *anything that is needed for soft-real time support*.
- The implementation should support robust mutexes similar to those of Solaris.
- The implementation should provide equivalent features at the kernel level for use by drivers and subsystems.

With this in mind, we aimed to satisfy the following detailed requirements:

1. mutexes and conditional variables must work according to real-time expectancies
   (a) All operations (lock, unlock, priority promotion and demotion, etc.) should be deterministic in time, and *O(1)* when possible (except of course, for waits).
   (b) The order of lock acquisition by waiters (in mutexes) and wake up (in conditional variables) has to be determined by the scheduling properties of each blocked task/thread.
   (c) Minimization of priority inversion (given the importance of this item, it will be treated in its own section):
      i. lock stealing: in SMP systems, during on the acquisition of the lock a lower priority thread can steal the lock from a higher priority thread.
      ii. when a high priority thread A is waiting for a lower priority owner B to relinquish the mutex and B is preempted by a medium priority thread C.
         A. priority protection
         B. priority inheritance
2. Robustness: when a mutex owner dies, the mutex switches to a *dead-owner* state and the first waiter gets ownership with a special error code.
3. Uncontested locks/unlocks must happen without kernel intervention.
4. Deadlock detection

As well, in order to provide the benefits of this infrastructure to all the levels of a Linux system, it must be possible to use it not only by the user space code, but also by the kernel code.

1.1 The real time expectancies

Real-time is all about being **deterministic**, so all algorithm execution times need to be as predictable or bounded as possible. Using $O(1)$ algorithms helps with this [2].

[2] It is possible to be deterministic with a $O(f(N))$ operation, as long as $f(N)$ is known; however, in most, if not all, of the cases involving mutex operation, it is highly impractical or plainly impossible to find out $f(N)$, and thus a possibly simpler implementation has to be replaced with one potentially more complex, but $O(1)$.

POSIX dictates that upon unlock of a mutex, the scheduling policy shall determine who is the next owner. An obvious way of doing this would be to wake up all of the waiters and let them compete for the lock–the scheduler would determine that the highest priority task would get there first.

However, this causes scheduling storms, unnecessary context switches and general avoidable overhead. It is easier and more effective to determine which is the highest priority waiter and only wake that one up. To implement this task in an $O(1)$ way, we need to queue the waiters in a sorted list that provides constant time queuing and unqueuing. On unlock or wake up time, the first waiter in the list will be the highest priority one.

1.2 Priority inversion

This condition happens when a lower priority thread blocks a higher priority one. The most general case (Figure 1) is the lower priority thread that holds a resource needed by the higher priority one–a situation that has to be avoided–as much as possible. As indicated before, we aim to solve three different flavors.

The first is **lock stealing**. For performance reasons, to avoid the convoy phenomenon[3] [1], the *unlock* operation is done by unlocking the mutex and then waking up the first waiter (eg: A). The waiter claims the mutex and then becomes owner. On a single CPU system it can be preempted only by higher priority tasks[4], so lock stealing is not a problem; however, on multi-CPU systems, a lower priority task C running on another CPU could claim the lock just before B had the chance to do it and it would create a priority inversion scenario (see Figure 2).

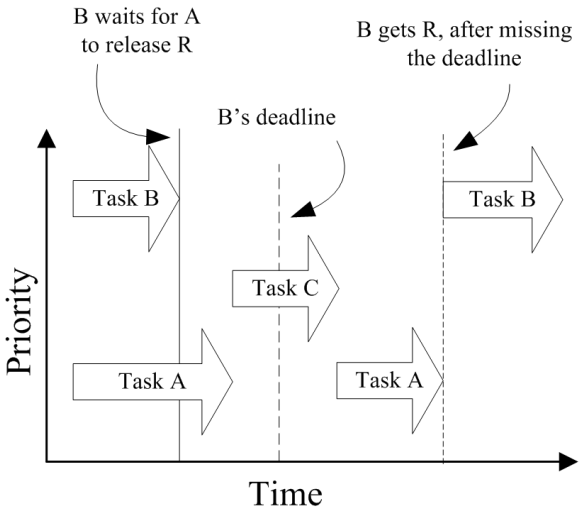


Figure 1: A case of priority inversion: high-priority task B misses its deadline because lower-priority task A holds for too long a resource it needs, as mid-priority task C preempted it. A lower priority task C blocks a high-priority task B.

The solution to this problem is simple: do not unlock the mutex, just transfer the ownership without unlocking it. We call this *serialized* unlock (versus *parallel*, wake and claim). This method severely limits performance in many cases, because it forces a context switch (causing the already mentioned convoy phenomenon). There has to be a compromise between protection and performance and by offering the option to unlock a mutex in either way, a developer can dynamically adapt according to her needs.

The other two cases (of priority inversion) are more complex. They solve the scenario depicted in Figure 1 where task B is waiting for a mutex owned by task A and task C preempts task A. When the priorities are $p(B) > p(C) > p(A)$, we have a priority inversion; task B will miss its deadline because C is blocking A from

[3] Summarizing: if task A (high priority) unlocks by transferring ownership to the first waiter B (lower priority), it forces a context switch to B, and if then A recontends for the lock it will create a convoy of waiters that is difficult to dissolve.

[4] We will use the terms tasks or threads indistinctly to refer to any entity that can acquire a mutex.

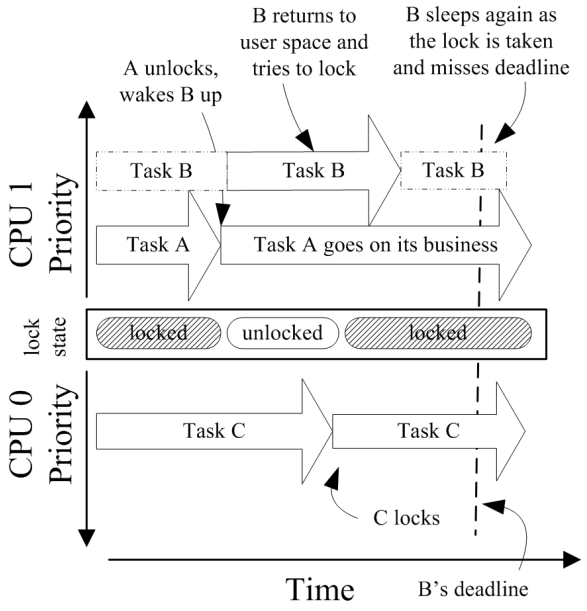


Figure 2: Low priority task C running on CPU0 steals the lock from higher priority task B running on CPU1.

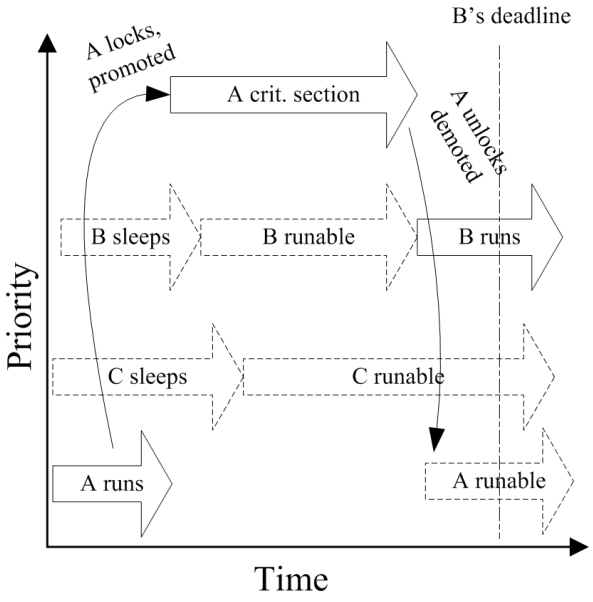


Figure 3: Priority protection: task A locks and its priority is promoted to the prioceiling; task C cannot preempt it and it finishes its critical section (and is demoted) in time for B to meet its deadline.

completing its mutex-protected critical section.

There are different ways to deal with this problem, but the most common involve bumping up the priority of the owner of the lock to a certain value.

In **priority protection** (or PP), a *priority ceiling* is determined as part of the design cycle. This is normally the highest of all the priorities among the threads that will share a given mutex; as soon as it is locked, the priority of the owner is raised to match that of the priority ceiling (see Figure 3. When a thread owns many priority-protected mutexes, its priority is that of the highest ceilings. This approach is simple and guaranteed to be trouble free. However, it is laborious; determining the priority ceiling might not be an easy task at all in a moderately complex system where modules from different parties need to interact.

Enter **priority inheritance** (PI): in this case we have a similar situation, but there is no priority ceiling. What happens in this case is that the priority of the owner is boosted up to that of the highest priority waiter, the first one (see Figure 4). Similarly to the previous case, if a task owns many PI-mutexes, its priority will be the highest of them all. There is no need now to do design-time analysis; the system solves it automatically. Of course, there are drawbacks– it does not come for free. This operation is more expensive, especially in the presence of owner/wait chains[5]. The propagation of the priority boost can be long (and will be $O(N)$ on the depth of the chain) and this can lead to unexpected surprises if the interaction across different threads and mutexes in the system is not kept on a tight leash (see [2] and [3]).

[5]Task A waits for mutex M that is owned by task B that is waiting for mutex N that is owned by task C that is waiting for mutex O…

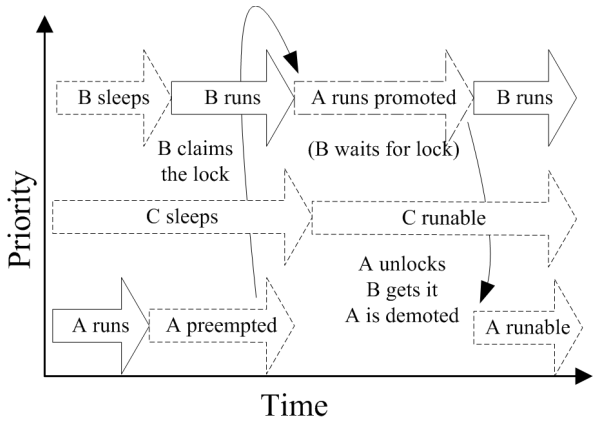


Figure 4: Priority inheritance: Task A (lock owner) is promoted to task B's priority when B waits for the lock; as soon as A unlocks, it gets demoted and B gets the lock. C never has a chance to preempt A.

Priority inheritance needs to be used with care– it is not a straight solution for a system with deadlock problems to make a mutex PI. What if that mutex is being shared with some low priority timesharing task that is not aware of the fact? In these cases, if a task does some kind of CPU spinning, the system is dead. The concept of priority inheritance and the simplicity it gives to designs provides enough rope as to hang oneself, as the effects can propagate way far more than expected.

1.3 Robustness

Mutex robustness is a key feature for implementing systems tolerant to certain kind of failures. A certain task A is holding a normal, *non-robust* mutex M with one or more waiters W_n blocked in the kernel. If it receives a fatal signal and is killed, the mutex will still be locked and the waiters will be never woken up. There are different ways to detect and recover from this situation, but they usually involve painful and complicated designs with watchdogs, timeouts, etc.

Robustness embeds all these in the mutex mechanism. When a task owns a mutex, the mutex knows who is its owner, and asks to be notified if the owner dies. If and when this happens, the mutex will be moved to an special *consistency state*, *dead-owner* and effectively unlocked; this will give control to the first waiter (or remain unlocked as *dead-owner* until somebody else claims it).

Threads claiming a dead mutex will receive ownership with a special error code, `-EOWNERDEAD`. This serves as a warning: *the data protected by this mutex might be inconsistent, it should be fixed*. The new locker can do different things at this point:

- it can ignore it (scary choice!)
- it might be unqualified for the job and pass the responsibility on to somebody else (by unlocking).
- it can try to fix the data and succeed–then it will *heal* the mutex, setting its consistency state back to normal and proceed.
- or it can fail and pass it on...
- or it can fail and deem the state completely broken; to notify about this situation, it can mark the mutex *not-recoverable*, so all waiters and future claimers will get a `-ENOTRECOVERABLE` error code so other recovery strategies can kick in.

The most important aspect to take into account is that the user of the mutex has means to detect this situation instantly without having to rely on timeouts or other overheads.

1.4 Deadlock detection

A situation of deadlock happens when a task A that owns a mutex M tries to lock it again.

This is the simplest case, of course. The general case is:

- task T_1 owns mutex M_1 and tries to lock mutex M_2
- task T_2 owns mutex M_2 and is waiting to lock mutex M_3
- task T_3 owns mutex M_3 and is waiting to lock mutex M_4
- ...
- task T_N owns mutex M_N and is waiting to lock mutex M_1

Construct like these are called *ownership-wait chains*. And it is obvious that in this particular case, this chain would deadlock, as M_1 would never be released if T_1 is allowed to block waiting for M_2.

The only way to detect this situation is, upon lock time, to walk the chain and verify if the task that is about to lock owns any lock on the chain.

By definition this is a linear operation that is going to take time to execute. The best way to avoid this expensive check is to make sure our design uses proper locking techniques (like for example, acquire and release multiple locks in LIFO order).

2 The first try: rtfutex

Once the requirements were laid out, we first tried modifying the futex code in `kernel/futex.c`, adding functionality while maintaining the original futex interface.

In a glimpse, the locking mode used with futexes works like this (see [4]): there is a word in user space that represents the mutex. The fast lock operation is performed entirely in user space; if the word is unlocked, then it becomes locked and work proceeds. If it is locked, the program sets a different value in the user space word and then goes down to the kernel and waits.

When the unlock operation is performed, the unlocker will check the value of the word; if it indicates that only a fast-lock was performed (and thus there are no waiters in the kernel), it will be simply unlocked in user space; otherwise, it will ask the kernel to wake up one or more waiters. These waiters will come back to user space and reclaim the lock; only one will get it, the rest will go back to the kernel to sleep[6].

With this in mind, we performed the following modifications:

- To allow wake-the-highest-priority waiter behavior on a bound time, the hash table model had to be modified.

  One node per waiter was replaced by one node per futex, and each node would have its own priority-ordered list of waiters. Although the lookup of the futex-node in the hash table is $O(N)$, at least the manipulation of the waiter-list (or wait list) can be made $O(1)$.

  This introduced the need of having to allocate the futex-node, as it could not live in the stack of some waiter[7].

- In order to support robustness, deadlock detection and priority inheritance, the

[6] note this means that the lock is actually unlocked for an unspecified amount of time in an unlock to lock transition.

[7] This raises extra issues; allocation can fail and is not time-predictable; it can be slow, so it is needed to cache the nodes (as normally they are frequently reused); caching means a strategy is needed to purge them (garbage collection).

concept of *ownership* had to be added to the futex. This would save *which* task owns the futex on each moment. It also required to note in the task struct which futex was being waited for, as well as a list of owned futexes.

- A different method had to be used for locking in user space, the fast lock and unlock paths.

  The user space word representing the futex would store the PID of the locker while on the fast path and indicate with a bit the presence of waiters in the kernel. This way if a locker died after having done a fast-lock operation in user space (and thus the kernel not having any notion of it), a potential waiter could check if the lock was stale[8]. When a futex went into the *dead-owner* or *not-recoverable* state, the kernel would modify the user space word with special values to mark these states.

  As well, the unlock operation had to always be serialized, with the kernel assigning ownership and modifying the user space word, ensuring robustness[9] and that no lock stealing happened.

This design (and its implementation) was broken: the futexes are designed to be queues, and they cannot be stretched to become mutexes– it is simply not the same. The result was a bloated implementation.

As well, the code itself missed many fine (and not so fine) details:

- it suffered from race conditions: the modification of the different back pointers in the task struct was being done without protection.

- the priority inheritance engine was very limited (to the most simple cases of inheritance) and it didn't support `SCHED_NORMAL` tasks.

- serialized unlocking is slow, it causes the convoy phenomenon, and the code did not provide flexibility to allow the user to balance performance vs. robustness or priority inversion protection depending on the situation.

- it didn't provide the functionality at the kernel level, for usage by kernel code.

- it didn't support changing the priority of a task while it was waiting for a futex while at the same time properly repositioning it on the wait list according to its new priority.

While broken, it was perfect as a prototype– it gave an indication of what was wrong, how things should not be done and hinted which methods were a good idea. It was time to rethink all over again.

3 Trying again: fusyn

With rtfutexes we found that stretched designs are not a good idea, however, experience tells layered designs are a better idea.

The fusyn design follows the same basic principles of the futexes, providing the same service in kernel and user space. Enforcing a strict modularity among the different units that comprise it, it is possible to accomplish much more with less bloat and complexity.

[8] This is a very simple method that cannot guarantee conflicts when PIDs are reused; we implemented a naive task-signature system to try to avoid this case. We didn't realize how broken it was until later.

[9] If waiters coming up from the kernel died before locking again and there were still some others waiting, the kernel would never know about it and the remaining tasks would wait for ever.

The four main blocks that comprise the fusyn architecture are:

- *fuqueues* are the wait queues (very similar to the Linux kernel's waitqueues) and are the basic building block.
- *fulocks* provide the mutex functionality by adding the concept of ownership on top of fuqueues and dealing with all the priority promotion.
- *vlocators* serve as the link between user space words and the kernel objects (fuqueues and fulocks) associated to them.
- *vfulock sync* maintains synchronization between the fulocks and the *vfulocks*, the user space word associated to them. It also is responsible for identifying owners from the cookies stored in the vfulocks.

fuqueues

We start with a simple queue structure, `struct fuqueue`, declared in `linux/fuqueue.h`. It merely contains a priority-sorted list where to register the waiters for the queue, a spinlock and an operations pointer. The operations are for managing a reference count (used when associated to user space), for canceling a task's wait on a fuqueue and notifying the fuqueue of a priority change on a waiter (most functions are defined in `kernel/fuqueue.c`).

A fuqueue can be initialized, waited on with `fuqueue_wait()` or a number of waiters for it can be woken up with `fuqueue_wake()`. All the functions for doing that are conveniently broken up so they can be used by other layers.

Whenever a task waits on a fuqueue, it registers itself by filling up a `struct fuqueue_waiter`; that structure and the fuqueue being waited for are linked to from the task struct (`struct fuqueue_waiter *fuqueue_waiter` and `struct fuqueue fuqueue_wait`), so that the signal delivery code (through `fuqueue_waiter_cancel()`) and the scheduler priority changing functions (through `fuqueue_waiter_chprio()`) can properly locate which fuqueue to act upon. A spinlock protects these pointers in the task structure.

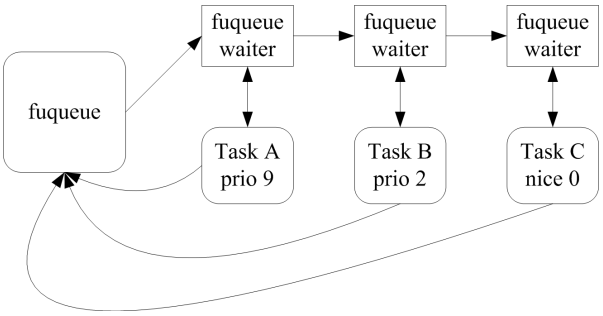


Figure 5: A fuqueue with three waiters, $p(A) > p(B) > p(C)$, showing the different pointers on each structure.

This satisfies the real-time requirements of wake-up order by priority. As well, the addition to the waiters list is bounded in time to the maximum number of different priority levels used–being this 140 for the Linux kernel, that makes the addition operation $O(140) \equiv O(1)$.

Note the fuqueue structure has to be protected, similarly to waitqueues with an IRQ-safe spinlock, as they will be accessed for wake-up from atomic contexts.

fulocks

Once we have a queue structure that is real-time friendly, we can build mutexes on top of them. Adding the concept of ownership, we create a `struct fulock` in `linux/fulock.h` that contains a fuqueue (for the waiters), a

pointer to a task struct (the owner), some flags and a node for a priority-sorted ownership list (to register all the fulocks owned by a task).

Let's ignore for a while the secondary effects of priority inheritance and protection. Come lock time, `fulock_lock()`: if the fulock is unlocked, the current task is assigned ownership by setting the owner pointer in the fulock to point to the task, the fulock is added to the `task->fulock_olist` ownership list through its `olist_node`.

If the fulock is locked (unless just try-locking) the task waits on the fulock's fuqueue; when woken up, depending on the result code of the wake up, it will own the lock (and thus proceed) or try again (serialized vs. parallelized unlocks).

The unlock operation, `fulock_unlock()` is quite simple: if the unlocker desires to perform a serialized wakeup, it just changes the owner to be the first waiter, removes it from the wait list and wakes him up with a 0 result code. If the unlock has to be parallelized, it unlocks the fulock and unqueues and wakes up the first waiter (or the first N waiters) with a `-EAGAIN` code–that will lead the sleeping `__fulock_lock()` call to retry. The unlock mode can be automatically determined based on the policy of the first waiting task: serialized for real-timers, parallelized for timesharers.

All this code is defined in `kernel/fulock.c`.

Robustness

Robustness comes into play with a hook in `kernel/exit.c:do_exit()`. When a process dies, `exit_fulocks()` goes over the list of fulocks owned by the exiting task; for each of them, the operation registered for task `exit` is executed, and that leads to setting the dead flag (`FULOCK_FL_DEAD`) and serially unlocking the fulock to the next waiter with the `-EOWNERDEAD` error code[10].

This introduces the need to have a way for the user to switch the fulock from one state to the other. `fulock_ctl()` provides this capability.

Deadlock detection

The process of checking for deadlocks is done via a hook in the `__fulock_lock()` function that calls `__fulock_check_deadlock()`.

This function will query the owner of the fulock the current task wants to wait for and inquire which fulock this owner is waiting for. If not waiting for anyone, there is no possible deadlock, so all resources are dropped and success is returned.

If it is waiting, the fulock is safely acquired (the ugliest part is to get the spinlocks properly as well as the reference counts); if the owner is the current task, then that is a deadlock; if not, then the operation repeats with the owner of the new fulock.

Priority inheritance and protection

Now let's take priority inheritance and protection into consideration. The key here is that in the priority-sorted list (plist), every node, including the list head, has a priority field, and that in a consistent plist, the priority of the list is that of the head, that in turn is that of the highest priority node queued.

Thus, by virtue of the priority-sorted list, each fuqueue has a *priority*. Fulocks inherit this property and when doing **priority inheritance**, they set that priority on the node for the priority-based ownership list. **Priority protected** fulocks set as priority that of the priority

[10] as well, a warning is issued if the fulock wasn't declared robust.

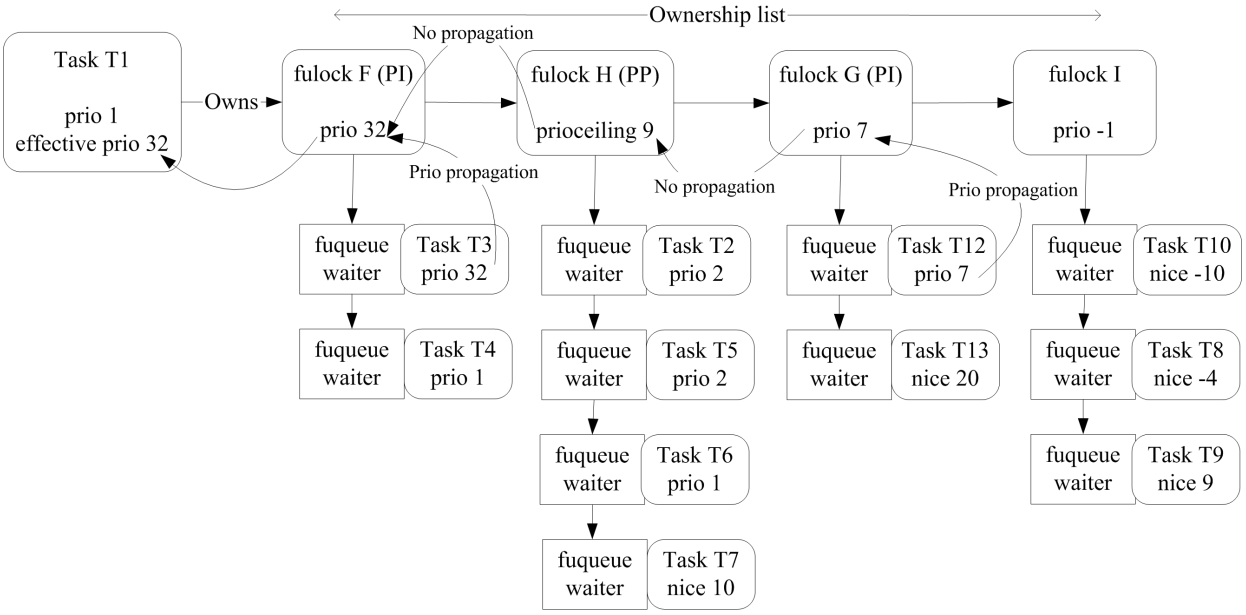


Figure 6: A task that owns four contested fulocks (two PI, one PP and one normal) showing the priority propagation flow. Note how fulock I, not being priority inheriting or protecting, has the minimal priority, -1 (which effectively disables all side effects).

ceiling of the fulock.

This way, each task has a list of the fulocks it owns sorted by priority. The ordering of the list means that the first fulock in the list has the minimum priority the task should have to meet the priority protection and/or priority inheritance criteria–and thus, the scheduler just has to select as effective task priority the highest between the task's final dynamic priority and that of the first fulock on its ownership list[11]. See Figure 6.

The process then becomes extremely simple: when a task queues waiting for a fulock (in `__fuqueue_waiter_queue()`), it might modify the plist priority because it sets a new *higher* priority–the function returns !0 in this case. This is propagated, with `__fulock_prio_update()` to the fulock's ownership list node, `fulock->olist_node`, that as we said above, is inserted in the ownership list of the fulock owner. The propagation could mean that a new maximum might be set in the ownership list, case in which the boost priority is updated for the scheduler to pick it up.

On top of that, the change might need to be propagated further on if the fulock owner is waiting for another fuqueue or fulock. `__fuqueue_waiter_chprio()` will take care of propagating that change until a task is reached that is higher priority or is not waiting for a priority-inheriting fulock.

Linking to user space

So far, the infrastructure presented is accessible only from kernel space. We have to allow user space programs to take advantage of these features, and for that, we copy the futex's method: associate a virtual address (word) to

[11] This is accomplished with a simple mechanism (improvement required to reduce invasiveness) that adds the concept of boost priority to the task struct (`boost_prio`), and modified through `__prio_boost()`.

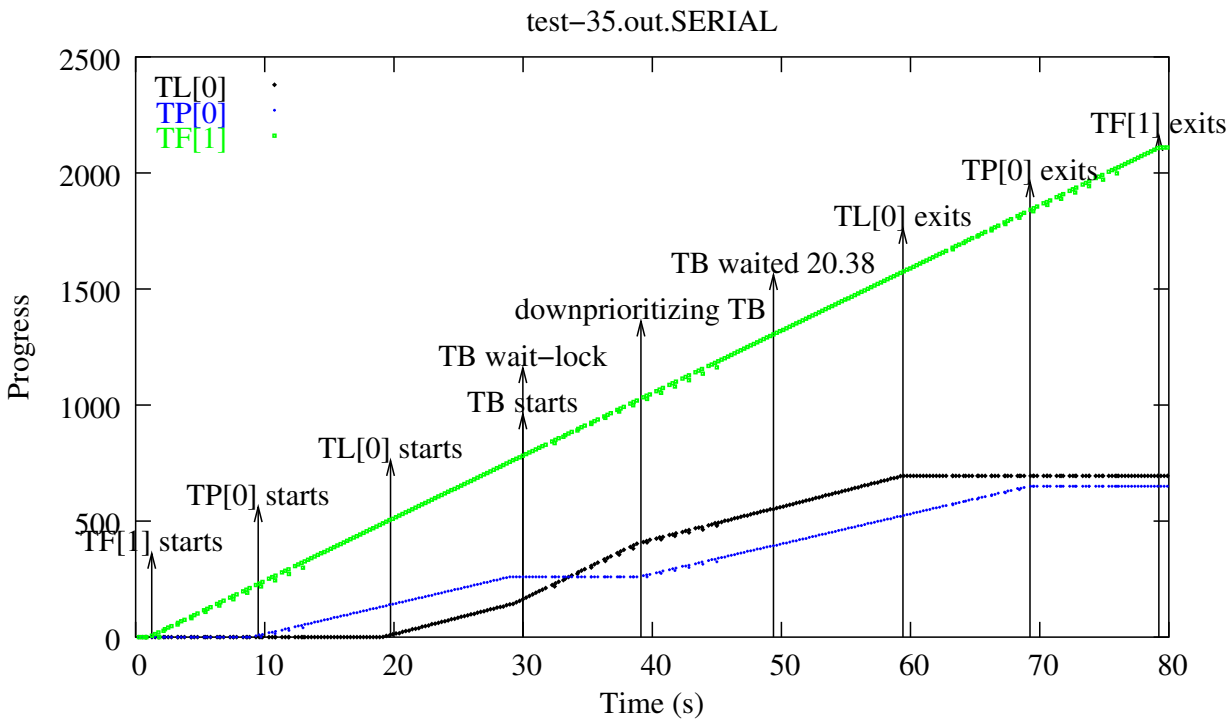


Figure 7: Testing priority inheritance: four threads of increasing priority (TL < TP < TB < TF) in an infinite loop counting up (progress); TF stays in CPU1 as a reference; TP sleeps from time to time in CPU0 to give TL a change; TL progresses what TP allows it. When TB starts (at $30s$), it claims a priority-inheriting fulock owned by TL and thus it gets boosted, TP doesn't progress any more. At almost $40s$, TB is down prioritized and that deboosts TL, allowing TP to progress again.

an object in memory (`struct vlocator`).

The API exposed in `linux/vlocator.h` provides a generic method for doing this by just embedding a vlocator; as well, this vlocator provides a reference-counting interface to simplify the object's life cycle management. And when it's use count is zero, it will be automatically disposed of[12].

This also improves scalability a little bit as the only global lock in the vlocator hash table is taken just to do the look up; once found, the vlocator is referenced before dropping the lock.

[12]Here is where the caching kicks in; the hash table is cleaned up of zero ref-counted items every certain amount of time, allowing for reuse.

ufuqueues and vfuqueues: imitating futexes

We need to create an interface equal to that of futexes for implementing conditional variables with real-time friendly functionality (for the wake up ordering).

We create a `struct ufuqueue` where we embed a vlocator and a fuqueue. A thin adaptation layer (`sys_ufuqueue_wait()` and `sys_ufuqueue_wake()`) will get the system call from user space, do the look up using the vlocator API, verify that the user space word (`vfuqueue`) hasn't changed and pass it down to the fuqueue layer.

The rest of the code in `kernel/ufuqueue.c` deals with creating the operation functions for

the vlocator structure.

Exposing the fulocks to user space

The same mechanism is used for exposing the fulocks to user space; in a similar fashion to fuqueues, we wrap together a vlocator and a fulock to create a `struct ufulock`.

However, more aspects have to be taken into consideration:

- If the fulock is not contested, the lock and unlock operations must happen entirely in user space (and thus the kernel will not know about it; this is the *fast-path*)
- When a lock has been locked through the fast path, the kernel has to be able to identify who locked it as well as its consistency status; this operation is called synchronization.
- When a lock becomes contested, the kernel has to update the user space word to indicate that future operations need to proceed in the kernel–as well, when it is eligible to be a fast-path only fulock again, the kernel must undo this, put the fulock structure in the cache tagged as requiring synchronization from user space and make sure the user space word has the consistency state of the fulock.
- The fulock structure in the kernel will be disposed if no task goes to the kernel querying about or operating on it for a while; as in the previous case, the information will be kept in user space word to enable proper synchronization.

For this we need some more information than the one used by the same futex mechanism for the fast path. A locker needs to identify itself in the user space word (that we call *vfulock*) by storing a cookie that can directly map to a task struct in the kernel space. The most obvious choice for the cookie would be the PID[13].

However, this operation must be atomic–this means that we need an atomic compare-and-exchange operation, and thus, the lock operation becomes the following: compare-and-exchange the cookie against 0 (meaning unlocked); if it succeeds, then the vfulock is locked, if not, dive into the kernel. The kernel will map the address to a fulock (possibly creating a new ufulock) get the value of the vfulock (`sys_ufulock_lock()` and `ufulock_lock()`) and map it to a task (in `__vfulock_sync()`. If the kernel is able to find the task, that task is made the owner and the caller is put to wait. As well, the vfulock is updated to a special value `VFULOCK_WP`, meaning waiters are present in the kernel.

If the kernel cannot find it, that will mean the task that fast-locked it in user space has died, the fulock will be declared *dead-owner* and the caller will get ownership. In this process, the vfulock will be set to another special value, `VFULOCK_DEAD` that indicates it as dead even across the kernel forgetting about its existence.

Unlocks are equally simple: atomically compare-and-exchange 0 (`VFULOCK_UNLOCKED`) against the cookie of the lock owner; if it succeeds, the job is done; else, the kernel does it. After mapping the vfulock to a ufulock, `ufulock_unlock()` is used to do the job and the vfulock is updated to reflect the new state: `VFULOCK_UNLOCKED` if unlocked, if there will be no waiters the new owner's cookie–enabling fast-path, `VFULOCK_WP` if waiters are still in the kernel, or `VFULOCK_DEAD` if the fulock is dead.

If parallelized unlocks are desired, the pro-

[13] This would break unique identification as PIDs are reused; a solution could be crypting the PID with the task creation date, but it needs to be tested.

cess is a little bit different. In the kernel, `__ufulock_unlock()` will unlock the vfulock and then wake up the first waiter, who then will contend (in the kernel) for the vfulock and possibly wait, as described above[14].

Note the two key moments in switching from fast-path enabled or not: the fulock becomes fast-path when it has no waiters in the kernel or when it is healed[15] without waiters. It looses the fast-path conditions as soon as a single waiter is queued. This means that to maintain proper semantics during the lifetime of a program that uses many locks, once a fulock has gone through the slow path, it needs to be destroyed in the kernel using the `sys_ufulock_ctl()` system call once it is not needed anymore. If not, there could be inconsistencies if a new lock is created in the same address where a previous one lived before.

KCO: When the fast-[un]lock path is not an option

The fast path, as we have seen, requires an atomic compare-and-exchange operation. Not all architectures provide this capability, so different strategies need to be considered here.

If robustness, and priority inversion protection[16] can be spared, the mutexes and conditional variables can be implemented as with futexes using fuqueues; the rest of the real-time featurettes are there (priority-based wake-ups and priority change semantics). If that can also be spared, futexes are still an option.

However, when that is not the case, the only possible choice is to use KCO mutexes, by ORing `FULOCK_FL_KCO` in the flags. That is an acronym for Kernel Controlled Ownership, or basically, the kernel takes care of everything. It needs to be called for locking and unlocking, there is no fast path (strictly speaking there is still a choice for fast path on some operations, as the vfulock is used to cache the consistency state of the fulock and any user space operation can check it before deciding if it should go to the kernel).

This feature also provides the highest level of protection for robustness. The per-thread cookie for the vfulock, be it the PID or any other, is not required, and the kernel deals directly with the task struct, so there is no possible collision conflict.

It has to be noted that priority-protected ufulocks always work in KCO mode. Even on uncontended acquisition or release the priority of the thread has to be changed to that of the prioceiling, and that task can only be done by the kernel.

4 Using it in the kernel

The fulock is a simple type like any other struct. To use it, we just need to do the following declarations:

```
...
#include <linux/fulock.h>
...

struct mystruct {
  struct fulock lock;
  ...
  my shared data;
};
```

It needs to be properly initialized before use, and of course, after releasing it (or more properly, telling all waiters to bail out) it shall not

[14] Not going back to user space to retry the operation has advantages: speed and maintaining the conditions for robustness.

[15] Moved from *dead-owner* consistency state back to normal (or healthy)

[16] Lock stealing avoidance, priority inheritance and priority protection.

be used. Note some flag combinations are not allowed (for example, querying for priority inheritance and protection at the same time is illegal) and will trigger a `BUG()`[17].

In this example we ask for a robust fulock with priority inheritance. It must be noted that fulocks are always robust–but clearly telling the kernel that we handle robust situations will suppress a kernel warning if the owner dies and it goes into *dead-owner* mode.

```
my_driver_probe(...)
{
 struct mystruct *my;
 ...
 my = kmalloc (...);
 if (my == NULL)
  goto err_alloc;
 fulock_init (&my->lock,
              FULOCK_FL_ROBUST
              | FULOCK_FL_PI);
 ...
};
```

As we see in the following snippet, the basic usage is the same as for every lock. However, in this case we add some recovery code for the case when some owner died[18]. Note also that the only fulock operation that is guaranteed to be safe in an atomic context is `fulock_unlock()`.

```
void my_something(
 struct mystruct *my) {
 ...
 result = fulock_lock(&my->lock,
                      0);
 if (result == -EOWNERDEAD
     && my_try_recover (my))
   goto notrecoverable;
 ...
 /* do our thing */
 ...
 fulock_unlock (&my->lock,
                FULOCK_FL_AUTO);
 ...
 return 0;

notrecoverable:
 /* Put it out of its misery,
  * release waiters, clean up,
  * user has to reload the
  * driver. */
 fulock_ctl (&my->lock,
             FULOCK_CTL_NR);
 my_put (my);
 return -ENOTRECOVERABLE;
};

int my_try_recover (struct
    *mystruct my) {
 int result, mode;
 ... try to recover *my ...
 if (successful) {
  result = 0;
  mode = FULOCK_CTL_HEAL;
 }
 else {
  result = !0;
  mode = FULOCK_CTL_NR);
 }
 fulock_ctl (&my->lock, mode);
 return result;
}
```

Finally, when we are done, we release all resources associated to the fulock to clean up. As indicated above, this merely makes sure that any waiter queued is woken up with an error condition and nobody can acquire it or queue again.

```
void my_cleanup (
 struct mystruct *my)
{
 ...
```

[17] For user space code, they will simply fail with `-EINVAL`.

[18] This is kind of an useless exercise, correct kernel code doesn't crash.

```
 fulock_release (&my->fulock);
 ...
}
```

The benefits that a fulock gives over a semaphore are the real-time characteristics, priority inheritance and protection and deadlock detection. The decision to use one or the other depends on the user needs, as it has to be taken into account that fulocks are somehow more heavyweight than semaphores.

5 Usage from user space

The main intention of the user space code is to do as little as possible in the fast path and delegate the rest to the slow path that will, in most cases, end up in the kernel.

Note these code snippets have been slightly simplified; for the authoritative reference, see the file `src/include/kernel-lock.h` in the test package `fusyn-package` available from the web site.

Locking

As mentioned, the fast lock operation needs an atomic compare and swap operation; for example, on i386:

```
unsigned acas (
 volatile unsigned *value,
 unsigned old_value,
 unsigned new_value)
{
 unsigned result;
 asm __volatile__ (
  "lock cmpxchg %3, %1"
  : "=a"(result),"+m"((*value))
  : "a"(old_value),"r"(new_value)
  : "memory");
 return result == old_value;
}
```

To simplify the code, this function returns true if it was successful in performing the swap operation. With this, we can create a generic, fast-path, user space lock operation:

```
int vfulock_timedlock (
 volatile unsigned *vfulock,
 unsigned flags, int pid,
 struct timespec *rel)
{
 if (acas(vfulock,
          VFULOCK_UNLOCKED,pid))
  return 0;
 return SYSCALL (ufulock_lock,
                 vfulock, flags,
                 rel);
}
```

We are using the thread's PID as the cookie for the vfulock, the user space memory word associated to the lock. Note the special syntax for timeouts understood by the kernel:

- Passing `NULL` means we don't want to wait, and this operation effectively becomes a trylock in the kernel.
- A `(void *)-1` timeout means block forever–no timeout.
- Any other specifies a pointer to a valid timeout structure.

From user space we have to always pass the same flags to the kernel for an specific vfulock, as it will check we are consistent during the lifetime of the fulock–when it dissapears from the cache, it is up to us to use still the same flags to maintain consistency in our program.

With a few additions, we can have a lock function that also works in KCO mode and that imitates the behavior of non-robust mutexes when owners die (*ie:* block forever):

```
int vfulock_timedlock (
 *vfulock, flags, pid, *rel)
{
 int result;
 if (!(flags & FULOCK_FL_KCO) &&
     acas(vfulock,
          VFULOCK_UNLOCKED,pid))
  return 0;
 result = SYSCALL (ufulock_lock,
                   vfulock, flags,
                   rel);
 if (!(flags & FULOCK_FL_RM) &&
     (result == -EOWNERDEAD
     |result == -ENOTRECOVERABLE))
  waiting_on_dead_fulock(vfulock);
 return result;
}
```

There are only two simple differences. First is to avoid the fast-path if we want to use KCO mode (and thus dive directly into the kernel). The second one takes care of non-robust mutexes returning in *dead-owner* state; in that case we block in `waiting_on_dead_fulock()`, a dummy function that blocks forever whose only purpose is to show up in program traces to indicate us the reason of a thread blocking.

Unlocking

The unlock operation is somehow more hairy. Although we could just make it simpler calling the kernel and letting it do all of the operations for us (as if it were in KCO mode), we want to have the fast-unlock path available:

```
int __vfulock_unlock (
  *vfulock, flags, unlock_type)
{
 unsigned old_value = *vfulock;

 if (flags & FULOCK_FL_KCO)
   goto straight;
retry:
 if (old_value < VFULOCK_WP) {
  if (acas (vfulock, old_value,
            VFULOCK_UNLOCKED))
   return 0;
  old_value = *vfulock;
  goto retry;
 }
straight:
 return old_value == VFULOCK_NR?
  -ENOTRECOVERABLE
  : SYSCALL (ufulock_unlock,
             vfulock, flags,
             unlock_type);
}
```

As with the lock() operation, we first check if the fulock is KCO; if so jump straight into the kernel (except if it is marked *not-recoverable*, in which case we fail).

In the case of the fast-path, we read the value of the vfulock; if it is looks like a cookie[19] then we try the fast-unlock, returning if successful. If it failed we retry from the beginning. When the value of the vfulock doesn't look like a cookie, we dive into the kernel, as it means that it is either dead or there are waiters (and thus the kernel handles it).

Note this unlock operation allows any thread to unlock the fulock, it doesn't need to be the owner.

Other operations

A `trylock()` operation is implemented in similar terms (please refer to the sample library code in the `fusyn-test` package, file `src/include/kernel-lock.h`; this package is available for download from the project's website).

Operations for manipulating or querying the

[19] The three values `VFULOCK_WP`, `VFULOCK_DEAD` and `VFULOCK_NR` are purposely chosen to be the last three values of the `unsigned` domain.

state of the fulock are implemented by calling the `ufulock_ctl()` system call directly, providing the vfulock and flags.

6 Integration with NPTL

The patches for integration with NPTL (that we call RTNPTL for short) allow any POSIX program to use these features, via a certain set of standard calls and ways to customize the operation mode of the fulock under the mutex's hood with other non-POSIX extensions.

RTNPTL uses the same or very similar user mode integration code than the one explained above, sitting down at the `lll_` layer in glibc. This code provides all the intended functionality only to the POSIX mutexes and conditional variables. Locks used internally by the library still need work (see the *future directions* section).

By default, RTNPTL provides non-robust fast-path enabled mutexes that unlock in automatic mode[20], without any priority inheritance and protection. However, by modifying the mutex attributes with the `pthread_mutexattr_set*()` calls, different parameters can be set:

- Manipulating the priority inversion protections:

  `pthread_mutexattr_setprotocol()` takes a mutex attribute and a protection protocol, `PTHREAD_PRIO_INHERIT` or `PTHREAD_PRIO_PROTECT`.

  `pthread_mutex_setprioceiling()` can be used to query and change the priority ceiling of a mutex.

  `pthread_mutexattr_setserial_np()` and `pthread_mutex_setserial_np()` allows setting the unlock method to use for lock-stealing avoidance out of `PTHREAD_MUTEX_SERIAL_NP`, `PTHREAD_MUTEX_PARALLEL_NP`, or `PTHREAD_MUTEX_AUTO_NP` (this one can be switched during the lifetime of the mutex).

- `pthread_mutexattr_setrobust_np()` enables robustness in the mutex to be. `pthread_mutex_setconsistency_np()` is used to heal or make *not-recoverable* a *dead-owner* mutex. The consistency state can be queried with `pthread_mutex_getconsistency_np()`.

- `pthread_mutexattr_setfast_np()` is used to select the use of a KCO fulock or not, effectively enabling/disabling fast-path operation.

The non-standard interfaces are still subject to some unlikely flux.

7 Current status and future direction

At the time of writing, the project has met most of the requirements that were set as targets, reaching stability and meeting performance goals of sub-millisecond latencies. The added overhead does not seem to affect too much compared to NPTL, being generally slightly slower.

Compatibility

We routinely test RTNPTL+fusyn by running:

[20] serialized or parallelized depending on the policy priority of the first waiter

- Miscellaneous multi-threaded applications (e.g.: Mozilla)
- SUN jdk-1.42_03 with SPECjbb2000[21].
- MySQL 2.23.58 with `super-smack` and `sql-bench`.

This has helped us to catch some bugs (with some pending for certain combinations) and to test the compatibility of our approach. Performance wise, no obvious differences have been found with plain NPTL running on futexes.

This set of macro benchmarks is incomplete and will be expanded in the future, time and resource availability permitting.

Latency

The current code performs fairly well latency wise (given the extra overhead). In an unloaded system[22], the latency of the serialized ownership change operation[23] is in the range of $60 \pm 10\mu s$. Adding some network load (ten simultaneous downloads of 40 MiB files) bumps it up to $110 \pm 10\mu s$. Simultaneous reading of 1 GiB from `/dev/hda` to `/dev/null` raises it up to $130 \pm 10\mu s$.

The code exposes a strange behavior when testing the ownership change latency in an unloaded system while increasing the number of waiters. The average latency stays stable for the first ten-to-fifteen waiters (threads of a single program) at around $18 \pm 10\mu s$ (see Figure 8).

However, when the number of queued waiters goes up to 2000 threads, the latency climbs up to $50 \pm 10\mu s$, stabilizing from there on, as seen on Figure 9. Of course this is an extremely unrealistic scenario, but it helps to test the scalability of the code, and nevertheless, we are trying to proof the root cause, being cache issues the most likely ones.

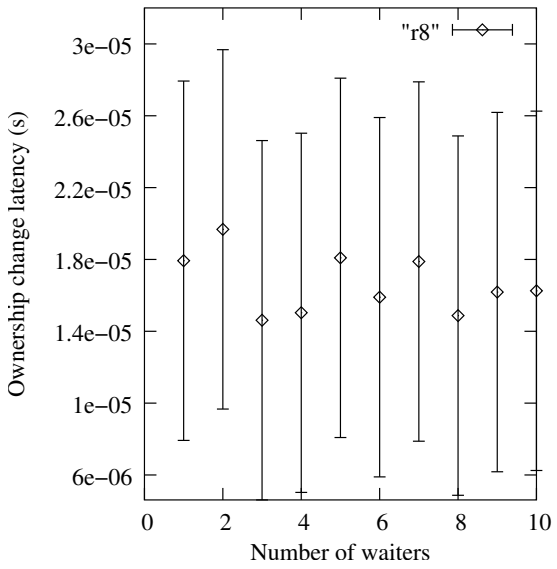


Figure 8: Scalability of the ownership-change latency *vs.* the number of waiters stays stable up until ten waiting threads.

.

Note: these numbers have been produced with a home-grown swiss-knife test program (to be published on the web site) called `ownership_change_latency`. Most of our timing efforts have concentrated in this particular case, although we have some other micro benchmarks planned.

Jitter

At this point, we haven't done yet any formal jitter studies.

Informally speaking, using the ownership change latency benchmark in unloaded systems, we have seen jitter increases over NPTL of about $1\mu s$, $0.3\mu s$ on a system fairly loaded with IDE and network traffic. However, bear in

[21] SPEC Java Business Benchmark 2000.

[22] as measured in a 2xP3 850 MHz 2.5 GiB RAM running version 2.3 of the code

[23] time since a serialized unlock is done until the first waiter gets the lock and executes.

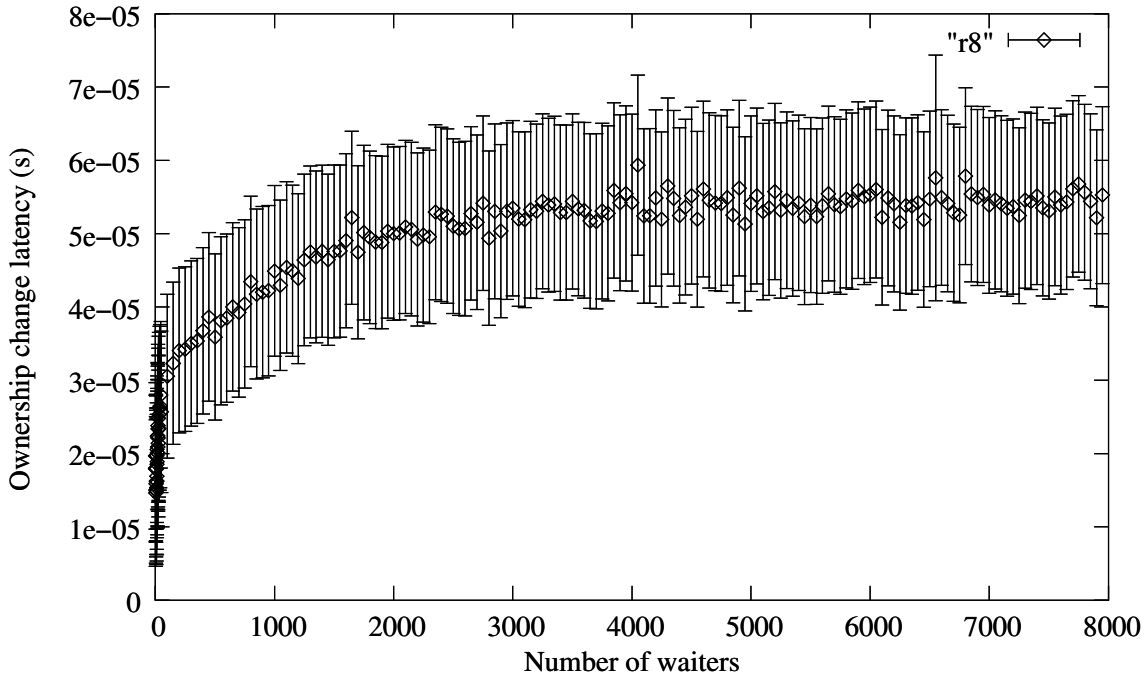


Figure 9: Scalability of the ownership-change latency *vs.* the number of waiters only stabilizes after two thousand waiters.

.

mind that these numbers are completely meaningless because the finest dependable clock resolution we can get (using the High Resolution Timers patch) is well higher, $10\mu s$. We can use them only to provide a hint.

Future direction

The project has reached an important milestone of maturity with the 2.2 release during the spring of 2004–nonetheless there is still much work to do. These some of the areas where we plan to target our future efforts:

- Some parties have asked for all these concepts (real-time, robustness, priority-protection) applied to read-write mutexes, much more complex than simple mutexes. We are still evaluation how worth is this.
- Some elusive bugs are still present.
- Accessing user space memory from the kernel by `kmapping` it poses some issues on architectures with *strange* cache consistency designs, such as some ARM and PA-RISC 8000. It is still not clear how to proceed for them and we would welcome any help.
- The kernel hash table for location of objects is a potential bottleneck in a system populated with many active user-space fusyn objects. We want to implement a proof of concept where a cookie identifying the object is placed in user space along the vfulock/vfuqueue. This cookie would consist of a two pointers crypted with two different keys by the kernel. In order to map a vfulock/vfuqueue to it's corresponding fusyn object, the kernel just has to decrypt the pointers. Having two crypted with different keys is used to enforce validity against garbage being writ-

ten by user space by mistake or to compromise the system.

- Providing a robust mutex infrastructure is OK, as long as it is used. Internally, glibc uses locks to protect many of its data structures—in order to be able to provide true robustness, we need to add robustness to those internal locks, as well as recovery strategies.

- Extend the coverage of our macro and micro benchmarks.

8 Downloading

The project maintains a website at:

`http://developer.osdl.org/dev/robustmutexes/`

from where all the current and older snapshots of the code can be obtained. As well, it offers pointers to the mailing list, bugzilla and CVS repositories.

We want to thank the Open Source Development Lab for making these resources available to us.

9 Conclusion

We have presented an infrastructure for providing real-time and robust synchronization services in the Linux kernel. We have been able to accomplish this with a minimum overhead impact over the current futex-based infrastructure and expect that it will be sufficient to satisfy the needs of multi-threaded, fault-proof and/or soft-real time designs.

10 Trademarks and acknowledgements

During development we kept discovering roadblocks, situations, and side effects we failed to spot or details we missed in the POSIX specifications. A lot of hair pulling that was counteracted by the thrill of the challenge, producing a love-and-hate relationship with the topic (and hence the title of this paper). We want to thank all of those who helped out by pointing out issues, contributing, reviewing, criticizing, and testing ideas and code.

References

[1] Mike Blasgen, Jim Gray, Mike Miltoma, and Tom Price. 1979. “The convoy phenomenon,” *ACM SIGOPS Operating Systems Review*, Volume 13, Issue 2 (April 1979). `http://portal.acm.org/citation.cfm?id=850659&jmp=cit&dl=GUIDE&dk=ACM`

[2] Arnd C. Heursch, Dirk Grambow, Dirk Roedel, and Helmut Rzehak. “Time-critical tasks in Linux 2.6,” `http://www.informatik.unibw-muenchen.de/inst3/index_de.php` Pages 6 and 7.

[3] Victor Yodaiken. 2001. “The dangers of priority inheritance,”

http://citeseer.ist.psdu.edu/yodaiken01dangers.html

[4] Hubertus Frankel, Rusty Russell, and Matthew Kirkwood. 2002. “Fuss, Futexes and Furwocks: Fast Userlevel Locking in Linux.” *Proceedings of the 2002 Ottawa Linux Symposium*, http://www.linux.org.uk/~ajh/ols2002_proceedings.pdf.gz Pages 479-495

Workload Dependent Performance Evaluation of the Linux 2.6 I/O Schedulers

Steven L. Pratt
IBM
slpratt@us.ibm.com

Dominique A. Heger
IBM
dheger@us.ibm.com

Abstract

The 2.6 release introduced the option to select a particular I/O scheduler at boot time. The 2.4 Linus elevator was retired, incorporated are now the anticipatory (AS), the deadline, the noop, as well as the completely fair queuing (CFQ) I/O schedulers. Each scheduler has its strengths and weaknesses. The question is under what workload scenarios does a particular I/O scheduler excel, as well as what is the performance gain that is possible by utilizing the available tuning options.

This study quantifies the performance of the 4 I/O schedulers under various workload scenarios (such as mail, web, and file server based conditions). The hardware is being varied from a single-CPU single-disk setup to machines with many CPUs that are utilizing large RAID arrays. In addition to characterizing the performance behavior and making actual recommendations on which scheduler to utilize under certain workload scenarios, the study looks into ways to actually improve the performance through either the existing tuning options or any potential code changes/enhancements.

Introduction

This study was initiated to quantify I/O performance in a Linux 2.6 environment. The I/O stack in general has become considerably more complex over the last few years. Contemporary I/O solutions include hardware, firmware, as well as software support for features such as request coalescing, adaptive prefetching, automated invocation of direct I/O, or asynchronous write-behind polices. From a hardware perspective, incorporating large cache subsystems on a memory, RAID controller, and physical disk layer allows for a very aggressive utilization of these I/O optimization techniques. The interaction of the different optimization methods that are incorporated in the different layers of the I/O stack is neither well understood nor been quantified to an extent necessary to make a rational statement on I/O performance. A rather interesting feature of the Linux operating system is the I/O scheduler [6]. Unlike the CPU scheduler, an I/O scheduler is not a necessary component of any operating system per se, and therefore is not an actual building block in some of the commercial UNIX® systems. This study elaborates how the I/O scheduler is embedded into the Linux I/O framework, and discusses the 4 (rather distinct) implementations and performance behaviors of the I/O schedulers that are available in Linux 2.6. Section 1 introduces the BIO layer, whereas Section 2 elaborates on the anticipatory (AS), the deadline, the noop, as well as the completely fair queuing (CFQ) I/O schedulers. Section 2 further highlights some of the performance issues that may surface based on which I/O scheduler is being utilized. Section 3 discusses some additional

hardware and software components that impact I/O performance. Section 4 introduces the workload generator used in this study and outlines the methodology that was utilized to conduct the analysis. Section 5 discusses the results of the project. Section 6 provides some additional recommendations and discusses future work items.

1 I/O Scheduling and the BIO Layer

The I/O scheduler in Linux forms the interface between the generic block layer and the low-level device drivers [2],[7]. The block layer provides functions that are utilized by the file systems and the virtual memory manager to submit I/O requests to block devices. These requests are transformed by the I/O scheduler and made available to the low-level device drivers. The device drivers consume the transformed requests and forward them (by using device specific protocols) to the actual device controllers that perform the I/O operations. As prioritized resource management seeks to regulate the use of a disk subsystem by an application, the I/O scheduler is considered an imperative kernel component in the Linux I/O path. It is further possible to regulate the disk usage in the kernel layers above and below the I/O scheduler. Adjusting the I/O pattern generated by the file system or the virtual memory manager (VMM) is considered as an option. Another option is to adjust the way specific device drivers or device controllers consume and manipulate the I/O requests.

The various Linux 2.6 I/O schedulers can be abstracted into a rather generic I/O model. The I/O requests are generated by the block layer on behalf of threads that are accessing various file systems, threads that are performing raw I/O, or are generated by virtual memory management (VMM) components of the kernel such as the kswapd or the pdflush threads. The producers of I/O requests initiate a call to `__make_request()`, which invokes various I/O scheduler functions such as `elevator_merge_fn()`. The enqueue functions in the I/O framework intend to merge the newly submitted block I/O unit (a bio in 2.6 or a `buffer_head` in the older 2.4 kernel) with previously submitted requests, and to sort (or sometimes just insert) the request into one or more internal I/O queues. As a unit, the internal queues form a single logical queue that is associated with each block device. At a later stage, the low-level device driver calls the generic kernel function `elv_next_request()` to obtain the next request from the logical queue. The `elv_next_request()` call interacts with the I/O scheduler's dequeue function `elevator_next_req_fn()`, and the latter has an opportunity to select the appropriate request from one of the internal queues. The device driver processes the request by converting the I/O submission into (potential) scatter-gather lists and protocol-specific commands that are submitted to the device controller. From an I/O scheduler perspective, the block layer is considered as the producer of I/O requests and the device drivers are labeled as the actual consumers.

From a generic perspective, every read or write request launched by an application results in either utilizing the respective I/O system calls or in memory mapping (mmap) the file into a process's address space [14]. I/O operations normally result in allocating PAGE_SIZE units of physical memory. These pages are being indexed, as this enables the system to later on locate the page in the buffer cache [10]. A cache subsystem only improves performance if the data in the cache is being reused. Further, the read cache abstraction allows the system to implement (file system dependent) read-ahead functionalities, as well as to construct large contiguous (SCSI) I/O commands that

can be served via a single direct memory access (DMA) operation. In circumstances where the cache represents pure (memory bus) overhead, I/O features such as direct I/O should be explored (especially in situations where the system is CPU bound).

In a general write scenario, the system is not necessarily concerned with the previous content of a file, as a `write()` operation normally results in overwriting the contents in the first place. Therefore, the write cache emphasizes other aspects such as asynchronous updates, as well as the possibility of omitting some write requests in the case where multiple `write()` operations into the cache subsystem result in a single I/O operation to a physical disk. Such a scenario may occur in an environment where updates to the same (or a similar) inode offset are being processed within a rather short time-span. The block layer in Linux 2.4 is organized around the `buffer_head` data structure [7]. The culprit of that implementation was that it is a daunting task to create a truly effective block I/O subsystem if the underlying `buffer_head` structures force each I/O request to be decomposed into 4KB chunks. The new representation of the block I/O layer in Linux 2.6 encourages large I/O operations. The block I/O layer now tracks data buffers by using struct page pointers. Linux 2.4 systems were prone to loose sight of the logical form of the writeback cache when flushing the cache subsystem. Linux 2.6 utilizes logical pages attached to inodes to flush dirty data, which allows multiple pages that belong to the same inode to be coalesced into a single bio that can be submitted to the I/O layer [2]. This approach represents a process that works well if the file is not fragmented on disk.

2 The 2.6 Deadline I/O Scheduler

The deadline I/O scheduler incorporates a per-request expiration-based approach and operates on 5 I/O queues [4]. The basic idea behind the implementation is to aggressively reorder requests to improve I/O performance while simultaneously ensuring that no I/O request is being starved. More specifically, the scheduler introduces the notion of a per-request deadline, which is used to assign a higher preference to read than write requests. The scheduler maintains 5 I/O queues. During the enqueue phase, each I/O request gets associated with a deadline, and is being inserted in I/O queues that are either organized by the starting logical block number (a sorted list) or by the deadline factor (a FIFO list). The scheduler incorporates separate sort and FIFO lists for read and write requests, respectively. The 5th I/O queue contains the requests that are to be handed off to the device driver. During a dequeue operation, in the case where the dispatch queue is empty, requests are moved from one of the 4 (sort or FIFO) I/O lists in batches. The next step consists of passing the head request on the dispatch queue to the device driver (this scenario also holds true in the case that the dispatch-queue is not empty). The logic behind moving the I/O requests from either the sort or the FIFO lists is based on the scheduler's goal to ensure that each read request is processed by its effective deadline, without starving the queued-up write requests. In this design, the goal of economizing the disk seek time is accomplished by moving a larger batch of requests from the sort list (logical block number sorted), and balancing it with a controlled number of requests from the FIFO list. Hence, the ramification is that the deadline I/O scheduler effectively emphasizes average read request response time over disk utilization and total average I/O request response time.

To reiterate, the basic idea behind the deadline

scheduler is that all read requests are satisfied within a specified time period. On the other hand, write requests do not have any specific deadlines associated with them. As the block device driver is ready to launch another disk I/O request, the core algorithm of the deadline scheduler is invoked. In a simplified form, the fist action being taken is to identify if there are I/O requests waiting in the dispatch queue, and if yes, there is no additional decision to be made what to execute next. Otherwise it is necessary to move a new set of I/O requests to the dispatch queue. The scheduler searches for work in the following places, BUT will only migrate requests from the first source that results in a hit. (1) If there are pending write I/O requests, and the scheduler has not selected any write requests for a certain amount of time, a set of write requests is selected (see tunables in Appendix A). (2) If there are expired read requests in the `read_fifo` list, the system will move a set of these requests to the dispatch queue. (3) If there are pending read requests in the sort list, the system will migrate some of these requests to the dispatch queue. (4) As a last resource, if there are any pending write I/O operations, the dispatch queue is being populated with requests from the sorted write list. In general, the definition of a certain amount of time for write request starvation is normally 2 iterations of the scheduler algorithm (see Appendix A). After two sets of read requests have been moved to the dispatch queue, the scheduler will migrate some write requests to the dispatch queue. A set or batch of requests can be (as an example) 64 contiguous requests, but a request that requires a disk seek operation counts the same as 16 contiguous requests.

2.1 The 2.6 Anticipatory I/O scheduler

The anticipatory (AS) I/O scheduler's design attempts to reduce the per thread read response time. It introduces a controlled delay component into the dispatching equation [5],[9],[11]. The delay is being invoked on any new read request to the device driver, thereby allowing a thread that just finished its read I/O request to submit a new read request, basically enhancing the chances (based on locality) that this scheduling behavior will result in smaller seek operations. The tradeoff between reduced seeks and decreased disk utilization (due to the additional delay factor in dispatching a request) is managed by utilizing an actual cost-benefit analysis [9].

The next few paragraphs discuss the general design of an anticipatory I/O scheduler, outlining the different components that comprise the I/O framework. Basically, as a read I/O request completes, the I/O framework stalls for a brief amount of time, awaiting additional requests to arrive, before dispatching a new request to the disk subsystem. The focus of this design is on applications threads that rapidly generate another I/O request that could potentially be serviced before the scheduler chooses another task, and by doing so, deceptive idleness may be avoided [9]. Deceptive idleness is defined as a condition that forces the scheduler into making a decision too early, basically by assuming that the thread issuing the last request has momentarily no further disk request lined up, and hence the scheduler selects an I/O request from another task. The design discussed here argues that the fact that the disk remains idle during the short stall period is not necessarily detrimental to I/O performance. The question of whether (and for how long) to wait at any given decision point is key to the effectiveness and performance of the implementation. In practice, the framework waits for the shortest possible period of time for which the scheduler expects (with a high probability) the benefits of actively waiting to outweigh the costs of keeping the disk subsystem in an idle state. An assessment of the costs and benefits is only pos-

sible relative to a particular scheduling policy [11]. To elaborate, a seek reducing scheduler may wish to wait for contiguous or proximal requests, whereas a proportional-share scheduler may prefer weighted fairness as one of its primary criteria. To allow for such a high degree of flexibility, while trying to minimize the burden on the development efforts for any particular disk scheduler, the anticipatory scheduling framework consists of 3 components [9]. (1) The original disk scheduler, which implements the scheduling policy and is unaware of any anticipatory scheduling techniques. (2) An actual scheduler independent anticipation core. (3) An adaptive scheduler-specific anticipation heuristic for seek reducing (such as SPTF or C-SCAN) as well as any potential proportional-share (CFQ or YFQ) scheduler. The anticipation core implements the generic logic and timing mechanisms for waiting, and relies on the anticipation heuristic to decide if and for how long to wait. The actual heuristic is implemented separately for each disk scheduler, and has access to the internal state of the scheduler. To apply anticipatory scheduling to a new scheduling policy, it is merely necessary to implement an appropriate anticipation heuristic.

Any traditional work-conserving I/O scheduler operates in two states (known as idle and busy). Applications may issue I/O requests at any time, and these requests are normally being placed into the scheduler's pool of requests. If the disk subsystem is idle at this point, or whenever another request completes, a new request is being scheduled, the scheduler's select function is called, whereupon a request is chosen from the pool and dispatched to the disk device driver. The anticipation core forms a wrapper around this traditional scheduler scheme. Whenever the disk becomes idle, it invokes the scheduler to select a candidate request (still basically following the same philosophy as always). However, instead of dequeuing and dispatching a request immediately, the framework first passes the request to the anticipation heuristic for evaluation. A return value (result) of zero indicates that the heuristic has deemed it pointless to wait and the core therefore proceeds to dispatch the candidate request. However, a positive integer as a return value represents the waiting period in microseconds that the heuristic deems suitable. The core initiates a timeout for that particular time period, and basically enters a new wait state. Though the disk is inactive, this state is considered different from idling (while having pending requests and an active timeout). If the timeout expires before the arrival of any new request, the previously chosen request is dispatched without any further delay. However, new requests may arrive during the wait period and these requests are added to the pool of I/O requests. The anticipation core then immediately requests the scheduler to select a new candidate request from the pool, and initiates communication with the heuristic to evaluate this new candidate. This scenario may lead to an immediate dispatch of the new candidate request, or it may cause the core to remain in the wait state, depending on the scheduler's selection and the anticipation heuristic's evaluation. In the latter case, the original timeout remains in effect, thus preventing unbounded waiting situations by repeatedly re-triggering the timeout.

As the heuristic being used is disk scheduler dependent, the discussion here only generalizes on the actual implementation techniques that may be utilized. Therefore, the next few paragraphs discuss a shortest positioning time first (SPTF) based implementation, where the disk scheduler determines the positioning time for each available request based on the current head position, and basically chooses the request that results into the shortest seek distance. In general, the heuristic has to evaluate the candidate request that was chosen by the scheduling policy. The intuition is that if

the candidate I/O request is located close to the current head position, there is no need to wait on any other requests. Assuming synchronous I/O requests initiated by a single thread, the task that issued the last request is likely to submit the next request soon, and if this request is expected to be close to the current request, the heuristic decides to wait for this request [11]. The waiting period is chosen as the expected YZ percentile (normally around 95%) think-time, within which there is a XZ probability (again normally 95%) that a request will arrive. This simple approach is transformed and generalized into a succinct cost-benefit equation that is intended to cover the entire range of values for the head positioning, as well as the think-times. To simplify the discussion, the adaptive component of the heuristic consists of collecting online statistics on all the disk requests to estimate the different time variables that are being used in the decision making process. The expected positioning time for each process represents a weighted-average over the time of the positing time for requests from that process (as measured upon request completion). Expected median and percentile think-times are estimated by maintaining a decayed frequency table of request think-times for each process.

The Linux 2.6 implementation of the anticipatory I/O scheduler follows the basic idea that if the disk drive just operated on a read request, the assumption can be made that there is another read request in the pipeline, and hence it is worth while to wait [5]. As discussed, the I/O scheduler starts a timer, and at this point there are no more I/O requests passed down to the device driver. If a (close) read request arrives during the wait time, it is serviced immediately and in the process, the actual distance that the kernel considers as close grows as time passes (the adaptive part of the heuristic). Eventually the close requests will dry out and the scheduler will decide to submit some of the write requests (see Appendix A).

2.2 The 2.6 CFQ Scheduler

The Completely Fair Queuing (CFQ) I/O scheduler can be considered to represent an extension to the better known Stochastic Fair Queuing (SFQ) implementation [12]. The focus of both implementations is on the concept of fair allocation of I/O bandwidth among all the initiators of I/O requests. An SFQ-based scheduler design was initially proposed (and ultimately being implemented) for some network scheduling related subsystems. The goal to accomplish is to distribute the available I/O bandwidth as equally as possible among the I/O requests. The implementation utilizes n (normally 64) internal I/O queues, as well as a single I/O dispatch queue. During an enqueue operation, the PID of the currently running process (the actual I/O request producer) is utilized to select one of the internal queues (normally hash based) and hence, the request is basically inserted into one of the queues (in FIFO order). During dequeue, the SFQ design calls for a round robin based scan through the non-empty I/O queues, and basically selects requests from the head of the queues. To avoid encountering too many seek operations, an entire round of requests is collected, sorted, and ultimately merged into the dispatch queue. In a next step, the head request in the dispatch queue is passed to the device driver. Conceptually, a CFQ implementation does not utilize a hash function. Therefore, each I/O process gets an internal queue assigned (which implies that the number of I/O processes determines the number of internal queues). In Linux 2.6.5, the CFQ I/O scheduler utilizes a hash function (and a certain amount of request queues) and therefore resembles an SFQ implementation. The CFQ, as well as the SFQ implementations strives to manage per-process I/O bandwidth, and provide fairness at the level of pro-

cess granularity.

2.3 The 2.6 noop I/O scheduler

The Linux 2.6 noop I/O scheduler can be considered as a rather minimal overhead I/O scheduler that performs and provides basic merging and sorting functionalities. The main usage of the noop scheduler revolves around non disk-based block devices (such as memory devices), as well as specialized software or hardware environments that incorporate their own I/O scheduling and (large) caching functionality, and therefore require only minimal assistance from the kernel. Therefore, in large I/O subsystems that incorporate RAID controllers and a vast number of contemporary physical disk drives (TCQ drives), the noop scheduler has the potential to outperform the other 3 I/O schedulers as the workload increases.

2.4 I/O Scheduler—Performance Implications

The next few paragraphs augment on the I/O scheduler discussion, and introduce some additional performance issues that have to be taken into consideration while conducting an I/O performance analysis. The current AS implementation consists of several different heuristics and policies that basically determine when and how I/O requests are dispatched to the I/O controller(s). The elevator algorithm that is being utilized in AS is similar to the one used for the deadline scheduler. The main difference is that the AS implementation allows limited backward movements (in other words supports backward seek operations) [1]. A backward seek operation may occur while choosing between two I/O requests, where one request is located behind the elevator's current head position while the other request is ahead of the elevator's current position.

The AS scheduler utilizes the lowest logical block information as the yardstick for sorting, as well as determining the seek distance. In the case that the seek distance to the request behind the elevator is less than half the seek distance to the request in front of the elevator, the request behind the elevator is chosen. The backward seek operations are limited to a maximum of MAXBACK (1024 * 1024) blocks. This approach favors the forward movement progress of the elevator, while still allowing short backward seek operations. The expiration time for the requests held on the FIFO lists is tuneable via the parameter's `read_expire` and `write_expire` (see Appendix A). When a read or a write operation expires, the AS I/O scheduler will interrupt either the current elevator sweep or the read anticipation process to service the expired request(s).

2.5 Read and Write Request Batches

An actual I/O batch is described as a set of read or write requests. The AS scheduler alternates between dispatching either read or write batches to the device driver. In a read scenario, the scheduler submits read requests to the device driver, as long as there are read requests to be submitted, and the read batch time limit (`read_batch_expire`) has not been exceeded. The clock on `read_batch_expire` only starts in the case that there are write requests pending. In a write scenario, the scheduler submits write requests to the device driver as long as there are pending write requests, and the write batch time limit `write_batch_expire` has not been exceeded. The heuristic used insures that the length of the write batches will gradually be shortened if there are read batches that frequently exceed their time limit.

When switching between read and write requests, the scheduler waits until all the requests from the previous batch are completed before scheduling any new requests. The read

and write FIFO expiration time is only being checked when scheduling I/O for a batch of the corresponding (read or write) operation. To illustrate, the read FIFO timeout values are only analyzed while operating on read batches. Along the same lines, the write FIFO timeout values are only consulted while operating on write batches. Based on the used heuristics and policies, it is generally not recommended to set the read batch time to a higher value than the write expiration time, or to set the write batch time to a greater value than the read expiration time. As the IO scheduler switches from a read to a write batch, the I/O framework launches the elevator with the head request on the write expired FIFO list. Likewise, when switching from a write to a read batch, the I/O scheduler starts the elevator with the first entry on the read expired FIFO list.

2.6 Read Anticipation Heuristic

The process of read anticipation solely occurs when scheduling a batch of read requests. The AS implementation only allows one read request at a time to be dispatched to the controller. This has to be compared to either the many write request scenario or the many read request case if read anticipation is deactivated. In the case that read anticipation is enabled (`antic_expire = 0`), read requests are dispatched to the (disk or RAID) controller one at a time. At the end of each read request, the I/O scheduler examines the next read request from the sorted read list (an actual rb-tree) [1]. If the next read request belongs to the same process as the request that just completed, or if the next request in the queue is close (data block wise) to the just completed request, the request is being dispatched immediately. Otherwise, the statistics (average think-time and seek distance) available for the process that just completed are being examined (cost-benefit analysis). The statistics are associated with each process, but these statistics are not associated with a specific I/O device *per se*. To illustrate, the approach works more efficiently if there is a one-to-one correlation between a process and a disk. In the case that a process is actively working I/O requests on separate devices, the actual statistics reflect a combination of the I/O behavior across all the devices, skewing the statistics and therefore distorting the facts. If the AS scheduler guesses right, very expensive seek operations can be omitted, and hence the overall I/O throughput will benefit tremendously. In the case that the AS scheduler guesses wrong, the `antic_expire` time is wasted. In an environment that consists of larger (HW striped) RAID systems and tag command queuing (TCQ) capable disk drives, it is more beneficial to dispatch an entire batch of read requests and let the controllers and disk do their magic.

From a physical disk perspective, to locate specific data, the disk drive's logic requires the cylinder, the head, and the sector information [17]. The cylinder specifies the track on which the data resides. Based on the layering technique used, the tracks underneath each other form a cylinder. The head information identifies the specific read/write head (and therefore the exact platter). The search is now narrowed down to a single track on a single platter. Ultimately, the sector value reflects the sector on the track, and the search is completed. Contemporary disk subsystems do not communicate in terms of cylinders, heads and sectors. Instead, modern disk drives map a unique block number over each cylinder/head/sector construct. Therefore, that (unique) reference number identifies a specific cylinder/head/sector combination. Operating systems address the disk drives by utilizing these block numbers (logical block addressing), and hence the disk drive is responsible for translating the block number into the appropriate cylinder/head/sector value. The culprit is

that it is not guaranteed that the physical mapping is actually sequential. But the statement can be made that there is a rather high probability that a logical block *n* is physically adjacent to a logical block *n+1*. The existence of the discussed sequential layout is paramount to the I/O scheduler performing as advertised. Based on how the read anticipatory heuristic is implemented in AS, I/O environments that consist of RAID systems (operating in a hardware stripe setup) may experience a rather erratic performance behavior. This is due to the current AS implementation that is based on the notion that an I/O device has only one physical (seek) head, ignoring the fact that in a RAID environment, each physical disk has its own physical seek head construct. As this is not recognized by the AS scheduler, the data being used for the statistics analysis is skewed. Further, disk drives that support TCQ perform best when being able to operate on *n* (and not 1) pending I/O requests. The read anticipatory heuristic basically disables TCQ. Therefore, environments that support TCQ and/or consist of RAID systems may benefit from either choosing an alternate I/O scheduler or from setting the `antic_expire` parameter to 0. The tuning allows the AS scheduler to behave similarly to the deadline I/O scheduler (the emphasis is on behave and not performance).

3 I/O Components that Affect Performance

In any computer system, between the disk drives and the actual memory subsystem is a hierarchy of additional controllers, host adapters, bus converters, and data paths that all impact I/O performance in one way or another [17]. Linux file systems submit I/O requests by utilizing `submit_bio()`. This function submits requests by utilizing the request function as specified during queue creation. Technically, device drivers do not have to use the I/O scheduler, however all SCSI devices in Linux utilize the scheduler by virtue of the SCSI mid-layer [1]. The `scsi_alloc_queue()` function calls `blk_init_queue()`, which sets the request function to `scsi_request_fn()`. The `scsi_request_fn()` function takes requests from the I/O scheduler (on dequeue), and passes them down to the device driver.

3.1 SCSI Operations

In the case of a simple SCSI disk access, the request has to be processed by the server, the SCSI host adapter, the embedded disk controller, and ultimately by the disk mechanism itself. As the OS receives the I/O request, it converts the request into a SCSI command packet. In the case of a synchronous request, the calling thread surrenders the CPU and transitions into a sleep state until the I/O operation is completed. In a next step, the SCSI command is transferred across the server's I/O bus to the SCSI host adapter. The host adapter is responsible for interacting with the target controller and the respective devices. In a first step, the host adapter selects the target by asserting its control line onto the SCSI-bus (as the bus becomes available). This phase is known as the SCSI selection period. As soon as the target responds to the selection process, the host adapter transfers the SCSI command to the target. This section of the I/O process is labeled as the command phase. If the target is capable of processing the command immediately, it either returns the requested data or the status information.

In most circumstances, the request can only be processed immediately if the data is available in the target controller's cache. In the case of a `read()` request, the data is normally not available. This results into the target disconnecting from the SCSI bus to allow other SCSI operations to be processed. If the I/O opera-

tion consists of a `write()` request, the data phase is followed immediately by a command phase on the bus, as the data is transferred into the target's cache. At that stage, the target disconnects from the bus. After disconnecting from the bus, the target resumes its own processing while the bus can be utilized by other SCSI requests. After the physical I/O operation is completed on the target disk, the target controller competes again for the bus, and reconnects as soon as the bus is available. The reconnect phase is followed by a data phase (in the case of `read()` operation) where the data is actually being moved. The data phase is followed by another status phase to describe the results of the I/O operation. As soon as the SCSI host adapter receives the status update, it verifies the proper completion of the request and notifies the OS to interrupt the requesting worker thread. Overall, the simple SCSI I/O request causes 7 phase changes consisting of a select, a command, a disconnect, a reconnect, a data, a status, and a disconnect operation. Each phase consumes time and contributes to the overall I/O processing latency on the system.

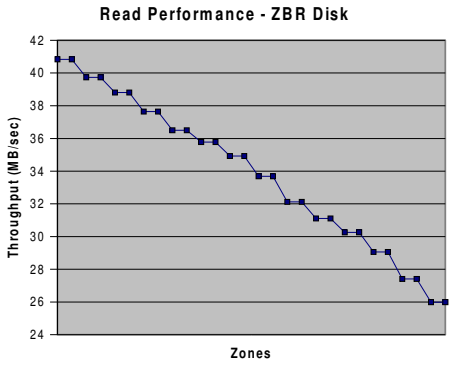
3.2 SCSI Disk Fence

When discussing SCSI disks, it is imperative to understand the performance impact of a relatively obscure disk control parameter that is labeled as the fence. When a SCSI disk recognizes a significant delay (such as a seek operation) in a `read()` request, the disk will surrender the bus. At the point where the disk is ready to transfer the data, the drive will again contend for the bus so that the `read()` request can be completed. The fence parameter determines the time at which the disk will begin to contend for the SCSI bus. If the fence is set to 0 (the minimum), the disk will contend for the SCSI bus after the first sector has been transferred into the disk controller's memory. In the case where the fence is set to 255 (the maximum), the disk will wait until almost all the requested data has been accumulated in the controller's memory before contending for the bus.

The performance implication of setting the fence to a low value is a reduced response time, but results in a data transfer that happens basically at disk speed. On the other hand, a high fence value will delay the start of the data transfer, but results in a data transfer that occurs at near burst speed. Therefore, in systems with multiple disks per adapter, a high fence value potentially increases overall throughput for I/O intensive workloads. A study by Shriver [15] observed fairness in servicing sufficiently large I/O requests (in the 16KB to 128KB range), despite the fact that the SCSI disks have different priorities when contending for the bus. Although each process attempts to progress through its requests without any coordination with other processes, a convoy behavior among all the processes was observed. Namely, all disk drives received a request and transmitted the data back to the host adapter before any disk received another request from the adapter (a behavior labeled as rounds). The study revealed that the host adapter does not arbitrate for the bus, despite having the highest priority, as long as any disk is arbitrating.

3.3 Zone Bit Recording (ZBR)

Contemporary disk drives utilize a technology called Zone Bit Recording to increase capacity [17]. Incorporating the technology, cylinders are grouped into zones, based on their distance from the center of the disk. Each zone is assigned a number of sectors per track. The outer zones contain more sectors per track compared to the inner zones that are located closer to the spindle. With ZBR disks, the actual data transfer rate varies depending on the physical sector location.



Note: Figure 1 depicts the average throughput per zone, the benchmark revealed 14 distinct performance steps.

Figure 1: ZBR Throughput Performance

Given the fact that a disk drive spins at a constant rate, the outer zones that contain more sectors will transfer data at a higher rate than the inner zones that contain fewer sectors. In this study, evaluating I/O performance on an 18.4 GB Seagate ST318417W disk drive outlined the throughput degradation for sequential read() operations based on physical sector location. The ZCAV program used in this experiment is part of the Bonnie++ benchmark suite. Figure 1 outlines the average zone read() throughput performance. It has to be pointed out that the performance degradation is not gradual, as the benchmark results revealed 14 clear distinct performance steps along the throughput curve. Another observation derived from the experiment was that for this particular ZBR disk, the outer zones revealed to be wider than the inner zones. The Seagate specifications for this particular disk cite an internal transfer rate of 28.1 to 50.7 MB/second. The measured minimum and maximum throughput `read()` values of 25.99 MB/second and 40.84 MB/second, respectively are approximately 8.1% and 19.5% (13.8% on average) lower, and represent actual throughput rates. Benchmarks conducted on 4 other ZBR drives revealed a similar picture. On average, the actual system throughput rates were 13% to 15% lower than what was cited in the vendor specifications. Based on the conducted research, this text proposes a first-order ZBR approximation nominal disk transfer rate model (for a particular request size *req* and a disk capacity *cap*) that is defined in Equation 1 as:

$$ntr_{zbr}(req) = 0.85 \cdot \left(tr_{max}\right) - \frac{req \cdot \left(tr_{max} - tr_{min}\right)}{cap} \quad (1)$$

tr_{max} = maximum disk specific internal transfer speed

tr_{min} = minimum disk specific internal transfer speed

The suggested throughput regulation factor of 0.85 was derived from the earlier observation that throughput rates adjusted for factors such as sector overhead, error correction, or track and cylinder skewing issues resulted in a drop of approximately 15% compared to the manufacturer reported transfer rates. This study argues that the manufacturer reported transfer rates could be more accurately defined as instantaneous bit rates at the read-write heads. It has to be emphasized that the calculated throughput rates derived from the presented model will have to be adjusted onto the target system's ability to sustain the I/O rate.

The theories of progressive chaos imply that anything that evolves out of a perfect order will over time become disordered due to outside forces. The progressive chaos concept can certainly be applied to I/O performance. The dynamic allocation (as well as de-allocation) of file system resources contributes to the progressive chaos scenario encountered in virtually any file system designs. Form a device driver and physical disk drive perspective, the results of disk access optimization strategies

are first, that the number of transactions per second is maximized and second, that the order in which the requests are being received is not necessarily the order the requests are getting processed. Thus, the response time of any particular request can not be guaranteed. A request queue may increase spatial locality by selecting requests in an order to minimize the physical arm movement (a workload transformation), but may also increase the perceived response time because of queuing delays (a behavior transformation). The argument made in this study is that the interrelationship of some the discussed I/O components has to be taken into consideration while evaluating and quantifying performance

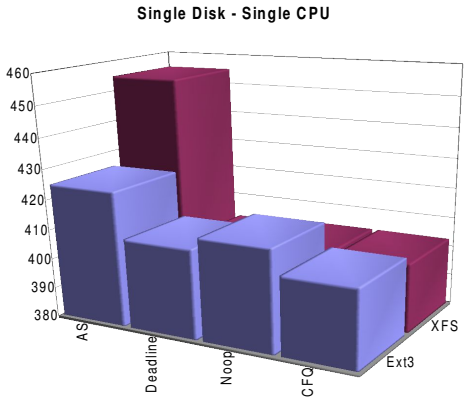
4 I/O Schedulers and Performance

The main goal of this study was to quantify I/O performance (focusing on the Linux 2.6 I/O schedulers) under varying workload scenarios and hardware configurations. Therefore, the benchmarks were conducted on a single-CPU single-disk system, a midrange 8-way NUMA RAID-5 system, and a 16-way SMP system that utilized a 28-disk RAID-0 configuration. The reader is referred to Appendix B for a more detailed description of the different benchmark environments. As a workload generator, the study utilized the flexible file system benchmark (FFSB) infrastructure [8]. FFSB represents a benchmarking environment that allows analyzing I/O performance by simulating basically any I/O pattern imaginable. The benchmarks can be executed on multiple individual file systems, utilizing an adjustable number of worker threads, where each thread may either operate out of a combined or a thread-based I/O profile. Aging the file systems, as well as collecting systems utilization and throughput statistics is part of the benchmarking framework. Next to the more traditional sequential read and sequential write benchmarks, the study used a filer server, a web server, a mail server, as well as a metadata intensive I/O profile (see Appendix B). The file, as well as the mail server workloads (the actual transaction mix) was based on Intel's Iometer benchmark [18], whereas the mail server transaction mix was loosely derived from the SPECmail2001 I/O profile [19]. The I/O analysis in this study was composed of two distinct focal points. One emphasis of the study was on aggregate I/O performance achieved across the 4 benchmarked workload profiles, whereas a second emphasis was on the sequential read and write performance behavior. The emphasis on aggregate performance across the 4 distinct workload profiles is based on the claim made that an I/O scheduler has to provide adequate performance in a variety of workload scenarios and hardware configurations, respectively. All the conducted benchmarks were executed with the default tuning values (if not specified otherwise) in an ext3 as well as an xfs file system environment. In this paper, the term response time represents the total run time of the actual FFSB benchmark, incorporating all the I/O operations that are executed by the worker threads.

5 Single-CPU Single-Disk Setup

The normalized results across the 4 workload profiles revealed that the deadline, the noop, as well as the CFQ schedulers performed within 2% and 1% percent on ext3 and xfs (see Figure 2). On ext3, the CFQ scheduler had a slight advantage, whereas on xfs the deadline scheduler provided the best aggregate (normalized) response time. On both file systems, the AS scheduler represented the least efficient solution, trailing the other I/O schedulers by 4.6% and 13% on ext3 and xfs, respectively. Not surprisingly, among the 4 workloads benchmarked in a single disk system, AS trailed the other 3 I/O schedulers by a rather significant

margin in the Web Server scenario (which reflects 100% random read operations). On sequential read operations, the AS scheduler outperformed the other 3 implementations by an average of 130% and 127% on ext3 and xfs. The sequential read results clearly support the discussion in this paper on where the design focus for AS was directed. In the case of sequential write operations, AS revealed the most efficient solution on ext3, whereas the noop scheduler provided the best throughput on xfs. The performance delta (for the sequential write scenarios) among the I/O schedulers was 8% on ext3 and 2% on xfs (see Appendix C).

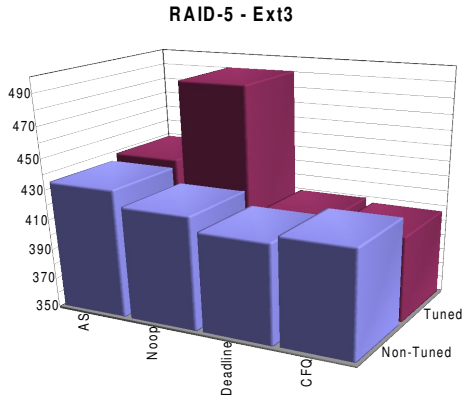


Note: In Figure 2, the x-axis depicts the I/O schedulers. The front row reflects the ext3 setup, whereas the back row shows xfs. The y-axis discloses the aggregate (normalized) response time over the 4 benchmarked profiles per I/O scheduler.

Figure 2: Aggregate Response Time (Normalized)

5.1 8-Way RAID-5 Setup

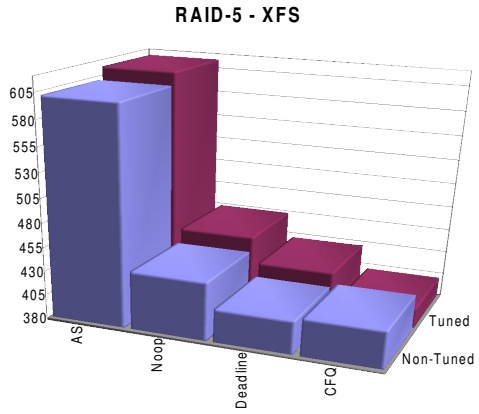
In the RAID-5 environment, the normalized response time values (across the 4 profiles) disclosed that the deadline scheduler provided the most efficient solution on ext3 as well as xfs (see Figure 3 and Figure 4). While executing in an ext3 environment, all 4 I/O schedulers were within 4.5%, with the AS I/O scheduler trailing noop and CFQ by approximately 2.5%. On xfs, the study clearly disclosed a profound AS I/O inefficiency while executing the metadata benchmark. The delta among the schedulers on xfs was much larger than on ext3, as the CFQ, noop, and AS implementations trailed the deadline scheduler by 1%, 6%, and 145%, respectively (see Appendix C). As in the single disk setup, the AS scheduler provided the most efficient sequential read performance. The gap between AS and the other 3 implementations shrunk though rather significantly compared to the single disk scenarios. The average sequential read throughput (for the other 3 schedulers) was approximately 20% less on both ext3 and xfs, respectively. The sequential write performance was dominated by the CFQ scheduler's response time that outperformed the other 3 solutions. The delta between the most (CFQ) and the least efficient implementation was 22% (AS) and 15% (noop) on ext3 and xfs, respectively (see Appendix C).



Note: In Figure 3, the x-axis depicts the I/O schedulers. The front-row reflects the non-tuned, and the back-row the tuned environments. The y-axis discloses the normalized response time (over the 4 profiles) per I/O scheduler.

Figure 3: EXT3 Aggregate Response Time (Normalized)

In a second phase, all the I/O scheduler setups were tuned by adjusting the (per block device)



Note: In Figure 4, the x-axis depicts the I/O schedulers. The front-row reflects the non-tuned, and the back-row the tuned environments. The y-axis discloses the normalized response time (over the 4 profiles) per I/O scheduler.

Figure 4: XFS Aggregate Response Time (Normalized)

tunable `nr_requests` (I/O operations in fly) from its default value of 128 to 2,560. The results revealed that the CFQ scheduler reacted in a rather positive way to the adjustment, and ergo was capable to provide on ext3 as well as on xfs the most efficient solution. The tuning resulted into decreasing the response time for CFQ in all the conducted (workload profile based) benchmarks on both file systems (see Appendix C). While CFQ benefited from the tuning, the results for the other 3 implementations were inconclusive. Based on the profile, the tuning either resulted in a gain or a loss in performance. As CFQ is designed to operate on larger sets of I/O requests, the results basically reflect the design goals of the scheduler [1]. This is in contrast to the AS implementation, where by design, any read intensive workload can not directly benefit from the change. On the other hand, in the case sequential write operations are being executed, AS was capable of taking advantage of the tuning as the response time decreased by 7% and 8% on ext3 and xfs, respectively. The conducted benchmarks revealed another significant inefficiency behavior in the I/O subsystem, as the write performance (for all the schedulers) on ext3 was significantly lower (by a factor of approximately 2.1) than on xfs. The culprit here is the ext3 reservation code. Ext3 patches to resolve the issue are available from kernel.org.

5.2 16-Way RAID-0 Setup

Utilizing the 28 disk RAID-0 configuration as the benchmark environment revealed that across the 4 workload profiles, the deadline implementation was able to outperform the other 3 schedulers (see Appendix C). It has to be pointed out though that the CFQ, as well as the noop scheduler, slightly outperformed the deadline implementation in 3 out of the 4 benchmarks. Overall, the deadline scheduler gained a substantial lead processing the Web server profile (100% random read requests), outperforming the other 3 implementations by up to 62%. On ext3, the noop scheduler reflected the most efficient solution while operating on sequential read and write requests, whereas on xfs, CFQ and deadline dominated the sequential read and write benchmarks. The performance delta among the schedulers (for the 4 profiles) was much more noticeable on xfs (38%) than on ext3 (6%), which reflects a similar behavior as encountered on the RAID-5 setup. Increasing nr_requests to 2,560 on the RAID-0 system led to inconclusive results (for all the I/O schedulers) on ext3 as well as xfs. The erratic behavior encountered in the tuned, large RAID-0 environment is currently being investigated.

5.3 AS Sequential Read Performance

To further illustrate and basically back up the claim made in Section 2 that the AS scheduler

design views the I/O subsystem based on a notion that an I/O device has only one physical (seek) head, this study analyzed the sequential read performance in different hardware setups. The results were being compared to the CFQ scheduler. In the single disk setup, the AS implementation is capable of approaching the capacity of the hardware, and therefore provides optimal throughput performance. Under the same workload conditions, the CFQ scheduler substantially hampers throughput performance, and does not allow the system to fully utilize the capacity of the I/O subsystem. The described behavior holds true for the ext3 as well as the xfs file system. Hence, the statement can be made that in the case of sequential read operations and CFQ, the I/O scheduler (and not the file system per se) reflects the actual I/O bottleneck. This picture is being reversed as the capacity of the I/O subsystem is being increased.

HW Setup	AS	CFQ
1 Disk	52 MB/sec	23 MB/sec
RAID-5	46 MB/sec	39 MB/sec
RAID-0	31 MB/sec	158 MB/sec

Table 1: AS vs. CFQ Sequential Read Performance

As depicted in Table 1, the CFQ scheduler approaches first, the throughput of the AS implementation in the benchmarked RAID-5 environment and second, is capable of approaching the capacity of the hardware in the large RAID-0 setup. In the RAID-0 environment, the AS scheduler only approaches approximately 17% of the hardware capacity (180 MB/sec). To reiterate, the discussed I/O behavior is reflected in the ext3 as well as the xfs benchmark results. From any file system perspective, performance should not degrade if the size of the file system, the number of files stored in the file system, or the size of the individual files stored in the file system increases. Further, the performance of a file system is supposed to approach the capacity of the hardware (workload dependent of course). This study clearly outlines that in the discussed workload scenario, the 2 benchmarked file systems are capable of achieving these goals, but only in the case the I/O schedulers are exchanged depending on the physical hardware setup. The fact that the read-ahead code in Linux 2.6 has to operate as efficiently as possible (in conjunction with the I/O scheduler and the file system) has to be considered here as well.

5.4 AS verses deadline Performance

Based on the benchmarked profiles and hardware setups, the AS scheduler provided in most circumstances the least efficient I/O solution. As the AS framework represents an extension to the deadline implementation, this study explored the possibility of tuning AS to approach deadline behavior. The tuning consisted of setting `nr_requests` to 2,560, `antic_expire` to 0, `read_batch_expire` to 1,000, `read_expire` to 500, `write_batch_expire` to 250, and `write_expire` to 5,000. Setting the `antic_expire` value to 0 (by design) basically disables the anticipatory portion of the scheduler. The benchmarks were executed utilizing the RAID-5 environment, and the results were compared to the deadline performance results reported this study. On ext3, the non-tuned AS version trailed the non-tuned deadline setup by approximately 4.5% (across the 4 profiles). Tuning the AS scheduler resulted into a substantial performance boost, as the benchmark results revealed that the tuned AS implementation outperformed the default deadline setup by approximately 6.5% (see Appendix C). The performance advantage was squandered though while comparing the tuned AS solution against the deadline environment with `nr_requests` set to 2,560. Across

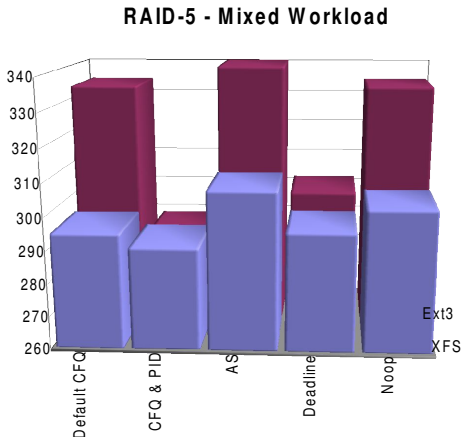
the 4 workload profiles, deadline again outperformed the AS implementation by approximately 17%. As anticipated, setting `antic_expire` to 0 resulted into lower sequential read performance, stabilizing the response time at deadline performance (see Appendix C). On xfs, the results were (based on the rather erratic metadata performance behavior of AS) inconclusive. One of the conclusions is that based on the current implementation of the AS code that collects the statistical data, the implemented heuristic is not flexible enough to detect any prolonged random I/O behavior, a scenario where it would be necessary to deactivate the active wait behavior. Further, setting `antic_expire` to 0 should force the scheduler into deadline behavior, a claim that is not backed up by the empirical data collected for this study. One explanation for the discrepancy is that the short backward seek operations supported in AS are not part of the deadline framework. Therefore, depending on the actual physical disk scheduling policy, the AS backward seek operations may be counterproductive from a performance perspective.

5.5 CFQ Performance

The benchmarks conducted revealed that the tuned CFQ setup provided the most efficient solution for the RAID-5 environment (see Section 5.1). Therefore, the study further explored varies ways to improve the performance of the CFQ framework. The CFQ I/O scheduler in Linux 2.6.5 resembles a SFQ implementation, which operates on a certain number or internal I/O queues and hashes on a per process granularity to determine where to place an I/O request. More specifically, the CFQ scheduler in 2.6.5 hashes on the thread group id (tgid), which represents the process PID as in POSIX.1 [1]. The approach chosen was to alter the CFQ code to hash on the Linux PID. This code change introduces fairness on a per thread (instead of per process) granularity, and therefore alters the distribution of the I/O requests in the internal queues. In addition, the `cfq_quantum` and `cfq_queued` parameters of the CFQ framework were exported into user space.

In a first step, the default tgid based CFQ version with `cfq_quantum` set to 32 (default equals to 8) was compared to the PID based implementation that used the same tuning configuration. Across the 4 profiles, the PID based implementation reflected the more efficient solution, processing the I/O workloads approximately 4.5% and 2% faster on ext3 and xfs, respectively. To further quantify the performance impact of the different hash methods (tgid verses PID based), in a second step, the study compared the default Linux 2.6.5 CFQ setup to the PID based code that was configured with `cfq_quantum` adjusted to 32 (see Appendix C). Across the 4 profiles benchmarked on ext3, the new CFQ scheduler that hashed on a PID granularity outperformed the status quo by approximately 10%. With the new method, the sequential read and write performance improved by 3% and 4%, respectively. On xfs (across the 4 profiles), the tgid based CFQ implementation proved to be the more efficient solution, outperforming the PID based setup by approximately 9%. On the other hand, the PID based solution was slightly more efficient while operating on the sequential read (2%) and write (1%) profiles. The ramification is that based on the conducted benchmarks and file system configurations, certain workload scenarios can be processed more efficiently in a tuned, PID hash based configuration setup.

To further substantiate the potential of the proposed PID based hashing approach, a mixed I/O workload (consisting of 32 concurrent threads) was benchmarked. The environment used reflected the RAID-5 setup. The I/O profile was decomposed in 4 subsets of 8 worker



Note: In Figure 5, the x-axis depicts the I/O schedulers. The front row reflects the xfs, whereas the back row depicts the ext3 based environment. The y-axis discloses the actual response time for the mixed workload profile.

Figure 5: Mixed Workload Behavior

threads, each subset executing either 64KB sequential read, 4KB random read, 4KB random write, or 256KB sequential write operations (see Figure 5). The benchmark results revealed that in this mixed I/O scenario, the PID based CFQ solution (tuned with `cfq_quantum = 32`) outperformed the other I/O schedulers by at least 5% and 2% on ext3 and xfs, respectively (see Figure 5 and Appendix C). The performance delta among the schedulers was greater on ext3 (15%) than on xfs (6%).

6 Conclusions and Future Work

The benchmarks conducted on varying hardware configurations revealed a strong (setup based) correlation among the I/O scheduler, the workload profile, the file system, and ultimately I/O performance. The empirical data disclosed that most tuning efforts resulted in reshuffling the scheduler performance ranking. The ramification is that the choice of an I/O scheduler has to be based on the workload pattern, the hardware setup, as well as the file system used. To reemphasize the importance of the discussed approach, an additional benchmark was conducted utilizing a Linux 2.6 SMP system, the jfs file system, and a large RAID-0 configuration, consisting of 84 RAID-0 systems (5 disks each). The SPECsfs [20] benchmark was used as the workload generator. The focus was on determining the highest throughput achievable in the RAID-0 setup by only substituting the I/O scheduler between SPECsfs runs. The results revealed that the noop scheduler was able to outperform the CFQ, as well as the AS scheduler. The result reverses the order, and basically contradicts the ranking established for the RAID-5 and RAID-0 environments benchmarked in this study. On the smaller RAID systems, the noop scheduler was not able to outperform the CFQ implementation in any random I/O test. In the large RAID-0 environment, the 84 rb-tree data structures that have to be maintained (from a memory as well as a CPU perspective) in CFQ represent a substantial, noticeable overhead factor.

The ramification is that there is no silver bullet (a.k.a. I/O scheduler) that consistently provides the best possible I/O performance. While the AS scheduler excels on small configurations in a sequential read scenario, the non-tuned deadline solution provides acceptable performance on smaller RAID systems. The CFQ scheduler revealed the most potential from a tuning perspective on smaller RAID-5 systems, as increasing the `nr_requests` parameter provided the lowest response time. As the noop scheduler represents a rather light-way solution, large RAID systems that consist of many individual logical devices may benefit from the reduced memory, as well as CPU overhead encountered by this solution. On large RAID systems that consist of many logical devices, the other 3 implementations have to maintain (by design) rather complex data structures as part of the operating framework. Further, the study

revealed that the proposed PID based and tunable CFQ implementation reflects a valuable alternative to the standard CFQ implementation. The empirical data collected on a RAID-5 system supports that claim, as true fairness on a per thread basis is being introduced.

Future work items include analyzing the rather erratic performance behavior encountered by the AS scheduler on xfs while processing a metadata intensive workload profile. Another focal point is an in-depth analysis of the inconsistent `nr_requests` behavior observed on large RAID-0 systems. Different hardware setups will be used to aid this study. The anticipatory heuristics of the AS code used in Linux 2.6.5 is the target of another study, aiming at enhancing the adaptiveness of the (status quo) implementation based on certain workload conditions. Additional research in the area of the proposed PID based CFQ implementation, as well as branching the I/O performance study out into even larger I/O subsystems represent other work items that will be addressed in the near future.

Legal Statement

References

1. The Linux Source Code

2. Arcangeli, A., “Evolution of Linux Towards Clustering,” EFD R&D Clamart, 2003.

3. Axboe, J., “Deadline I/O Scheduler Tunables,” SuSE, EDF R&D, 2003

4. Corbet, J., “A new deadline I/O scheduler.” `http://lwn.net/Articles/10874`.

5. Corbet, J., “Anticipatory I/O scheduling.” `http://lwn.net/Articles/21274`.

6. Corbet, J., “The Continuing Development of I/O Scheduling.” `http://lwn.net/Articles/21274`.

7. Corbet, J., “Porting drivers to the 2.5 kernel,” Linux Symposium, Ottawa, Canada, 2003.

8. Heger, D., Jacobs, J., McCloskey, B., Stultz, J., “Evaluating Systems Performance in the Context of Performance Paths,” IBM Technical White Paper, Austin, 2000.

9. Iyer, S., Drushel, P., “Anticipatory Scheduling – A disk scheduling framework to overcome deceptive idleness in synchronous I/O,” SOSP 2001

10. Lee Irwin III, W., “A 2.5 Page Clustering Implementation,” Linux Symposium, Ottawa, 2003

11. Nagar, S., Franke, H., Choi, J., Seetharaman, C., Kaplan, S., Singhvi, N., Kashyap, V., Kravetz, M., “Class-Based Prioritized Resource Control in Linux,” 2003 Linux Symposium.

12. McKenney, P., “Stochastic Fairness

Queueing," INFOCOM, 1990

13. Molnar, I., "Goals, Design and Implementation of the new ultra-scalable O(1) scheduler." (`sched-design.txt`).

14. Mosberger, D., Eranian, S., "IA-64 Linux Kernel, Design and Implementation," Prentice Hall, NJ, 2002.

15. Shriver, E., Merchant, A., Wilkes, J., "An Analytic Behavior Model with Readahead Caches and Request Reordering," Bell Labs, 1998.

16. Wienand, I., "An analysis of Next Generation Threads on IA64," HP, 2003.

17. Zimmermann, R., Ghandeharizadeh, S., "Continuous Display Using Heterogeneous Disk-Subsystems," ACM Multimedia, 1997.

18. `http://www.iometer.org/`

19. `http://www.specbench.org/osg/mail2001`

20. `http://www.specbench.org/sfs97r1/docs/chapter1.html`

Appendix A: Scheduler Tunables

Deadline Tunables

The `read_expire` parameter (which is specified in milliseconds) is part of the actual deadline equation. As already discussed, the goal of the scheduler is to insure (basically guarantee) a start service time for a given I/O request. As the design focuses manly on read requests, each actual read I/O that enters the scheduler is assigned a deadline factor that consists of the current time plus the `read_expire` value (in milliseconds).

The `fifo_batch` parameter governs the number of request that are being moved to the dispatch queue. In this design, as a read request expires, it becomes necessary to move some I/O requests from the sorted I/O scheduler list into the block device's actual dispatch queue. Hence the `fifo_batch` parameter controls the batch size based on the cost of each I/O request. A request is qualified by the scheduler as either a seek or a stream request. For additional information, please see the discussion on the `seek_cost` as well as the `stream_unit` parameters.

The `seek_cost` parameter quantifies the cost of a seek operation compared to a `stream_unit` (expressed in Kbytes). The `stream_unit` parameter dictates how man Kbytes are used to describe a single stream unit. A stream unit has an associated cost of 1, hence if a request consists of XY Kbytes, the actual cost can be determined as *cost = (XY + stream_unit - 1)/ stream_unit*. To reemphasize, the combination of the `stream_unit`, `seek_cost`, and `fifo_batch` parameters, respectively, determine how many requests are potentially being moved as an I/O request expires.

The `write_starved` parameter (expressed in number of dispatches) indicates how many times the I/O scheduler assigns preference to read over write requests. As already discussed, when the I/O scheduler has to move requests to the dispatch queue, the preference scheme in the design favors read over write requests. However, the write requests can not be staved indefinitely, hence after the read requests were favored for `write_starved` number of times, write requests are being dispatched.

The `front_merges` parameter controls the request merge technique used by the scheduler. In some circumstances, a request may enter the scheduler that is contiguous to a request that is already in the I/O queue. It is feasible to assume that the new request may have a correla-

tion to either the front or the back of the already queued request. Hence, the new request is labeled as either a front or a back merge candidate. Based on the way files are laid out, back merge operations are more common than front merges. For some workloads, it is unnecessary to even consider front merge operations, ergo setting the `front_merges` flag to 0 disables that functionality. It has to be pointed out that despite setting the flag to 0, front merges may still happen due to the cached `merge_last` hint component. But as this feature represents an almost 0 cost factor, this is not considered as an I/O performance issue.

AS Tunables

The parameter `read_expire` governs the timeframe until a read request is labeled as expired. The parameter further controls to a certain extent the interval in-between expired requests are serviced. This approach basically equates to determining the timeslice a single reader request is allowed to use in the general presence of other I/O requests. The approximation `100 * ((seek time / read_expire) + 1)` describes the percentile of streaming read efficiency a physical disk should receive in a environment that consists of multiple concurrent read requests.

The parameter `read_batch_expire` governs the time assigned to a batch (or set) of read requests prior to serving any (potentially) pending write requests. Obviously, a higher value increases the priority allotted to read requests. Setting the value to less than `read_expire` would reverse the scenario, as at this point the write requests would be favored over the read requests. The literature suggests setting the parameter to a multiple of the `read_expire` value. The parameters `write_expire` and `write_batch_expire`, respectively, describe and govern the above-discussed behavior for any (potential) write requests.

The `antic_expire` parameter controls the maximum amount of time the AS scheduler will idle before moving on to another request. The literature suggests initializing the parameter slightly higher for large seek time devices.

Appendix B: Benchmark Environment

The benchmarking was performed in a Linux 2.6.4 environment. For this study, the CFQ I/O scheduler was back-ported from Linux 2.6.5 to 2.6.4.

1.16-way 1.7Ghz Power4+™ IBM p690 SMP system configured with 4GB memory. 28 15,000-RPM SCSI disk drives configured in a single RAID-0 setup that used Emulex LP9802-2G Fiber controllers (1 in use for the actual testing). System was configured with the Linux 2.6.4 operating system.

2.8-way NUMA system. IBM x440 with Pentium™ IV Xeon 2.0GHz processors and 512KB L2 cache subsystem. Configured with 4 qla2300 fiber-cards (only one was used in this study). The I/O subsystem consisted of 2 FAStT700 I/O controllers and utilized 15,000-RPM SCSI 18GB disk drives. The system was configured with 1GB of memory, setup as a RAID-5 (5 disks) configuration, and used the Linux 2.6.4 operating system.

3.Single CPU system. IBM x440 (8-way, only one CPU was used in this study) with Pentium™ IV Xeon 1.5GHz processor, and 512k L2 cache subsystem. The system was configured with a Adaptec aic7899 Ultra160 SCSI adapter and a single 10,000 RPM 18GB disk. The system used the Linux 2.6.4 operating system and was configured with 1GB of memory.

Workload Profiles

1. Web Server Benchmark. The benchmark utilized 4 worker threads per available CPU. In a first phase, the benchmark created several hundred thousand files ranging from 4KB to 64KB. The files were distributed across 100 directories, The goal of the create phase was to exceed the size of the memory subsystem by creating more files than what can be cached by the system in RAM. Each worker thread executed 1,000 random read operations on randomly chosen files. The workload distribution in this benchmark was derived from Intel's Iometer benchmark.

2. File Server Benchmark. The benchmark utilized 4 worker threads per available CPU. In a first phase, the benchmark created several hundred thousand files ranging from 4KB to 64KB. The files were distributed across 100 directories. The goal of the create phase was to exceed the size of the memory subsystem by creating more files than what can be cached by the system in RAM. Each worker thread executed 1,000 random read or write operations on randomly chosen files. The ratio of read to write operations on a per thread basis was specified as 80% to 20%, respectively. The workload distribution in this benchmark was derived from Intel's Iometer benchmark.

3. Mail Server Benchmark. The benchmark utilized 4 worker threads per available CPU. In a first phase, the benchmark created several hundred thousand files ranging from 4KB to 64KB. The files were distributed across 100 directories. The goal of the create phase was to exceed the size of the memory subsystem by creating more files than what can be cached by the system in RAM. Each worker thread executed 1,000 random read, create, or delete operations on randomly chosen files. The ratio of read to create to delete operations on a per thread basis was specified as 40% to 40% to 20%, respectively. The workload distribution in this benchmark was (loosely) derived from the SPECmail2001 benchmark.

4. MetaData Benchmark. The benchmark utilized 4 worker threads per available CPU. In a first phase, the benchmark created several hundred thousand files ranging from 4KB to 64KB. The files were distributed across 100 directories. The goal of the create phase was to exceed the size of the memory subsystem by creating more files than what can be cached by the system in RAM. Each worker thread executed 1,000 random create, write (append), or delete operations on randomly chosen files. The ratio of create to write to delete operations on a per thread basis was specified as 40% to 40% to 20%.

(i) Sequential Read Benchmark. The benchmark utilized 4 worker threads per available CPU. In a first phase, the benchmark created several hundred 50MB files in a single directory structure. The goal of the create phase was to exceed the size of the memory subsystem by creating more files than what can be cached by the system in RAM. Each worker thread executed 64KB sequential read operations, starting at offset 0 reading the entire file up to offset 5GB. This process was repeated on a per worker thread basis 20 times on randomly chosen files.

(ii) Sequential Write (Create) Benchmark. The benchmark utilized 4 worker threads per available CPU. Each worker thread executed 64KB sequential write operations up to a target file size of 50MB. This process was repeated on a per worker-thread basis 20 times on newly created files.

Appendix C: Raw Data Sheets (Mean Response Time in Seconds over 3 Test Runs)

	AS - ext3	DL- ext3	NO - ext3	CFQ - ext3	AS - xfs	DL - xfs	NO - xfs	CFQ - xfs
File Server	610.9	574.6	567.7	579.1	613.5	572.9	571.3	569.9
	AS - ext3	DL- ext3	NO - ext3	CFQ - ext3	AS - xfs	DL - xfs	NO - xfs	CFQ - xfs
MetaData	621	634.1	623.6	597.5	883.8	781.8	773.3	771.7
	AS - ext3	DL- ext3	NO - ext3	CFQ - ext3	AS - xfs	DL - xfs	NO - xfs	CFQ - xfs
Web Server	531.4	502.1	498.3	486.8	559	462.7	461.6	462.9
	AS - ext3	DL- ext3	NO - ext3	CFQ - ext3	AS - xfs	DL - xfs	NO - xfs	CFQ - xfs
Mail Server	508.9	485.3	522.5	505.5	709.3	633	648.5	650.4
	AS - ext3	DL- ext3	NO - ext3	CFQ - ext3	AS - xfs	DL - xfs	NO - xfs	CFQ - xfs
Seq. Read	405	953.2	939.4	945.4	385.2	872.8	881.3	872.4
	AS - ext3	DL- ext3	NO - ext3	CFQ - ext3	AS - xfs	DL - xfs	NO - xfs	CFQ - xfs
Seq. Write	261.3	276.5	269.1	282.6	225.7	222.6	220.9	222.4

Table 2: Single Disk Single CPU – Mean Response Time in Seconds

	AS - ext3	DL- ext3	NO - ext3	CFQ - ext3	AS - xfs	DL - xfs	NO - xfs	CFQ - xfs
File Server	77.2	81.2	86.5	82.7	83.8	90.3	96.6	90.7
	AS - ext3	DL- ext3	NO - ext3	CFQ - ext3	AS - xfs	DL - xfs	NO - xfs	CFQ - xfs
MetaData	147.8	148.4	133	145.3	205.8	90.8	101.6	100.8
	AS - ext3	DL- ext3	NO - ext3	CFQ - ext3	AS - xfs	DL - xfs	NO - xfs	CFQ - xfs
Web Server	70.2	58.4	66.2	59.2	82.1	81.3	78.8	75.2
	AS - ext3	DL- ext3	NO - ext3	CFQ - ext3	AS - xfs	DL - xfs	NO - xfs	CFQ - xfs
Mail Server	119.2	114.8	115.3	119.3	153.9	92.1	100.7	92.2
	AS - ext3	DL- ext3	NO - ext3	CFQ - ext3	AS - xfs	DL - xfs	NO - xfs	CFQ - xfs
Seq. Read	517.5	631.1	654.1	583.5	515.8	624.4	628.7	604.5
	AS - ext3	DL- ext3	NO - ext3	CFQ - ext3	AS - xfs	DL - xfs	NO - xfs	CFQ - xfs
Seq. Write	1033.2	843.7	969.5	840.5	426.6	422.3	462.6	400.4

Table 3: RAID-5 8-Way Setup – Mean Response Time in Seconds

	AS - ext3	DL- ext3	NO - ext3	CFQ - ext3	AS - xfs	DL - xfs	NO - xfs	CFQ - xfs
File Server	78.3	72.1	87.1	70.7	94.1	75	89.2	76
	AS - ext3	DL- ext3	NO - ext3	CFQ - ext3	AS - xfs	DL - xfs	NO - xfs	CFQ - xfs
MetaData	127.1	133	137.3	124.9	189.1	101.1	104.6	99.3
	AS - ext3	DL- ext3	NO - ext3	CFQ - ext3	AS - xfs	DL - xfs	NO - xfs	CFQ - xfs
Web Server	62.4	58.8	75.3	57.5	79.4	72.83	80.6	71.7
	AS - ext3	DL- ext3	NO - ext3	CFQ - ext3	AS - xfs	DL - xfs	NO - xfs	CFQ - xfs
Mail Server	110.2	92.9	118.8	99.6	152.5	100.2	95.1	81
	AS - ext3	DL- ext3	NO - ext3	CFQ - ext3	AS - xfs	DL - xfs	NO - xfs	CFQ - xfs
Seq. Read	523.8	586.2	585.3	618.7	518.5	594.8	580.7	594.4
	AS - ext3	DL- ext3	NO - ext3	CFQ - ext3	AS - xfs	DL - xfs	NO - xfs	CFQ - xfs
Seq. Write	968.2	782.9	1757.8	813.2	394.3	395.6	549.9	436.4

Table 4: RAID-5 8-Way Setup – `nr_requests = 2,560` – Mean Response Time in Seconds

	AS - ext3	DL - ext3	AS Tuned - ext3	AS - xfs	DL - xfs	AS Tuned - xfs
File Server	77.2	81.2	72.1	83.8	90.3	84.5
	AS Default	DL Default	AS Tuned	AS Default	DL Default	AS Tuned
MetaData	147.8	148.4	133.7	205.8	90.8	187.4
	AS Default	DL Default	AS Tuned	AS Default	DL Default	AS Tuned
Web Server	70.2	58.4	62	82.1	81.3	75.9
	AS Default	DL Default	AS Tuned	AS Default	DL Default	AS Tuned
Mail Server	119.2	114.8	103.5	153.9	92.1	140.2
	AS Default	DL Default	AS Tuned	AS Default	DL Default	AS Tuned
Seq. Read	517.5	631.1	634.5	515.8	624.4	614.1
	AS Default	DL Default	AS Tuned	AS Default	DL Default	AS Tuned
Seq. Write	1033.2	843.7	923.4	426.6	422.3	389.1

Table 5: RAID-5 8-Way - Default AS, Default deadline, and Tuned AS Comparison - Mean Response Time in Seconds

	CFQ-ext3	PID-Tuned-ext3	CFQ Tuned-ext3	CFQ-xfs	PID-Tuned-xfs	CFQ Tuned-xfs
File Server	70.7	71.1	70.6	76	75.9	74.3
	CFQ	PID - Tuned	CFQ Tuned	CFQ	PID - Tuned	CFQ Tuned
MetaData	124.9	122	125.1	99.3	92.9	97.4
	CFQ	PID - Tuned	CFQ Tuned	CFQ	PID - Tuned	CFQ Tuned
Web Server	57.5	55.8	58	71.7	73	72.5
	CFQ	PID - Tuned	CFQ Tuned	CFQ	PID - Tuned	CFQ Tuned
Mail Server	99.6	94.5	93.3	81	93.6	93.3
	CFQ	PID - Tuned	CFQ Tuned	CFQ	PID - Tuned	CFQ Tuned
Seq. Read	618.7	599.5	595.4	594.4	583.7	604.1
	CFQ	PID - Tuned	CFQ Tuned	CFQ	PID - Tuned	CFQ Tuned
Seq. Write	813.2	781.1	758.4	436.4	432.1	414.6

Table 6: RAID-5 8-Way- Default CFQ, PID Hashed CFQ & `cfq_quantum=32`, Default CFQ & `cfq_quantum=32` – Mean Response Time in Seconds

	AS - ext3	DL- ext3	NO - ext3	CFQ - ext3	AS - xfs	DL - xfs	NO - xfs	CFQ - xfs
File Server	44.5	40	41.9	40.8	42.5	43	45.9	42.5
	AS - ext3	DL- ext3	NO - ext3	CFQ - ext3	AS - xfs	DL - xfs	NO - xfs	CFQ - xfs
MetaData	66.7	64.6	66.2	64	101.8	71.7	72.4	66.7
	AS - ext3	DL- ext3	NO - ext3	CFQ - ext3	AS - xfs	DL - xfs	NO - xfs	CFQ - xfs
Web Server	43.4	38.2	37.9	42.9	68.3	42.8	69.3	64.5
	AS - ext3	DL- ext3	NO - ext3	CFQ - ext3	AS - xfs	DL - xfs	NO - xfs	CFQ - xfs
Mail Server	60.3	58.5	58.7	58.1	100.3	66.2	65.8	65.1
	AS - ext3	DL- ext3	NO - ext3	CFQ - ext3	AS - xfs	DL - xfs	NO - xfs	CFQ - xfs
Seq. Read	2582.1	470.4	460.2	510.9	2601.2	541	576.1	511.2
	AS - ext3	DL- ext3	NO - ext3	CFQ - ext3	AS - xfs	DL - xfs	NO - xfs	CFQ - xfs
Seq. Write	1313.8	1439.3	1171.1	1433.5	508.5	506.2	508.5	509.8

Table 7: RAID-0 16 – Default I/O Schedulers, No Tuning, Mean Response Time in Seconds

	CFQ	CFQ-T	AS	DL	NO
Mixed ext3	334.1	288.1	371.2	301.2	333.5
	CFQ	CFQ-T	AS	DL	NO
Mixed xfs	295	291	308.4	296	302.8

Table 8: RAID-5 8-Way Mixed Workload Behavior, Mean Response Time in Seconds

Creating Cross-Compile Friendly Software

Sam Robb
TimeSys
sam.robb@timesys.com

Abstract

Typical OSS packages make assumptions about their build environment that are not necessarily true when attempting to cross compile the software. There are two significant contributors to cross compile problems: platform specific code, and build/host confusion. Several examples of problems existing in current OSS packages are presented for each of these root causes, along with explanations of how they can be identified, how they can have been avoided, and how they can be resolved.

1 Why Cross Compile?

Cross compiling is the process of building software on a particular platform (architecture and operating system), with the intent of producing executables that will run on an entirely different platform. Generally, the platform the software is built on is referred to as the "build" system, while the platform the executables are run on is referred to as the "host" system.[1]

The process of cross compiling software is somewhat related to, but distinct from, the process of porting software to run on a different platform. The critical distinction is in the difference between the build and host system characteristics. Often times, software that can be built natively on different platforms will exhibit problems when cross compiling. These problems arise because the software fails to distinguish between the build system and the host system during one or more of the four distinct stages in the process of cross compiling software: configuration, compilation, installation, and verification.

Cross compiling is an absolute necessity for a very small number of software packages. In the OSS world, there are several software packages that are specifically designed with cross compiling in mind (binutils, gcc, busybox, the Linux kernel itself, etc.) These packages are often used to bootstrap a new system, providing a high-quality, low-cost way of obtaining a minimal working system with a small amount of effort. Once a minimal OS and related utilities are present on a system, a developer can then build additional software for the system as required.

As Linux becomes more prevalent in the embedded market space, there is an increased desire among embedded systems developers for more cross compile friendly software packages. While modern embedded systems are often resource rich in terms of processing power, I/O capabilities, memory, and disk space when compared to embedded systems of only a few years ago, compiling software natively on such a system still poses problems for an embedded developer. In extreme cases, compiling a moderately complex software package on an em-

[1] Unfortunately, not everyone chooses the same terminology. For example, the Scratchbox documentation (http://www.scratchbox.org/) uses the terms "host" and "target" where this paper uses "build" and "host" to refer to the same concepts.

bedded system natively may take hours instead of minutes.

Embedded developers therefore prefer cross compiling. Most significantly, it gives the embedded developer the advantage of working in a more comfortable, resource-rich environment—typically on a high-end workstation or desktop system—where they can take advantage of superior hardware to reduce their compile/link/debug cycles. Also importantly, cross compiling makes it easier to set up a system by which an entire system can easily be built from scratch in a reproducible manner.

2 Terminology and Assumptions

Cross compiling is a specialized subset of the software development world, and as such, employs its own terminology in an attempt unambiguously identify certain concepts. The following terms are definitions based on those provided by the GNU autoconf documentation [2], and used commonly in OSS projects such as binutils, gcc, etc.

platform - an architecture and OS combination

build system - the platform that a software package will be *configured* and *compiled* on

host system - the platform that a software package will *run* on

target system - the platform that the software package will *produce* output for

toolchain - the collection of tools (compiler, linker, etc.) along with the headers, libraries, etc. needed to build software for a platform

cross compiler - a **toolchain** that runs on a **host system**, but produces output for a **target system**

Typically, the target system is really only of interest to those working on compilers and related tools, where that extra degree of precision is needed in order to specify the final binary format those tools are intended to produce. In the OSS world, aside from binutils, gcc, and similar software packages, one can usually ignore the additional possibilities and complications introduced by variations in the target system.

The remainder of this paper will assume the existence of a cross compiler[3] that runs on an unspecified build system, and is capable of producing executables that will run on a different unspecified host system. The paper ignores the process of porting software to run on a new platform, in order to concentrate solely on issues that arise from the process of cross compiling the software.

3 Configuration Issues

All but the most simple software packages generally require some means of configuration. This is a process by which the software determines how it should be built—which libraries it should reference, which headers it may include, any particular quirks or workarounds in system calls it needs to deal with, etc.

Configuration is an area ripe for introducing cross compile problems. It provides software packages with the unique opportunity to completely confuse a build by assuming that the build system and the host system are one and the same. All cross compile configuration

[2] Available at `http://www.gnu.org/manual/`

[3] Those interested in building their own cross compiler may wish to consult the 'Resources' section at the end of this paper.

problems are some reflection of this confusion between the identity of the build and host systems.

3.1 Avoid using the wrong tools

This particular problem is caused by misidentifying which tools are to be used as part of the build process. Some software packages expect to be able to build and execute utility programs as part of their build process; a good example of this is the Linux kernel configuration utility. While the final output of the software package will need to run on the host system, these utility programs will need to be run on the build system.

Figure 1 shows an example of this problem. In this case, `CC_FOR_BUILD` is set to the same value as `CC`, which would be appropriate if it wasn't for the fact that earlier in the configuration process, `CC` was explicitly set to reference the cross compiler being used for the build.

```
 # compilers to use to create programs
 # which must be run in the build environment.
-CC_FOR_BUILD = $(CC)
-CXX_FOR_BUILD = $(CXX)
+CC_FOR_BUILD = gcc
+CXX_FOR_BUILD = g++

 SUBDIRS = "this is set via configure, \
            don't edit this"
 OTHERS =
```

Figure 1: Using the wrong tools

In this particular instance, there are several solutions. The most correct, and most expensive, is to update the makefile templates to use the proper variables (`CC_FOR_BUILD` and `CC`) in their proper context. Another possible solution is to override the definition of `CC_FOR_BUILD` and `CC` prior to invoking the makefile. The solution presented in Figure 1 is a simple, straightforward, get-it-working approach where `CC_FOR_BUILD` is simply set to an appropriate value for the majority of build systems.

3.2 Be cautious when executing code on the build system

As part of the configuration process, many software packages—particularly those built on top of `autoconf`—will try to compile, link, or even execute code on the host system.

For `autoconf` based projects, most of the standard `autoconf` macros (`AC_CHECK_LIB`, `AC_CHECK_HEADER`, etc.) do a good job of dealing with cross compile issues. In some instances, though, these standard macros fail when trying to test for the presence of an uncommon header file or library. Developers typically deal with these case by writing custom `autoconf` macros.

If the developer is not cautious, s/he may produce a custom macro that ends up performing a more extensive check than what is really needed. Often times, a developer will create a custom macro that makes use of the `autoconf AC_TRY_RUN` macro. This macro attempts to compile, link, and execute an arbitrary code fragment. The problem here is that the conditions being tested for may not actually require that the resulting binary be executed.

When cross compiling a package that uses custom macros, this leads to a situation where test code will compile and link properly (thanks to the cross compiler), but will then fail to run, or will run and produce incorrect output. In either case, it is highly unlikely that the configure script will reach the proper conclusion about whether or not the header file or library is actually available.

A simple solution to this problem is to check and see if the output from the test program is ever actually used. If not, then the call to `AC_TRY_RUN` in the test macro can be re-

placed with a call to `AC_TRY_COMPILE` or `AC_TRY_LINK`, as shown in Figure 2. These two macros implement checks for the ability to compile and link the provided code fragment, respectively.

```
      SKEY_MSG="yes"

      AC_MSG_CHECKING([for s/key support])
-     AC_TRY_RUN(
+     AC_TRY_LINK(
              [
 #include <stdio.h>
 #include <skey.h>
```

Figure 2: Avoiding execution when linking will suffice

3.3 Allow the user to override a 'detected' configuration value

In some cases, use of `AC_TRY_RUN` is absolutely essential; the automatic configuration process may need to be able to compile, link, and execute code in order to determine the characteristics of the host system. This is a definite stumbling block when trying to configure a software package for cross compiling.

A good configuration script allows the user to explicitly identify or override what would otherwise be an automatically detected value. For `autoconf` based projects, this typically means adding `AC_ARG_ENABLE` macros to your `configure.in` file that allow the user to explicitly set the value of questionable `autoconf` variables.

In the case of existing software packages, there may not be an explicit method for setting a questionable variable. In this case, it may be possible to set the appropriate variable by hand before configuring the software package, in order to force the desired outcome. This may still fail under some circumstances; for example, some configuration scripts do not bother to check to see if the a configuration variable has been set before attempting to automatically deduce its value.

In those cases, the configuration script may be modified[4] to guard the detection code by checking to see if the variable has already been assigned a value. If a value has already been assigned, the configuration script can use the specified value, and skip executing the detection code. In other cases, it may be more appropriate to fix the detection code itself so that it sets the variable to the proper value.

4 Compilation Issues

For the majority of portable software packages, attempting to cross compile will generally not uncover any issues with the code itself.[5] Even though individual source files may compile when pushed through the cross compiler, though, the overall way in which the software is built can still exhibit problems.

4.1 Avoid hard-coded tool names

Figure 4 shows a makefile fragment that originally made an explicit call to `ar`. In a package that is otherwise cross compile friendly, this is a particularly annoying occurrence. Depending on the specifics of the cross compiler, the call to `ar` may succeed, but produce an unusable static library.

Correcting this kind of problem is straightforward—replace the hard-coded tool name with a reference to a make variable

[4] For `autoconf` based software packages, keep in mind that the `configure` script is generated by processing `configure.in`. Editing the `configure` script direclty can be helpful for testing fixes, but changes will have to be made to `configure.in` as well to ensure they persist if the `configure` script is regenerated.

[5] Provided, of course, that the software has already been ported to the host platform.

that names the appropriate tool for the system the binary is intended to run on.

4.2 Avoid decorated tool names

Occaisionally, project makefiles will avoid hardcoded tool names by defining a variable, but then attempt to eliminate the an "unneeded" variable by combining a tool reference with the default flags that should be passed along to the tool, as shown in Figure 3.

While the intent was noble, this type of definition makes it difficult for a user to supply a different definition for a tool. Instead of simply setting the value of of the tool when invoking the makefile (ex, `make AR=ppc7xx-linux-ar`), a user now has to know to define AR in a way that includes the default arguments (ex, `make AR='ppc7xx-linux-ar cr'`).

Again, correcting this type of problem is straightforward—split the definition of the tool reference into a reference to the simple tool name and a variable that indicates the default flags that should be passed to the tool.

```
-AR = @AR@ cq
+AR = @AR@
+ARFLAGS = cq

 all: $(OBJS)
        -rm -f libsupport.a
-       $(AR) libsupport.a $(OBJS)
+       $(AR) $(ARFLAGS) libsupport.a $(OBJS)
        @RANLIB@ libsupport.a
```

Figure 3: Avoiding execution when linking will suffice

4.3 Avoid hard-coded paths

It is very easy for an otherwise cross compile friendly software package to mistakenly set up an absolute include path that looks reasonable. In many situations, the added include path may in fact be harmless, particularly if the build system and host system have roughly the same OS version, library versions, etc. However, even slight differences in structure definitions, enumerated constants, etc. between build system and host system headers can very easily result in either compilation errors, or in the cross compiler producing an unusable binary.

Figures 5 and 6 shows a simple and straightforward solution—remove the hard-coded include path. If the include path is required, then you will need to alter it so that it can be specified relative to the location of the include files appropriate for the host system.

4.4 Avoid assumptions about the build system

While this is nominally a porting issue, sometimes a software package will make what seems to be a reasonable assumption about the build system. In particular, software packages that are intended to run only on a particular class of operating systems (Linux, POSIX complaint systems, etc.) may assume that even if they are cross compiled, they will at least be cross compiled on a build system that has characteristics similar to the host system.

Figure 7 illustrates this problem. This makefile fragment assumes that the build system will have a case-sensitive file system, and that the file patterns `'*.os'` and `'*.oS'` will therefore refer to a distinct set of files—in this case, files for inclusion in a static library and files for inclusion in a shared library, respectively.

This particular assumption breaks down when compiling on a case-insensitive file system like VFAT, NTFS, or HPFS.[6] When encountering this type of problem, there is no easy workaround—the build logic for the software

[6] While these file systems are case-insensitive, they are case preserving, which sometimes helps mask potential case-sensitivity issues.

will need to be altered in order to adjust to the conditions of the unexpected build system.

In this case, the solution was to replace '`*.oS`' with '`*.on`', a file pattern that is distinct from '`*.os`' on either a case-insensitive or a case-sensitive file system.

5 Installation Issues

Software installation is sometimes seen as a simple problem. After all, how hard can it be to just copy files around and make sure they all end up in the right place? As with configuration and compilation, though, cross compiling software introduces additional complexities when installing software.

5.1 Avoid `install -s`

Figure 8 shows a makefile fragment that at first glance looks reasonable; as originally written, it attempted to install a binary using the detected version of the `install` program available on the build system.

The problem here is that the original `install` command specified the `-s` option, which instructs `install` to strip the binary after installing it. Because the command uses the build system's version of `install`, this means that the stripping will be accomplished using the build system's version of `strip`. Depending on the version of `strip` installed on the build system, this command may appear to succeed, yet result in a useless binary being installed.

The solution here is to avoid the use of `install -s`, and instead explicitly strip the binary after installation using the version of `strip` provided with the cross compile toolchain that built the binary.

5.2 Avoid hard-coded installation paths

When cross compiling software, it is often convenient to treat a directory on the build system as the logical root of the host system's file system.[7] This allows a developer to "install" the software into this logical root file system (RFS); often times, the RFS is made available to the host system via NFS.

Autoconf packages typically use variables to specify the prefix for installation paths, which makes installing them into an RFS a simple matter. As Figure 9 shows, non-`autoconf` makefiles may need to be modified to make the same sort of adjustments to installation paths.

Even if the software package already makes use of `prefix` or a similar variable, it may overload the meaning of that variable. This can happen in any type of software package, `autoconf` based or not. For example, a package may use the `prefix` variable to both control the installation path, and also generate `#define` statements that specify paths to configuration files or other important data. In this case, it may still be necessary to modify the makefile to introduce the idea of an installation prefix, as shown in Figure 10.

5.3 Create the required directory structure

Often times, software packages assume that they are being installed on an existing, full-featured system—which implies the existence of a certain directory structure. A cross compiled software package may be installed on the build system into a location that is lacking part or all of a normal directory structure. In this case, the install steps of the software package must be pessimistic, and assume that it will always be necessary to create whatever directory

[7] See the Scratchbox website (`http://www.scratchbox.org`) for more information on the hows and whys of build sandboxing.

structure it requires for the installation to succeed.

Figure 11 shows a patch for a makefile fragment that originally assumed the pre-existence of a particular directory structure. Appropriate calls to `mkdir -p` are enough to ensure that the existing directory structure is in place prior to the install.

6 Verification Issues

There are a number of OSS packages that very conveniently provide self-test capabilities. Along with the usual targets in their makefiles, they include targets that allow the user to build and run a test suite against the software after it is built, but before it is installed.

The main problem here is that these test targets generally run each individual test in the suite using a "compile, execute, analyze" cycle. Even if the compilation and result analysis steps succeed on the build system, test execution will most likely fail if the package has been cross compiled, since the tests were built with the host system in mind. If you are fortunate, these tests will simply fail; otherwise, you will not be able to gauge the accuracy of the tests, as they may be picking up information or artifacts from the build system.

A simple solution is to rewrite test targets to separate test compilation from test execution and result analysis. Providing a distinct install or packaging target for the test suite so that it can be easily moved over to a host system for execution is an added bonus.

Don't assume that you can execute self-tests as part of the normal build cycle (see Figure 12). If you do include a test target as part of your default target dependencies, at least make sure that it is only enabled or run if it knows that it can execute the tests on the build system.

7 Conclusions

By now, it should be apparent that while there are any number of subtle ways that cross compiling software can fail, they are for the most part simple problems with simple solutions.

Developers interested in supporting cross compiling of software packages they maintain can use these problems as a guideline of potential problem areas in their own projects. Detecting potential cross compile issues is often a simple matter of examining project source code and identifying the potential for confusing the meaning of build and host systems.

Finally—the best possible way to examine a software package to see if (or how well) it supports cross compiling is to actually try and cross compile it. While the truly adventurous may wish to try and build their own cross compiler, there are any number of locations on the web where an interested developer can obtain a pre-built toolchain for this purpose. Those working primarily on an x86 Linux host may wish to consider using one of the available pre-built cross compilers that can be found through the rpmfind (`http://www.rpmfind.net`) service. For those interested in building their own cross compiler, or in researching other cross compile issues, are a number of resources (see Table 8) on the net that deal specifically with cross compile issues. The emphasis of these resources is generally on embedded system development, though much of the information available is still applicable when discussing cross compiling in general.

8 Appendix—Code Examples

The following figures are referred to in the paper, and are collected here (instead of presented inline) for the sake of providing clarity in the text. Each figure represents a patch (or a partial patch) for a common OSS package that was used at TimeSys to work around cross compile problems. These selections were chosen to illustrate, in a compact fashion, both the problems described in the text and some possible solutions.

```
       decompress.o \
       bzlib.o

-all: libbz2.a bzip2 bzip2recover test
+all: libbz2.a bzip2 bzip2recover #test

 bzip2: libbz2.so bzip2.c
        $(CC) $(CFLAGS) -o bzip2 $\^
```

Figure 12: Avoid making tests part of the default build target

The CrossGCC Mailing List	`http://sources.redhat.com/ml/crossgcc/` A list for discussing embedded ('cross') programming using the GNU tools.
The CrossGCC FAQ	`http://www.sthoward.com/CrossGCC/`
crosstool	`http://www.kegel.com/crosstool/` A set of scripts to build gcc and glibc for most architectures supported by glibc.
Linux from Scratch	`http://www.linuxfromscratch.org/` A project that provides you with the steps necessary to build your own custom Linux system.
Scratchbox	`http://www.scratchbox.org/` A cross-compile toolkit for embedded Linux application development.
Embedded Gentoo	`http://www.gentoo.org/proj/en/base/embedded/index.xml` Gentoo project concerned with cross compiling and embedded systems.
The GNU configure and build system	`http://www.airs.com/ian/configure/` Document describing the GNU configure and build systems. A bit out of date (circa 1998), but still very useful.
GNU Autoconf, Automake, and Libtool	`http://sources.redhat.com/autobook/` Online version of the classic book covering GNU autotools.

Table 1: Selected internet resources on cross compiling

```
 libbz2.a: $(OBJS)
        rm -f libbz2.a
-       ar cq libbz2.a $(OBJS)
-       @if ( test -f /usr/bin/ranlib -o -f /bin/ranlib -o \
-               -f /usr/ccs/bin/ranlib ) ; then \
-               echo ranlib libbz2.a ; \
-               ranlib libbz2.a ; \
-       fi
+       $(AR) cq libbz2.a $(OBJS)
+       $(RANLIB) libbz2.a
+       #@if ( test -f /usr/bin/ranlib -o -f /bin/ranlib -o \
+       #       -f /usr/ccs/bin/ranlib ) ; then \
+       #       echo ranlib libbz2.a ; \
+       #       ranlib libbz2.a ; \
+       #fi

 libbz2.so: libbz2.so.$(somajor)
```

Figure 4: Avoiding hard-coded tool references

```
 export GCC_WARN   =  -Wall -W -Wstrict-prototypes -Wshadow $(ANAL_WARN)
-export INCDIRS    =  -I/usr/include/ncurses
-export CC         =  gcc
+#export INCDIRS    =  -I/usr/include/ncurses
+#export CC         =  gcc
 export OPT        =  -O2
 export CFLAGS     =  -D_GNU_SOURCE $(OPT) $(GCC_WARN) -I$(shell pwd) $(INCDIRS)
```

Figure 5: Avoiding hard-coded include paths

```
 INSTALL        = install -o $(BIN_OWNER) -g $(BIN_GROUP)

 # Additional libs for Gnu Libc
-ifneq ($(wildcard /usr/lib/libcrypt.a),)
 LCRYPT         = -lcrypt
-endif

 all:           $(PROGS)
```

Figure 6: Avoiding tests for hard-coded path names

```
 # Bounded pointer thunks are only built for *.ob
 elide-bp-thunks = $(addprefix $(bppfx),$(bp-thunks))

-elide-routines.oS += $(filter-out $(static-only-routines),\
+elide-routines.on += $(filter-out $(static-only-routines),\
    $(routines) $(aux) $(sysdep_routines)) \
       $(elide-bp-thunks)
 elide-routines.os += $(static-only-routines) $(elide-bp-thunks)
```

Figure 7: Avoiding assumptions about the build system

```
-       $(INSTALL) -m 0755 -s ssh $(DESTDIR)$(bindir)/ssh
+       $(INSTALL) -m 0755 ssh $(DESTDIR)$(bindir)/ssh
+       $(STRIP) $(DESTDIR)$(bindir)/ssh
```

Figure 8: Replacing install -s with an explicit call to strip

```
 NAME       =  proc

 # INSTALLATION OPTIONS
-TOPDIR     = /usr
+TOPDIR     = $(DESTDIR)/usr
 HDRDIR     = $(TOPDIR)/include/$(NAME)#        where to put .h files
 LIBDIR     = $(TOPDIR)/lib#       where to put library files
-SHLIBDIR   = /lib#                where to put shared library files
+SHLIBDIR   = $(DESTDIR)/lib#      where to put shared library files
 HDROWN     = $(OWNERGROUP) #      owner of header files
 LIBOWN     = $(OWNERGROUP) #      owner of library files
 INSTALL    = install
```

Figure 9: Avoiding hard-coded install paths

```
 # Where is include and dir located?
 prefix=/
+installdir=/

 .c.o:
        $(CC) $(CFLAGS) -c $<
@@ -47,28 +48,32 @@
        -if [ ! -d pic ]; then mkdir pic; fi

 install: lib install-dirs install-data
-       -if [ -f $(prefix)/lib/$(SHARED_LIB) ]; then \
-          mkdir -p $(prefix)/lib/backup; \
-          mv $(prefix)/lib/$(SHARED_LIB) \
-               $(prefix)/lib/backup/$(SHARED_LIB).$$$$; \
+       -if [ -f $(installdir)/$(prefix)/lib/$(SHARED_LIB) ]; then \
+          mkdir -p $(installdir)/$(prefix)/lib/backup; \
+          mv $(installdir)/$(prefix)/lib/$(SHARED_LIB) \
+               $(installdir)/$(prefix)/lib/backup/$(SHARED_LIB).$$$$; \
        fi
-       cp $(SHARED_LIB) $(prefix)/lib
-       chown $(OWNER) $(prefix)/lib/$(SHARED_LIB)
+       cp $(SHARED_LIB) $(installdir)/$(prefix)/lib
+       chown $(OWNER) $(installdir)/$(prefix)/lib/$(SHARED_LIB)
        if [ -x /sbin/ldconfig -o -x /etc/ldconfig ]; then \
          ldconfig; \
```

Figure 10: Working around the use of an overloaded prefix variable

```
 install-only:
   n=‘echo gdbserver | sed ’$(program_transform_name)’‘; \
   if [ x$$n = x ]; then n=gdbserver; else true; fi; \
+  mkdir -p $(bindir); \
+  mkdir -p $(man1dir); \
   $(INSTALL_PROGRAM) gdbserver $(bindir)/$$n; \
   $(INSTALL_DATA) $(srcdir)/gdbserver.1 $(man1dir)/$$n.1
```

Figure 11: Creating required directories at install time

Page-Flip Technology for use within the Linux Networking Stack

John A. Ronciak
Intel Corporation
john.ronciak@intel.com

Jesse Brandeburg
Intel Corporation
jesse.brandeburg@intel.com

Ganesh Venkatesan
Intel Corporation
ganesh.venkatesan@intel.com

Abstract

Today's received network data is copied from kernel-space to user-space once the protocol headers have been processed. What is needed is to provide a *hardware (NIC) to user-space* zero-copy path. This paper discusses a page-flip technique where a page is *flipped* from kernel memory into user-space via page-table manipulation. Gigabit Ethernet was used to produce this zero-copy receive path within the Linux stack which can then be extrapolated to 10 Gigabit Ethernet environments where the need is more critical. Prior experience in the industry with page-flip methodologies is cited.

The performance of the stack and the overall system is presented along with the testing methodology and tools used to generate the performance data. All data was collected using a modified TCP/IP stack in a 2.6.x kernel. The stack modifications are described in detail. Also discussed is what hardware and software features are required to achieve page-flipping.

The issues involving page-flipping are described in detail. Also discussed are problems related to this technology concerning the Virtual Memory Manager (VMM) and processor cache. Another issue that is discussed is what would be needed in an API or code changes to enable user-space applications.

The consequences and possible benefits of this technology are called out within the conclusions of this study. Also described are the possible next steps needed to make this technology viable for general use. As faster networks like 10 Gigabit Ethernet become more commonplace for servers and desktops, understanding and developing zero-copy receive mechanisms within the Linux kernel and networking stack is becoming more critical.

Introduction

Data arriving at a network port undergoes two copy operations (a) from the device memory to kernel memory as a DMA by the device into host memory and (b) from kernel memory to application memory, copied by the processor. Techniques that avoid the second copy are designated zero-copy; no additional copy operations are involved once the data is copied into host memory. Avoiding the second copy can potentially improve throughput and reduce CPU utilization. This has been demonstrated in [Hurd] [Duke] and [Gallatin]. Several techniques have been discussed in the literature for

avoiding the second copy namely page flipping, direct data placement (DDP) and remote DMA (RDMA).

Significant performance benefits were demonstrated with the zero copy implementation in the transmit path. We investigate the effectiveness of the page flipping on newer platforms (faster processor(s) and faster memory). Additional motivation for this experiment and paper came from a discussion on the netdev (and linux-kernel) mailing list where David Miller mentioned his idea of

> On receive side, clever RX buffer flipping tricks are the way to go and require no protocol changes and nothing gross like TOE or weird buffer ownership protocols like RDMA requires.[1]

Approaches

Our initial approach consisted of attempting to modify the 2.4 kernel to support direct modification of PTE's in user and kernel space. This method was based on the assumption that any PTE could represent any location in memory which we later found out not to be true. Our findings indicated that we needed to rely more upon the OS abstraction layers to complete our page-flip implementation. This had the side benefit of making our changes less x86 specific as well. Eventually we settled upon a 2.6 based kernel and effectively implemented our original idea but instead just install a new page into the application space in much the same way as the swapper does. The biggest hurdles came from understanding how the Linux memory manager and its various kernel structures work and relate to each other.

[1] `http://marc.theaimsgroup.com/?l=linux-netdev&w=2&r=1&s=TCP+offloading+interface&q=b`

For our final experiments we used 2.6.4 or newer kernels with what eventually amounted to small changes to the kernel to support page flipped PAGE_SIZE data.

The kernel code consisted of these changes (see patch at the end of this document):

1. Driver modifications to support header and data portions of a packet in separate buffers, where the data buffer is always aligned to a PAGE_SIZE boundary.
2. Add a flag to the skb structure to indicate to the stack that the hardware and driver prepared a zero copy capable receive structure.
3. Modifications to the `skb_copy_datagram_iovec()` function to support calling the new `flip_page_mapping()` function when zero copy capable skbs are received.
4. A new `flip_page_mapping()` function that executes the installation of the driver page into the user's receive data space. This routine handles fixing up permissions.
5. A modification was made to the skb free routines to handle a frags[i] where the .page member was zero after that page had changed ownership to user space.

Experiment

Our test platform consisted of a pre-release system with a dual 2.4 GHz Intel® Pentium® 4 processor supporting Hyper-Threading Technology, and 512 megabytes of RAM. This machine had a network card that supported splitting the header and data portions of a packet into different buffers, and validating the IP, TCP and Ethernet checksums.

Assumptions

For this experiment we made some assumptions to simplify and to work with the hardware that we had available.

- Our application had to allocate a receive data area in multiples of 4K bytes, and that memory had to be PAGE_SIZE aligned.
- We modified the freely available nttcp-1.47 to use valloc instead of malloc, resulting in PAGE_SIZE aligned memory starting addresses.
- Our network used Maximum Transmission Units (MTU) to allow for 4KB or 8KB of data to be packaged in every packet.
- Upon splitting of the packet into header and data portions, this resulted in an aligned data block
- The 2.6.4 kernel was configured for standard 4KB PAGE_SIZE and debugging options were turned off.

Methodologies

After making the required code changes and debugging, we measured the performance of the new "page flip" code against the "copy once" method of receiving data.

These measurements consisted of two major test runs, one where the application never touched the data (notouch) being received, and the other where the application did a comparison of the data to an expected result (touch), effectively forcing the data into the cache and also validating that data was not corrupted in any way through this process.

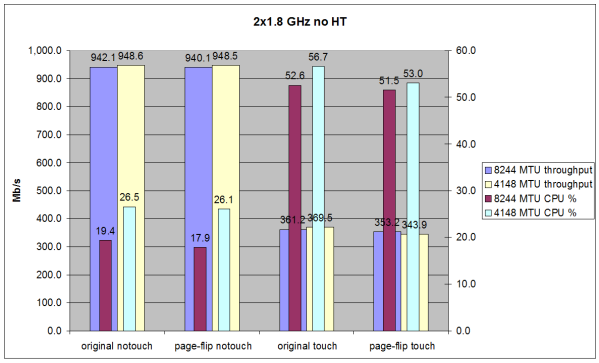


Figure 1: 1.8 GHz comparison

For every instance of the test, three runs were done and the results were averaged for each data point.

Oprofile was used to record the hot-spots for each run.

CPU utilization and network utilization were measured with sar from the sysstat package. NOTE: Our initial results were skewed by a version of sar that incorrectly measured CPU and network utilization (showing more than 1Gb/s transferred in a single direction), be aware that some versions of sar that shipped with your distribution may need to be updated.

Results of Performance Analysis

It is apparent from the touch graphs in Figure 1 that the page flip slightly reduces CPU on slower processors. However, the touch throughput decreases as well, with a decrease in efficiency (Mbits/CPU = eff) for the 4148 MTU from 6.52 (original) to 6.48 (page-flip). The decrease in efficiency is even smaller for 8244 MTUs, where the efficiency went from 6.86 to 6.85. The difference in CPU from the 8244 to the 4148 MTU case is most likely due to header processing as the data throughput is very similar.

The difference between Figure 1 and Figure 2 is simply the processor's speed being adjusted

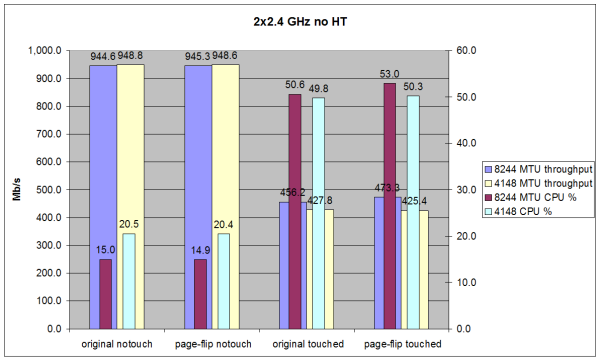


Figure 2: 2.4 GHz comparison

in the bios using a multiplier change. The results from Figure 2 show that the faster processor is more efficient overall, but that even if there is a slight increase in throughput for the page-flip case, the efficiency is still less than if the copy was being done. The efficiency for the 4148 MTU touch data case went from 8.59 to 8.45. For the 8244 byte MTU the efficiency goes from 9.02 to 8.93, even though the throughput goes up.

Surprises and Unexpected Results

We expected that the copy may actually have some beneficial side effects, and our data shows that it does. Especially as processor clock rate increases, the copy becomes less costly in CPU-utilization, while the page flip maintains a constant load which is heavier than the copy was initially.

Oprofile analysis indicated that the locks associated with the page-flip code cause the majority of the stalls in this code path.

Oprofile also showed that the stall associated with the TLB (translation look-aside buffer) flush was very painful.

Conclusions

We had several surprises along the way, but feel confident that at least with our current code base, we can conclude that using a page-flip methodology to receive network data is less efficient than simply doing a copy. The major contributors to this counterintuitive result seem to be cache issues (especially obvious in the "touched data" tests), and a heavier cost associated with the work necessary to prepare and complete the page-flip.

There may be environments such as embedded systems and slower processors where page-flipping will help significantly in decreasing CPU utilization or increasing performance.

Our feeling is that page flipping will not scale in CPU utilization as well as a plain copy does.

There is much room however for optimization of the page-flip code path, which will be followed up with the community. Our expectation is that this optimization will be fighting an uphill battle just to achieve parity with a copy, and then will mostly likely not be able to keep up with speed advances in the processor.

Also, we had to remind ourselves that the cache warming cost must be paid somewhere along all receive paths. Using page-flip methods only moves the cost of the cache miss to the application instead of taking the cost of the miss in the kernel. If the application is waiting impatiently for data, its likely that the cache will be seeded with the data and the application will get all of its data out of cache and have very fast access at that point.

Current issues

The current patch has several outstanding issues that we worked around.

1. There isn't much (if any) commercially available hardware that supports header split receives.
2. Ideally hardware (as mentioned by David Miller) would be able to have flow identification and fill `PAGE_SIZE` buckets with data. This would eliminate the requirement for specific MTU sizes.
3. The current code has a bug when a network data consumer causes a `clone_skb()` to occur. If a page-flipped page pointer `nr_frags[].page` is referenced in the skb being cloned, then a zero pointer is read and the system faults. This is due to the ownership of the page changing from kernel space to user space before the clone is completed. It is not immediately clear if this is an easily surmountable problem, but is easy to work around for our tests.
4. The assumptions we made to enable testing this new code path, like specifying MTU, recompiling the application, etc, create such strict requirements that the usefulness of this code outside of an academic environment is severely limited.

Future directions possible

It is likely that on a system with lots of context switching going on (high load) that the page-flip would be more beneficial. Testing in these environments would provide useful results.

If tested on other architectures besides x86, such as x86-64, IA64 and PPC this code may yield significantly different results.

We did create a driver patch (Appendix B) for the currently available e1000 driver and hardware that prepares packets (using a copy) for processing through the page-flip modified network stack to the user application. We saw that the copy necessary in the driver to do this made the differences between "driver with copy followed by a flip in the stack" and a "driver with a copy followed by another copy in the stack" almost nonexistent. We believe this is because of the cache warming done by the Appendix B driver as it prepares the flip capable structure. Making this code behave more like the flip capable hardware (possibly with a cache flush) would be very useful to increase the amount of experimentation that could be done with the non-hardware specific kernel patches.

References

[Hurd] Dana Hurd Zero-copy interfacing to TCP/IP Dr. Dobbs Journal Sep 1995

[Duke] Trapeze Project: `http://www.cs.duke.edu/ari/trapeze/slides/freenix/sld001.htm`

[Gallatin] Drew Gallatin `http://people.freebsd.org/~ken/zero_copy/`

Appendix A Kernel Patch

This patch will be available at `http://www.aracnet.com/~micro/flip/flip_2_6_4.patch.bz2`

Appendix B mock zero copy e1000 patch

This patch will be available at `http://www.aracnet.com/~micro/flip/e1000_flip.patch.bz2`

Linux Kernel Hotplug CPU Support

Zwane Mwaikambo
FSMLabs
zwane@fsmlabs.com

Ashok Raj
Intel
ashok.raj@intel.com

Rusty Russell
IBM
rusty@rustcorp.com.au

Joel Schopp
IBM
jschopp@austin.ibm.com

Srivatsa Vaddagiri
IBM
vatsa@in.ibm.com

Abstract

During the 2.5 development series, many people collaborated on the infrastructure to add (easy) and remove (hard) CPUs under Linux. This paper will cover the approaches we used, tracing back to the initial PowerPC hack with Anton Blanchard in February 2001, through the multiple rewrites to inclusion in 2.6.5.

After the brief history lesson, we will describe the approach we now use, and then the authors of the various platform-specific code will describe their implementations in detail: Zwane Mwaikambo (i386) Srivatsa Vaddagiri (i386, ppc64), Joel Schopp (ppc64), Ashok Raj (ia64). We expect an audience of kernel programmers and people interested in dynamic cpu configuration in other architectures.

1 The Need for CPU Hotplug

Linux is growing steadily in the mission critical data-center type installations. Such installations requires Reliability, Availability and Serviceability (RAS) features. Modern processor architectures are providing advanced error correction and detection techiniques. CPU hotplug provides a way to realize these features in mission critical applications. CPU hotplug feature adds the following ability to Linux to compete in the high end applications.

- **Dynamic Partitioning**

  Within a single system multiple Linux partitions can be running. As workloads change CPUs can be moved between partitions without rebooting and without interrupting the workloads.

- **Capacity Upgrade on Demand**

  Machines can be purchaced with extra CPUs, without paying for those CPUs until they are needed. Customers can at a later date purchase activiation codes that enable these extra CPUs to match increases in demand, without interrupting service. These activiation codes can either be for temporary activation or permanant activation depending on customer needs.

- **Preventive CPU Isolation**

  Advanced features such as CPU Guard in PPC64 architectures, and Machine Check Abort (MCA) features in Itanium® Product Family (IPF) permit the hardware to catch recoverable failures that are symptomatic of a failing CPU and remove that CPU before an unrecoverable failure occurs. An unused CPU can later be brought online to replace the failed CPU.

2 The Initial Implementation

In February 2001, Anton Blanchard and Rusty Russell spent a weekend modifying the ppc32 kernel to switch CPUs on and off. Stress tests on a 4-way PPC crash box showed it to be reasonably stable. The resulting 60k patch to 2.4.1 was posted to the linux-kernel on February the 4th: `http://www.uwsg.iu.edu/hypermail/linux/kernel/0102.0/0751.html`.

Now we know that the problem could be solved, we got distracted by other things. Upon joining IBM, Rusty had an employer who actually had a use for hotplugging CPUs, and in 2002 the development started up again.

The 2.4 kernels used `cpu_number_map()` to map from the CPU number given by `smp_processor_id()` (between 0 and `NUM_CPUS`) to a unique number between 0 and `smp_num_cpus`. This allows simple iteration between 0 and `smp_num_cpus` to cover all the CPUs, but this cannot be maintained easily in the case where CPU are coming and going. Given my experience that `cpu_number_map()` and `cpu_logical_map()` (which are noops on x86) are a frequent source of errors, Rusty chose to eliminate them, and introduce a `cpu_online()` function which would indicate if the CPU was actually online. Much of the original patch consisted of removing the number remapping, and rewriting loops appropriately.

This change went into Linus' tree in 2.5.24, June 2002, which made the rest of the work much less intrusive.

In the next month, as we were trying to get the `cpu_up()` function used for booting, Linus insisted that we also change the boot order so that we boot as if we were uni-processor, and then bring the CPUs up. Unfortunately, this patch broke Linus' machine, and he partially reverted it, leaving us with the current situation where a little initialization is done before secondary CPUs come online, and normal `__initcall` functions are done with all CPUs enabled. This change also introduced the `cpu_possible()` macro, which can be used to detect whether a CPU could ever be online in the future.

The old boot sequence for architectures was:

1. `smp_boot_cpus()` was called to initialize the CPUs, then
2. `smp_commence()` was called to bring them online.

In addition, each arch optionally implemented a "maxcpus" boot argument. This was made into an arch-independent boot argument, and the boot sequence became:

1. `smp_prepare_cpus(maxcpus)` was called to probe for cpus and set up `cpu_present(cpu)`[1], then

[1] On arch's that dont fill in `cpu_present(cpu)` the function `fixup_cpu_present_map` just uses what `cpu_possible_map` was set during probe. See the section in IA64 for more details.

2. `__cpu_up(cpu)` was called for each CPU where `cpu_present(cpu)` was true, then

3. `smp_cpus_done(maxcpus)` was called after every CPU has been brought up.

At this stage, the CPU notifier chain and the `cpu_up()` function existed, but CPU removal was not in the mainstream kernel. Indeed, significant scheduler changes occurred, preemption went into the kernel, and Rusty was distracted by the module reworking. The result: hotplug CPU development floundered outside the main tree for over a year.

3 The Problem of CPU Removal

The initial CPU removal patch was very simple: the process scheduled on the dying CPU, moved interrupts away, set `cpu_online()` to false, and then scheduled on every other CPU to ensure that noone was looking at the old CPU values. The scheduler's `can_schedule()` macro was changed to return false if the CPU was offline, so the CPU would always run the idle task during this time. Finally, the arch-specific `cpu_die()` function actually killed the CPU.

Three things made this approach harder as the 2.5 kernel developed:

1. Ingo Molnar's O(1) scheduler was included. Rather than checking if the CPU was offline every time we ran `schedule()`, we wanted to avoid touching the highly-optimized code paths.

2. The kernel became preemptible. This means that scheduling on every CPU is not sufficient to ensure that noone is using the old online CPU information.

3. Workqueue and other infrastructure was introduced which used per-cpu threads, which had to be cleanly added and removed.

4. More per-CPU statistics were used in the kernel, which sometimes need to be merged when a CPU went offline (or each sum must be for every possible CPU, not just currently online ones)

5. Sysfs was included, meaning that the interface should be there, instead of in proc, along with structure for other CPU features

Various approaches were discussed and tried: some architectures (like i386) merely simulate CPUs going away, by looping in the idle thread. This is useful for testing. Others (like PPC64 and IA64) actually need to re-start CPUs.

The following were the major design points which were tested and debated, and the resolution of each:

- How should we handle userspace tasks bound to a single CPU?

  Our original code sent a SIGPWR to tasks which were bound such that we couldn't move them to another CPU. This has the default behaviour of killing the task, which is unfortunate if the task merely inherited the binding from its parent. The ideal would be a new signal which would also be delivered on other reconfiguration events (like addition of CPUs, memory), but the Linux ABI does not allow the addition of new signals.

  The final result was to rely on the hotplug scripts to handle this information, and rely on userspace to ensure that removing a CPU was OK before telling the kernel to switch it off.

- How should we handle kernel threads bound to a single CPU?

  Unlike userspace, kernel threads often have a correctness requirement that they run on a particular CPU. Our original approach used a notifier between marking the CPU offline, and actually taking it down; these threads would then shut themselves down. This two-stage approach caused other complications, and the legendary Ingo Molnar recommended a single-stage takedown, and that the kernel threads could be cleaned up later. While that simplified things in general, it involved some new considerations for such kernel threads.

- Issues Creating And Shutting Down Kernel Threads

  In general, the amount of code required to stop kernel threads proved to be significant: barriers and completions at the very least. The other issue is that most kernel threads assume they are started at boot: they don't expect to be started from whatever random process which brought up the CPU.

  This lead Rusty to develop the "kthread" infrastructure, which encapsulated the logic of starting and stopping threads in one place. In particular, it uses keventd (which is always started at boot) to create the new thread, ensuring that there is no contamination by forking the userspace process. The `daemonize()` function attempts to do this, but it's more certain to start from a clean slate than to try to fix a existing one.

- Issues Using keventd for CPU Hotplug

  keventd is used as a general purpose kernel thread for performing some deferred work in a thread context. The "kthread" infrastructure uses this framework to start and stop threads. In addition when various kernel code attempts to call user-space scripts and agents use `call_usermode_helper()`. This function used the keventd thread to spawn the user space program. This approach caused a dead lock situation when the `call_usermode_helper()` is called as part of the `_cpu_disable()`, since keventd threads are per-CPU threads. This results in queueing work to keventd thread via `schedule_work()`, then waiting for completion. This results in blocking the keventd thread. Unless the work queued gets to run, this keventd thread would never be woken again. To avoid this scenario, Rusty introduced the `create_singlethread_workqueue` which now provides a separate thread that is not bound to any particular CPU.

- How to Avoid Having To Lock Around Every Access to Online Map

  Naturally, we wanted to avoid locking around every access to `cpu_online_map` (via `cpu_online()` for example). The method was one Rusty invented for the module code: the so-called "bogolock". To make a change, we schedule a thread on every CPU and have them all simultaneously disabled interrupts, then make the change. This code was generalized from the module code, and called `stop_machine_run()`. This means that we only need to disable preemption to access `cpu_online_map` reliably. If you need to sleep, the `cpu_control` semaphore also protects the CPU hotplug code, so there is a slow-path alternative.

- How to Avoid Doing Too Much Work With the Machine Stopped

  While all CPUs are not taking interrupts,

we don't want to take too long. The initial code walked the task list while the machine was frozen, moving any tasks away from the dying CPU. Nick Piggin came up with an improvement which only migrated the tasks on the CPU's runqueue, and then ensured no other tasks were migrated to the CPU, which reduced the hold time by an order of magnitude. Finally Srivatsa Vaddagiri went one better: by simply raising the priority of the idle task with a special `sched_idle_next()` function, we ensure that nothing else runs on the dying CPU.

The process by which the CPU actually goes offline is as follows:

1. Take `cpu_control` semaphore,
2. Check more than one CPU is online (a bug Anton discovered in the first implementation!),
3. Check that the CPU which they are taking down is actually online,
4. Take the target CPU out of the CPU mask of this process. When the other steps are finished, they will wake us up, and we must not migrate back onto the dead CPU!
5. Use `stop_machine_run()` to freeze the machine and run the following steps on the target CPU
6. Take the CPU out of `cpu_online_map` (easier for arch code to do this first).
7. Call the arch-specific `__cpu_disable()` which must ensure that no more hardware interrupts are received by this CPU (by reprogramming interrupt controllers, or whatever),
8. If that call fails, we restore the `cpu_online_map`. Otherwise we call `sched_idle_next()` to ensure that when we exit the CPU will be idle.
9. At this point, back in the caller, we wait for the CPU to become idle, then call the arch-specific `__cpu_die()` which actually kills the offline CPU, by setting a flag which the idle task polls for, or using an IPI, or some other method.
10. Finally, the `CPU_DEAD` notifier is called, which the scheduler uses to migrate tasks off the dead CPU, the workqueues use to remove the unneeded thread, etc.

The implementation specifics of each architecture can be found in the following sections.

4 Remaining Issues

The main remaining issue is the interaction of the NUMA topology and addition of new CPUs. An architecture can choose a static NUMA topology which covers all the possible CPUs, but for logical partitioning this might not be possible (we might not know in advance).

- Per-CPU variables are allocated using `__alloc_bootmem_node()` at boot, for performance reasons. Unknown CPUs are usually assumed to be in the boot node, which will impact performance.
- sysfs node topology entries need to be updated when a CPU comes online, if the node association is not known at boot.
- The NUMA topology itself should be updated if it is only known when a CPU comes online. This is now possible, using the `stop_machine_run()` function,

but no architectures, other than PPC64, currently do this.

- There are likely some tools in use today that would require minor changes as well. One such tool identified is the top(1) utility, which has trouble dealing with the fact that CPU's available in the system are not logically contiguous. For e.g in a 4-way system, if logical cpu2 was offlined, when cpu0, cpu1, cpu3 were still functional, top would display some error information. Also the tool does not update the CPU information and not able to dynamically update them when new CPU's are added, or removed from the system.

5 i386 Implementation

Commercial i386 hardware available today offer very limited support for CPU Hotplug. Hence the i386 implementation, as it exists, is more of a toy for fun and experimentation. Nevertheless, it was used intensively during development for exercising various code paths and, needless to say, it exposed numerous bugs. Most of these bugs were in arch-independent code.

Since the hardware does not support physical hotplugging of CPUs, only logical removal of a CPU is possible. Once removed from the system, a dead CPU does not participate in any OS activity. Instead, it keeps spinning, waiting for a online command, in the context of its idle thread. Once it gets the online command, it breaks out of the spin loop, puts itself in `cpu_online_map`, flushes TLB and comes alive!

Some important i386 specific issues faced during development are described below:

- **Boot processor**
  There are a few interrupt controller configurations, which necessitate that we not offline the boot processor. Systems may be running with the I/O APIC disabled in which case all interrupts are being serviced by the boot processor via the i8259A, which cannot be programmed to direct interrupts to other processors. Another being interrupts which may be configured to go via the boot processor's LVT (Local Vector Table) such as various timer interrupt setups.

- **smp_call_function**
  smp_call_function is one tricky function which haunted us a long time. Since it deals with sending IPIs to online CPUs and waiting for acknowledgement, number of races was found in this function wrt CPUs coming and going while this function runs on some CPU. Fortunately, when CPU offline was made atomic, most of these race conditions went away. CPU online operation, being still non-atomic, exposes a race wherein an IPI can be sent to a CPU coming online and the sender will not wait for it to acknowledge the IPI. The race was fixed by taking a spinlock (`call_lock`) before putting CPU in the online_map.

- **Interrupt redirection**
  If I/O APIC is enabled, then its redirection table entries (RTEs) need to be reprogrammed every time a CPU comes and goes. This is so that interrupts are delivered to only online CPUs.

  According to Ashok Raj, a safe time to reprogram I/O APIC RTE for any interrupt is when that interrupt is pending, or when the interrupt is masked in RTE.

  Going by the first option, we would have to wait for each interrupt to become pending before reprogramming its RTE. Waiting like this for all interrupts to become

pending may not be a viable solution during CPU Hotplug. Hence the method followed currently is to reprogram RTEs from the dying CPU and wait for a small period (20 microseconds) with interrupts enabled to flush out any pending interrupts. This, in practice, has been enough to avoid lost interrupts.

The right alternative however would be to mask the interrupt in RTE before reprogramming it, but also accounting for the case where the interrupt might have been lost during the interval the entry was left masked. A detailed description of this method is provided in IA64 implementation section.

- **Disabling Local Timer Ticks**
  Local timer ticks are local to each CPU and are not affected by I/O APIC reprogramming. Hence when a CPU is brought down, we have to stop local timer ticks from hitting the dying CPU. This feature is not implemented in the current code. As a consequence, local timer ticks keep hitting and are discarded in software by a `cpu_is_offline` check in its interrupt handler. There are a few solutions under consideration in order to avoid adding a conditional in the timer interrupt path. One method was setting up an offline processor IDT (Interrupt Descriptor Table) which would be loaded when the processor was in the final offline state. The offline IDT would be populated with an entry stub which simply returns from the interrupt. This method would mean that any interrupts hitting the offline processor would be blindly discarded, something which may cause problems if an ACK was required. So what may be safer and sufficient is simply masking the timer LVT for that specific cpu and unmasking it again on the way out of the offline loop.

6 IA64 Implementation

6.1 What is Required to Support CPU Hotplug in IA64?

IA64 CPU hotplug code was developed once Rusty had the base infrastructure support ready. Some of the work that was done to bring the code to stable state include:

- Remove section identifiers marked with `__init` that are required after completing SMP boot. for e.g `cpu_init()`, `do_boot_cpu()` used to wakeup a CPU from SAL_BOOT_RENDEZ mode, `fork_by_hand()` used to fork idle threads for newly added CPUs on the fly.
- Perform a safe interrupt migration from the CPU being removed to another CPU without loss of interrupts.
- Handing off the CPU being removed to SAL_BOOT_RENDEZ mode back to SAL.
- Handling platform level dependencies that trigger physical CPU hotplug in a platform capable of performing it.

6.2 Handling IA64 CPU removal

The arch-specific call `_cpu_disable()` implements the necessary functionality to offline a CPU. The different steps taken are:

1. Check if the platform has any restrictions on this CPU being removed. Returning an error from `_cpu_disable()` ensures that this CPU is still part of the `cpu_online_map`.
2. Turn of local timer interrupt. In IA64 there is a timer interrupt per CPU and not

an external interrupt as in i386 case. It is required that the `timer_interrupt` does not happen any further. It is possible there is one pending, hence check if this interrupt is from an this is an offline CPU, and ignore the interrupt, but just return IRQ_HANDLED, so that the local SAPIC can honour other interrupt vectors now.

3. Ensure that all IRQs bound to this CPU are now targeted to a different CPU by programming the RTEs for a new CPU destination. On return from this step, there must be no more interrupts sent to this CPU being removed from any IOSAPIC.

4. Now the idle thread gets scheduled last, and waits until the CPU state indicates that this CPU must be taken down. Then it hands the CPU to SAL.

6.3 Managing IA64 Interrupts

6.3.1 When Is It Safe to Reprogram an IOSAPIC?

IOSAPIC RTE entries should not be programmed when its actively receiving interrupt signals. The recommended method is to mask the RTE, reprogram for new destination, and then re-enable the RTE. The `/proc/irq` write handlers were calling the set affinity handlers immediately which can cause loss of interrupts, including IOAPIC lockups. In `i386` the introduction of IRQ_BALANCE did this the right way, which is to perform the reprograming operation when an interrupt is pending by storing the intend to change interrupt destinations in a deferred array `pending_irq_balance`.

The same concept was extended to `ia64` as well for the proc write handlers. With the CPU hotplug patches, the write to `/proc/irq` entries are stored in an array and performed when the interrupt is serviced, rather than calling it potentially when an interrupt can also be fired. Due to the delayed nature of these updates, with CPU hotplug, the new destination CPU may be offlined before an interrupt fired and the RTE can be re-programmed. Hence before setting IRQ destination CPU for an RTE, the code should check if the new destination processor is in the `cpu_online_map`.

6.3.2 Why Turn Off Interrupt Redirection Hint With CPU Hotplug?

Interrupt destination in any IOSAPIC RTE must be re-programmed to a different CPU if the CPU being removed is a possible interrupt destination. Since we cannot wait for the interrupt to fire to do the reprogramming, we must force the interrupt destination in safe way. IA64 interrupt architecture permits a platform chipset to perform redirection based on lowest priority based on a hint in the interrupt vector (bit 31) provided by the operating system. If platform interrupt redirection is enabled, it would imply that we need to reprogram all the interrupt destinations, because hotplug code in OS cannot be sure which CPU the chipset is going to direct this interrupt to. Hence if CONFIG_HOTPLUG_CPU is enabled, then we disable platform redirection hint at boot time.

6.3.3 Safely Migrating Interrupt Destinations

The function `fixup_irqs()` performs all the necessary tasks for safely migrating interrupts, and reprogramming interrupt destinations for which this CPU being removed was a destination. The handling of IRQ is managed in 3 distinct phases.

- `migrate_irqs()` performs the job of identifying all IRQs with this CPU as the interrupt destination. This iteration also keeps track of IRQs identified in `vectors_in_migration[]` for later processing to cover cases of missed interrupts, since we mask RTEs during reprogramming, if the device asserted an interrupt during that time, they get lost.

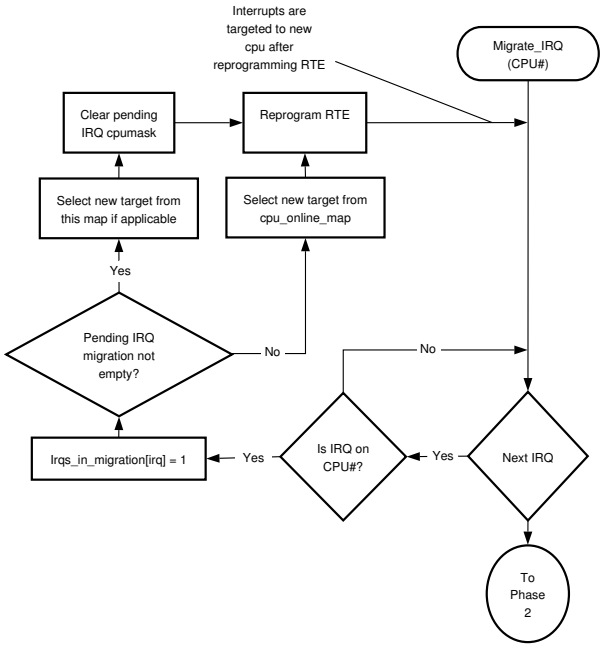


Figure 1: Phase1: Migrate IRQ

- `ia64_process_pending_intr()` Does normal interrupt style processing. During this phase, we look at the local APIC interrupt vector register `ivr` and process all pending interrupts on this CPU. For each processed interrupt, we also clear the bits set in `vectors_in_migration[]`.

- Phase 3 accounts for cases where a device possibly attempted to assert an interrupt, but got lost during the window the RTE was also being re-programmed. This phase looks at entries not accounted for in phase 2, and issues interrupt handler callbacks as if an interrupt happened. It is likely there were no interrupts asserted. We rely on the fact that most device drivers can tolerate calls even if there was no work to perform due to the fact that IRQs may be shared.

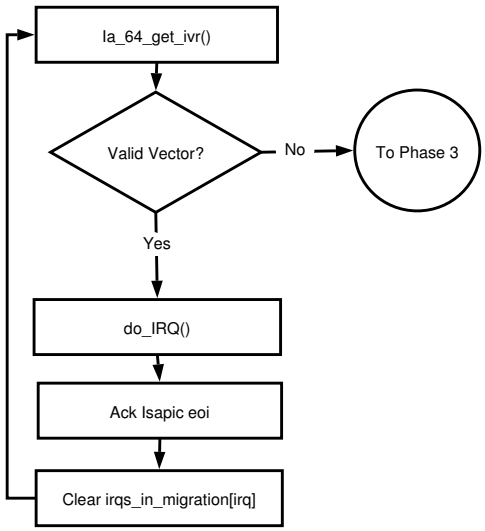


Figure 2: Phase2: Processing Pending intr

6.3.4 Managing Platform Interrupt Sources

IA64 architecture specifies platform interrupt sources to report corrected platform errors to the OS. ACPI specifies these sources via the Platform Interrupt Source Structures. These are communicated to the OS with data such as the following.

- Interrupt Type, indicating if the interrupt is Platform Management Interrupt (PMI), INIT, or CPEI.
- IOSAPIC vector the OS should program.
- The processor that should receive this interrupt, by specifying the APIC id.

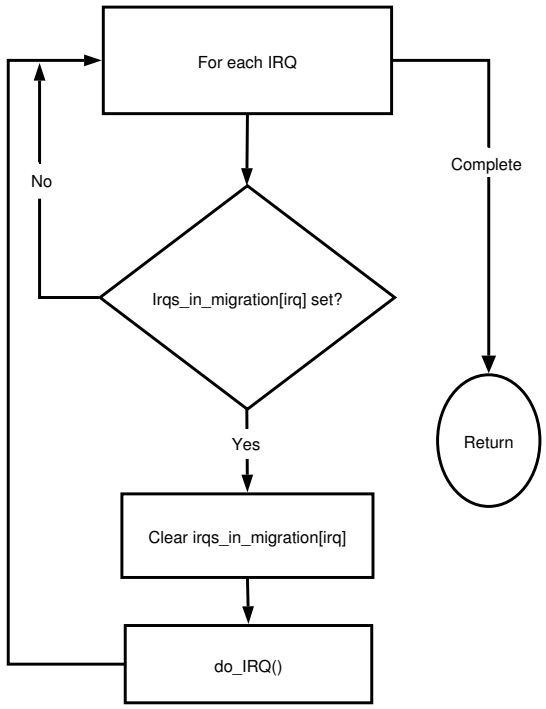


Figure 3: Phase3: Account for Lost Interrupts

- The interrupt line used to signal the interrupts by specifying the global system interrupt.

Some platforms do not support an interrupt model for retrieving platform errors via CPEI. Such platforms provide support via specifying polling tables that list all processors that can poll for Correctable Platform Errors by using the Correctable Platform Error Polling (CPEP) tables.

The issue with both above schemes is that CPEI specifies just one entry for a destination processor. This automatically restricts the target CPU that handles CPEI not removable. On the other hand with CPEP polling tables, although the scheme permits specifying more than one processor, the tables are static and cannot be expanded dynamically as new processors capable of handling polling to be updated.

The motivation for restricting certain processors was that for some platforms that are asymmetric, not all CPUs can retrieve the platform error registers. Hence it is required that only certain processors are permitted. Most platforms that support interruptible model are symmetric in nature. Hence any CPU is capable of accepting the interrupt for CPEI.

We are working with the ACPI specification team to try and address this capability to support platforms supporting CPU hotplug. In the interim before a specification change permits either specifying any CPU as a target, or a method to dynamically update the processors before a CPU gets removed, the code would fail removal of a CPU that is a target of CPEI. In the case of polling, the last processor in the list would be made non-removable.

6.4 Why Should the CPU be handed off to SAL?

The Itanium® processor architecture provides a machine check abort mechanism for reporting and recovering from a variety of errors that can be detected by the processor or chipset. In the event of global MCA, it is required that the slave processors perform checkin with the monarch processor, before which the master could call the recovery to resume execution. SAL would exclude processors in SAL_BOOT_RENDEZ mode. Hence it is important that we return the offlined processors to SAL to avoid processing MCA events on the offlined processor, as the OS would not have it in the active map of CPUs.

6.5 Handling Boot CPU Removal

IA64 architecture does not have any direct dependency that would preclude the boot CPU being removable. There may be some platform level issues such as the boot CPU is usually the target of CPEI or some such dependency that

would make the boot CPU from being removable. In the existing IA64 code base, there is one dependency, that the boot CPU (CPU0) is the master time keeper. This dependency can be easily removed by electing a new CPU as the master timekeeper.

6.6 Recovering the Idle Thread After CPU Removal

Idle threads are created on demand when a new CPU is added to the OS image. These threads are special, since when we return the processor back to SAL, this is done from the context of the idle thread. These calls don't return, and don't have a natural exit path as other threads. The simplest thing to do would be to keep these free idle threads, and just reuse them the next time we need to create a new idle thread for a new CPU.

6.7 Why Was `cpu_present_map` Introduced?

There are several pieces of kernel code that size resources upfront. Before the advent of CPU hotplug, the variable `cpu_possible_map` also indicated the CPUs physically available in the system and would eventually be booted via `smp_init()`. It is very intrusive to make all these callers behave dynamically to CPU hotplug code. There are some issues around this is use of `boot_mem_allocator`. In order to simplify these issues the map `cpu_possible_map` was set to all bits indicating `NR_CPUS`. In order to start only CPUs that are physically present in the system, the new map `cpu_present_map` was added. On platforms capable of supporting CPU hotplug, this map would dynamically change depending on a new CPU being added or removed from the system. In order to accommodate systems that don't directly populate `cpu_present_map` the function `fixup_cpu_present_map` was introduced to just copy the bits from `cpu_possible_map` to `cpu_present_map`.

6.8 ACPI and Platform Issues With CPU hotplug

Any platform capable of supporting hotpluggable CPUs must provide a mechanism to initiate hotplug. Platforms supporting ACPI aware OSs could use ACPI mechanisms to initiate hotplug activity which I would call Physical CPU hotplug. The CONFIG_HOTPLUG_CPU provides the kernel capability and could still be useful if a CPU can be taken offline based on say, the number of correctable error rate.

A typical sequence of operations on a platform supporting a physical CPU is described below. Each specific platform may have additional steps, the following is only a possible sequence and applies to the ACPI based implementations as well.

1. Insert the CPU or the module that contains the CPU into the platform.
2. Platform BIOS does some preparation, and notifies the OS. The kernel platform component such as ACPI that registered to receive the notification, processes this event.
3. Platform dependent OS component prepares necessary information required to bring this CPU to the OS image. For example, in IA64, the code would initialise the following data structures before calling the `cpu_up()`:
   - `ia64_cpu_to_sapicid[]`, in the case of NUMA also populate `node_to_cpu_mask` and `cpu_to_node_map` necessary for NUMA kernels.

- Populate `cpu_present_map` so that kernel now knows about this new CPU is present in the system.

4. Create the necessary entries such as `/sys/devices/system/cpu/cpu#`.

5. Launch the`/sbin/hotplug` script that will now invoke the CPU hotplug agent, which in turn would use the sysfs entry just created to bring up the new CPU.

7 PPC64 Implementation

7.1 What PPC64 Specific Tasks Occur During a CPU Removal?

The architecure specific kernel pieces of a CPU removal focus on three functions mentioned previously: `__cpu_disable()`, `cpu_down()`, and `__cpu_die()`.

In `__cpu_disable()` all interrupts are disabled and migrated, with the exception of interprocessor interrupts (IPIs).

1. The process of disabling interrupts starts off by writing 0 into the processor's current processor priority register (CPPR) to reject any possible queued interrupts.

2. With the CPPR set to 0 it is safe to remove ourselves from the global interrupt queue server, which is done via a Run-Time Abstraction Service (RTAS) set-indicator call that is provided by the firmware. This has the effect of refusing new interrupts from being added to the processor.

3. After new interrupts are refused the next step is to set the CPPR back to default priority, which allows us to recieve IPIs again.

4. All interrupts are iterated through, checking via an RTAS "get-xive" call if any of the interrupts are specific to the target processor.

5. If an interrupt is specific to the target processor it is migrated via an RTAS "set-xive" call.

6. With the processor removed from the global interrupt queue server and all interrupts migrated it would be safe to remove the target processor without affecting the delivery of interrupts. Success is returned.

During `__stopmachine_run()` the online attribute of a CPU is set to to 0. On PPC64 we stop the CPU at this point by calling `cpu_die()` (not to be confused with `__cpu_die()`)

1. Depending on the machine model and kernel configuration, the idle function will be `default_idle()`, `dedicated_idle()`, or `shared_idle()`. All three idle functions check `cpu_is_offline()` and if it is true call `cpu_die()`.

2. `cpu_die` first disables IRQs.

3. After disabling IRQs it clears the CPPR.

4. Finally `rtas_stop_self()` is called, stopping the processor.

Most architectures use `__cpu_die()` to stop the processor. Because on PPC64 we poll for offline CPUs we only need to wait and confirm the CPU has been stopped while in this function.

1. We confirm the CPU has been stopped by using the RTAS `query-cpu-stopped-state` call.

2. Because this call can return busy, and because the CPU may not yet be stopped we loop and schedule timeouts.

3. After confirming the CPU is stopped we do a little extra cleanup by clearing the corresponding entry in the `cpu_callin_map` and `xProcStart` in the PACA.

7.2 What About Adding CPUs?

The initial structure of PPC64 CPU bringup required a lot of modification to be able to add CPUs after the system was already running. Most of the changes are trivial and straightforward, but one bears mentioning.

PPC64 used to number CPUs based on their physical id. With CPU hotplug it would have been necessary to reserve a CPU entry and corresponding structures for each possible physical CPU. It was quite possible that the machine could have more CPUs than the kernel was compiled to work with, as many CPUs would be assigned to other partitions. Furthermore, the number of CPUs in the machine was not necessarily a static number. Also, from a usability point of view there were going to be far too many entries in `/sys/devices/system/cpu/` compared to how many CPUs were actually online.

The CPU numbering was logically abstracted so that for kernel use there was a logical number, and when interfacing to the hardware there was a corresponding physical number. The kernel is able to read at boot time the maximum number of CPUs the partition is configured to be able to grow to. Thus it reserves less space in structures that must be allocated at boot time, allows reuse of logical CPUs for different physical CPUs, and presents a cleaner directory structure.

7.3 Other Software

While outside the scope of this paper it is worth mentioning that there is other software running on PPC64 platforms to enable customers halfway around the world from the machines they administer to use their mouse and move CPUs. This software is downloadable from IBM, and should be available on the bonus CD shipped with new machines.

Issues with Selected Scalability Features of the 2.6 Kernel

Dipankar Sarma
Linux Technology Center
IBM India Software Lab
dipankar@in.ibm.com

Paul E. McKenney
Linux Technology Center and Storage Software Architecture
IBM Beaverton
paulmck@us.ibm.com

Abstract

The 2.6 Linux™ kernel has a number of features that improve performance on high-end SMP and NUMA systems. Finer-grain locking is used in the scheduler, the block I/O layer, hardware and software interrupts, memory management, and the VFS layer. In addition, 2.6 brings new primitives such as RCU and per-cpu data, lock-free algorithms for route cache and directory entry cache as well as scalable user-level APIs like `sys_epoll()` and futexes. With the widespread testing of these features of the 2.6 kernel, a number of new issues have come to light that needs careful analysis. Some of these issues encountered thus far are: overhead of multilple lock acquisitions and atomic operations in critical paths, possibility of denial-of-service attack on subsystems that use RCU-based deferred free algorithms and degradation of realtime response due to increased softirq load.

In this paper, we analyse a select set of these issues, present the results, workaround patches and future courses of action. We also discuss applicability of some these issues in new features being planned for 2.7 kernel.

1 Introduction

Support for symmetric multi-processing (SMP) in the Linux kernel was first introduced in 2.0 kernel. The 2.0 kernel had a single `kernel_flag` lock AKA Big Kernel Lock (BKL) which essentially single threaded almost all of the kernel [Love04a]. The 2.2 kernel saw the introduction of finer-grain locking in several areas including signal handling, interrupts and part of I/O subsystem. This trend continued in 2.4 kernel.

A number of significant changes were introduced in during the development of the 2.6 kernel that helped boost performance of many workloads. Some of the key components of the kernel were changed to have finer-grain locking. For example, the global `runqueue_lock` lock was replaced by the locks on the new per-cpu runqueues. Gone was `io_request_lock` with the introduction of the new scalable `bio`-based block I/O subsystem. BKL was peeled off from ad-

ditional commonly used paths. Use of data locking instead of code locking became more widespread. In addition, Read-Copy Update(RCU) [McK98a, McK01a] allowed further optimization of critical sections by avoiding locking while reading data structures which are updated less often. RCU enabled lock-free lookup of the directory-entry cache and route cache, which provided considerable performance benefits [Linder02a, Blanchard02a, McK02a]. While these improvements targeted high-end SMP and NUMA systems, the vast majority of the Linux-based systems in the computing world are small uniprocessor or low-end SMP systems that remain the main focus of the Linux kernel. Therefore, scalability enhancements must not cause any performance regressions in these smaller systems, and appropriate regression testing is required [Sarma02a]. This effort continues and has since thrown light on interesting issues which we discuss here.

Also, since the release of the 2.6 kernel, its adoption in many different types of systems has called attention to some interesting issues. Section 2 describes the 2.6 kernel's use of fine-grained locking and identifies opportunities in this area for the 2.7 kernel development effort. Section 3 discusses one such important issue that surfaced during Robert Olsson's router DoS testing. Section 4 discusses another issue important for real-time systems or systems that run interactive applications. Section 5 explores the impact of such issues and their workarounds on new experiments planned during the development of 2.7 kernel.

2 Use of Fine-Grain Locking

Since the support for SMP was introduced in the 2.0 Linux kernel, granularity of locking has gradually changed toward finer critical sections. In 2.4 and subsequently 2.6 kernel, many of the global locks were broken up to improve scalability of the kernel. Another scalability improvement was the use of reference counting in protecting kernel objects. This allowed us to avoid long critical sections. While these features benefit large SMP and NUMA systems, on smaller systems, benefit due to reduction of lock contention is minimal. There, the cost of locking due to atomic operations involved needs to be carefully evaluated. Table 1 shows cost of atomic operations on a 700MHz Pentium™ III Xeon™ processor. The cost of atomic increment is more than 4 times the cost of an L2 hit. In this section, we discuss some side effects of such finer-grain locking and possible remedies.

Operation	Cost (ns)
Instruction	0.7
Clock Cycle	1.4
L2 Cache Hit	12.9
Atomic Increment	58.2
cmpxchg Atomic Increment	107.3
Atomic Incr. Cache Transfer	113.2
Main Memory	162.4
CPU-Local Lock	163.7
cmpxchg Blind Cache Transfer	170.4
cmpxchg Cache Transfer and Invalidate	360.9

Table 1: 700 MHz P-III Operation Costs

2.1 Multiple Lock Acquisitions in Hot Path

Since many layers in the kernel use their own locks to protect their data structures, we did a simple instrumentation (Figure 1) to see how many locks we acquire on common paths. This counted locks in all variations of spinlock and rwlock. We used a running counter which we can read using a system call `get_lcount()`. This counts only locks acquired by the task in non-interrupt context.

With this instrumented kernel, we measured writing 4096 byte buffers to a file on ext3 filesystem. Figure 2 shows the test code

```
+static inline void _count_lock(void)
+{
+   if ((preempt_count() & 0x00ffff00) == 0) {
+           current_thread_info()->lcount++;
+   }
+}

....

 #define spin_lock(lock)         \
 do { \
+   _count_lock(); \
    preempt_disable(); \
    _raw_spin_lock(lock); \
 } while(0)
```

Figure 1: Lock Counting Code

```
if (get_lcount(&lcount1) != 0) {
        perror("get_lcount 1 failed\n");
        exit(-1);
}
write(fd, buf, 4096);
if (get_lcount(&lcount2) != 0) {
        perror("get_lcount 2 failed\n");
        exit(-1);
}
```

Figure 2: Lock Counting Test Code

that reads the lock count before and after the `write()` system call.

4K Buffer	Locks Acquired
0	19
1	11
2	10
3	11
4	10
5	10
6	10
7	10
8	16
9	10
Average	11.7

Table 2: Locks acquired during 4K writes

Table 2 shows the number of locks acquired during each 4K write measured on a 2-way Pentinum IV HT system running 2.6.0 kernel. The first write has a lock acquisition count of 19 and an average of 11.7 lock round-trips per 4K write. This does not count locks associated with I/O completion handling which is done from interrupt context. While this indicates scalability of the code, we still need to

```
1 struct file * fget(unsigned int fd)
2 {
3       struct file * file;
4       struct files_struct *files =
5                       current->files;
6
7       read_lock(&files->file_lock);
8       file = fcheck(fd);
9       if (file)
10              get_file(file);
11      read_unlock(&files->file_lock);
12      return file;
13 }
```

Figure 3: fget() Implementation

analyze this to see which locks are acquired in such hot path and check if very small adjacent critical sections can be collapsed into one. The modular nature of some the kernel layers may however make that impossible without affecting readability of code.

2.2 Refcounting in Hot Path

As described in Section 2.1, atomic operations can be costly. In this section, we discuss such an issue that was addressed during the development of the 2.6 kernel. Andrew Morton [Morton03a] pointed out that in 2.5.65-mm4 kernel, CPU cost of writing a large amount of small chunks of data to an ext2 file is quite high on uniprocessor systems and takes nearly twice again as long on SMP. It also showed that a large amount of overheads there were coming from `fget()` and `fput()` routines. A further look at Figure 3 shows how `fget()` was implemented in 2.5.65 kernel.

Both `read_lock()` and `read_unlock()` involve expensive atomic operations. So, even if there is no contention for `->file_lock`, the atomic operations hurt performance [McKenney03a]. Since most programs do not share their file-descriptor tables, the reader-writer lock is usually not really necessary. The lock need only be acquired when the reference count of the `file` structure indicates sharing. We optimized this as shown in Fig-

```
1  struct file *fget_light(unsigned int fd,
2          int *fput_needed)
3  {
4      struct file *file;
5      struct files_struct *files = current->files;
6
7      *fput_needed = 0;
8      if (likely((atomic_read(&files->count)
9                          == 1))) {
10         file = fcheck(fd);
11     } else {
12         read_lock(&files->file_lock);
13         file = fcheck(fd);
14         if (file) {
15             get_file(file);
16             *fput_needed = 1;
17         }
18         read_unlock(&files->file_lock);
19     }
20     return file;
21 }
```

Figure 4: fget_light() Implementation

ure 4.

By optimizing the fast path to avoid atomic operation, we reduced the system time use by 11.2% in a UP kernel while running Andrew Morton's micro-benchmark with the command `dd if=/dev/zero of=foo bs=1 count=1M`. The complete results measured in a 4-CPU 700MHz Pentium III Xeon system with 1MB L2 cache and 512MB RAM is shown in Table 3

Kernel	sys time	Std Dev
2.5.66 UP	2.104	0.028
2.5.66-file UP	1.867	0.023
2.5.66 SMP	2.976	0.019
2.5.66-file SMP	2.719	0.026

Table 3: fget_light() results

However, the reader-writer lock must still be acquired in `fget_light()` fast path when the file descriptor table is shared. This is now being further optimized using RCU to make the file descriptor lookup fast path completely lock-free. Optimizing file descriptor look-up in shared file descriptor table will improve performance of multi-threaded applications that do a lot of I/Os. Techniques such as this are extremely useful for improving performance in both low-end and high-end SMP systems.

```
1  static __inline__ void rt_free(
2          struct rtable *rt)
3  {
4    call_rcu(&rt->u.dst.rcu_head,
5      (void (*)(void *))dst_free,
6      &rt->u.dst);
7  }
8
9  static __inline__ void rt_drop(
10           struct rtable *rt)
11 {
12   ip_rt_put(rt);
13   call_rcu(&rt->u.dst.rcu_head,
14       (void (*)(void *))dst_free,
15       &rt->u.dst);
16 }
```

Figure 5: `dst_free()` Modifications

3 Denial-of-Service Attacks on Deferred Freeing

[McK02a] describes how RCU is used in the IPV4 route cache to void acquiring the per-bucket reader-writer lock during lookup and the corresponding speed-up of route cache lookup. This was included in the 2.5.53 kernel. Later, Robert Olsson subjected a 2.5 kernel based router to DoS stress tests using `pktgen` and discovered problems including starvation of user-space execution and out-of-memory conditions. In this section, we describe our analysis of those problems and potential remedies that were experimented with.

3.1 Potential Out-of-Memory Situation

Starting with the 2.5.53 kernel, the IPv4 route cache uses RCU to enable lock-free lookup of the route hash table.

The code in Figure 5 shows how route cache entries are freed. Because each route cache entry's freeing is deferred by `call_rcu()`, it is not returned to its slab immediately. However

```
CLONE_SKB="clone_skb 1"
PKT_SIZE="pkt_size 60"
COUNT="count 10000000"
IPG="ipg 0"
PGDEV=/proc/net/pktgen/eth0
echo "Configuring $PGDEV"
pgset "$COUNT"
pgset "$CLONE_SKB"
pgset "$PKT_SIZE"
pgset "$IPG"
pgset "flag IPDST_RND"
pgset "dst_min 5.0.0.0"
pgset "dst_max 5.255.255.255"
pgset "flows 32768"
pgset "flowlen 10"
```

Figure 6: pktgen parameters

the route cache imposes a limit of total number of in-flight entries at `ip_rt_max_size`. If this limit is exceeded, subsequent allocation of route cache entries are failed. We reproduced Robert's experiment in a setup where we send 100,000 packets/sec to a 2.4GHz Pentium IV Xeon 2-CPU HT system with 256MB RAM running 2.6.0 kernel set up as a router. Figure 6 shows the parameters used in `pktgen` testing. This script sends 10000000 packets to the router with random destination addresses in the range 5.0.0.0 to 5.255.255.255. The router has an outgoing route set up to sink these packets. This results in a very large number of route cache entries along with pruning of the cache due to aging and garbage collection.

We then instrumented RCU infrastructure to collect lengths of RCU callback batches invoked after grace periods and corresponding grace period lengths. As indicated by the graph plotted based on this instrumentation (Figure 7), it is evident that every spike in RCU batch length as an associated spike in RCU grace period. This indicates that prolonged grace periods are resulting in very large numbers of pending callbacks.

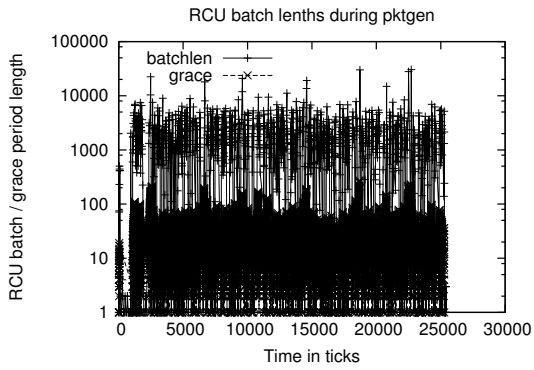


Figure 7: Effect of pktgen testing on RCU

Next we used the same instrumentation to understand what causes long grace periods. We measured total number of softirqs received by each cpu during consecutive periods of 4 `jiffies` (approximately 4 milliseconds) and plotted it along with the corresponding maximum RCU grace period length seen during that period. Figure 8 shows this relationship. It clearly shows that all peaks in RCU grace period had corresponding peaks in number of softirqs received during that period. This conclusively proves that large floods of softirqs holds up progress in RCU. An RCU grace period of 300 milliseconds during a 100,000 packets/sec DoS flood means that we may have up to 30,000 route cache entries pending in RCU subsystem waiting to be freed. This causes us to quickly reach the route cache size limits and overflow.

In order to avoid reaching the route cache entry limits, we needed to reduce the length of RCU grace periods. We then introduced a new mechanism named *rcu-softirq* [Sarma04a] that considers completion of a softirq handler a *quiescent* state. It introduces a new interface `call_rcu_bh()`, which is to be used when the RCU protected data is mostly used from softirq handlers. The update function will be invoked as soon as all CPUs have performed a context switch or been seen in the

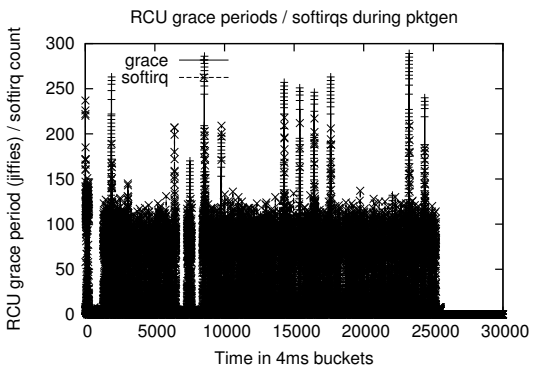


Figure 8: Softirqs during pktgen testing

idle loop or in a user process or or has exited a softirq handler that it may have been executing. The reader side of critical section that use call_rcu_bh() for updating must be protected by `rcu_read_lock_bh()` and `rcu_read_unlock_bh()`. The IPv4 route cache code was then modified to use these interfaces instead. With this in place, we were able to avoid route cache overflows at the rate of 100,000 packets/second. At higher packet rates, route cache overflows have been reported. Further analysis is being done to determine if at higher packet rates, current softirq implementation doesn't allow route cache updates to keep up with new route entries getting created. If this is the case, it may be necessary to limit softirq execution in order to permit user-mode execution to continue even in face of DoS attacks.

3.2 CPU Starvation Due to softirq Load

During the `pktgen` testing, there was another issue that came to light. At high softirq load, user-space programs get starved of CPU. Figure 9 is a simple piece of code that can be used to test this under severe `pktgen` stress. In our test router, it indicated user-space starvation for periods longer that 5 seconds. Application of the *rcu-softirq* patch reduced it by a few seconds. In other words, introduction of quicker

```
gettimeofday(&prev_tv, NULL);

for (;;) {
    gettimeofday(&tv, NULL);
    diff = (tv.tv_sec - prev_tv.tv_sec)*
        1000000 +
        (tv.tv_usec - prev_tv.tv_usec);
        if (diff > 1000000)
            printf("%d\n", diff);
    prev_tv = tv;
}
```

Figure 9: user-space starvation test

RCU grace periods helped by reducing size of pending RCU batches. But the overall softirq rate remained high enough to starve user-space programs.

4 Realtime Response

Linux has been use in realtime and embedded applications for many years. These applications have either directly used Linux for soft realtime use, or have used special environments to provide hard realtime, while running the soft-realtime or non-realtime portions of the application under Linux.

4.1 Hard and Soft Realtime

Realtime applications require latency guarantees. For example, such an application might require that a realtime task start running within one millisecond of its becoming runnable. Andrew Morton's `amlat` program may be used to measure an operating system's ability to meet this requirement. Other applications might require that a realtime task start running within 500 microseconds of an interrupt being asserted.

Soft realtime applications require that these guarantees be met *almost* all the time. For example, a building control application might require that lights be turned on within 250 milliseconds of motion being detected within a

given room. However, if this application occasionally responds only within 500 milliseconds, no harm is likely to be done. Such an application might require that the 250 millisecond deadline be met 99.9% of the time.

In contrast, hard realtime applications require that guarantees *always* be met. Such applications may be found in avionics and other situations where lives are at stake. For example, Stealth aircraft are aerodynamically unstable in all three axes, and require frequent computer-controlled attitude adjustments. If the aircraft fails to receive such adjustments over a period of two seconds, it will spin out of control and crash [Rich94]. These sorts of applications have traditionally run on "bare metal" or on a specialized realtime OS (RTOS).

Therefore, while one can validate a soft-realtime OS by testing it, a hard-realtime OS must be validated by inspection and testing of *all* non-preemptible code paths. Any non-preemptible code path, no matter how obscure, can destroy an OS's hard-realtime capabilities.

4.2 Realtime Design Principles

This section will discuss how preemption, locking, RCU, and system size affect realtime response.

4.2.1 Preemption

In theory, neither hard nor soft realtime require preemption. In fact, the realtime systems that one of the authors (McKenney) worked on in the 1980s were all non-preemptible. However, in practice, preemption can greatly reduce the amount of work required to design and validate a hard realtime system, because while one must validate *all* code paths in a non-preemptible system, one need only validate all *non-preemptible* code paths in a preemptible system.

4.2.2 Locking

The benefits of preemption are diluted by locking, since preemption must be suppressed across any code path that holds a spinlock, even in UP kernels. Since most long-running operations are carried out under the protection of at least one spinlock, the ability of preemption to reduce the Linux kernel's hard realtime response is limited.

That said, the fact that spinlock critical sections degrade realtime response means that the needs of the hard realtime Linux community are aligned with those of the SMP-scalability Linux community.

Traditionally, hard-realtime systems have run on uniprocessor hardware. The advent of hyperthreading and multicore dies have provided cheap SMP, which is likely to start finding its way into realtime and embedded systems. It is therefore reasonable to look at SMP locking's effects on realtime response.

Obviously, a system suffering from heavy lock contention need not apply for the job of a realtime OS. However, if lock contention is sufficiently low, SMP locking need not preclude hard-realtime response. This is shown in Figure 10, where the maximum "train wreck" lock spin time is limited to:

$$S_{max} = (N_{CPU} - 1)C_{max} \quad (1)$$

where N_{CPU} is the number of CPUs on the system and C_{max} is the maximum critical section length for the lock in question. This maximum lock spin time holds as long as each CPU spends at least S_{max} time outside of the critical section.

It is not yet clear whether Linux's lock contention can be reduced sufficiently to make this

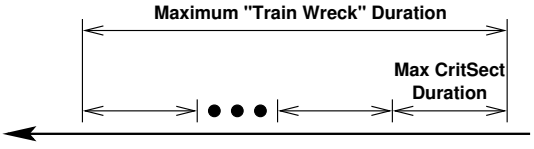


Figure 10: SMP Locking and Realtime Response

level of hard realtime guarantee, however, this is another example of a case where improved realtime response benefits SMP scalability and vice versa.

4.2.3 RCU

Towards the end of 2003, Robert Love and Andrew Morton noted that the Linux 2.6 kernel's RCU implementation could degrade realtime response. This degradation is due to the fact that, when under heavy load, literally thousands of RCU callbacks will be invoked at the end of a grace period, as shown in Figure 11.

The following three approaches can each eliminate this RCU-induced degradation:

1. If the batch of RCU callbacks is too large, hand the excess callbacks to a preemptible per-CPU kernel daemon for invocation. The fact that these daemons are preemptible eliminates the degradation.

2. On uniprocessors, in cases where pointers to RCU-protected elements are not held across calls to functions that remove those elements, directly invoke the RCU callback from within the `call_rcu_rt()` primitive, which is identical to the `call_rcu()` primtive in SMP kernels. The separate `call_rcu_rt()` primitive is necessary because direct invocation is not safe in all cases.

3. Throttling RCU callback invocation so that only a limited number are invoked at a given time, with the remainder being invoked later, after there has been an opportunity for realtime tasks to run.

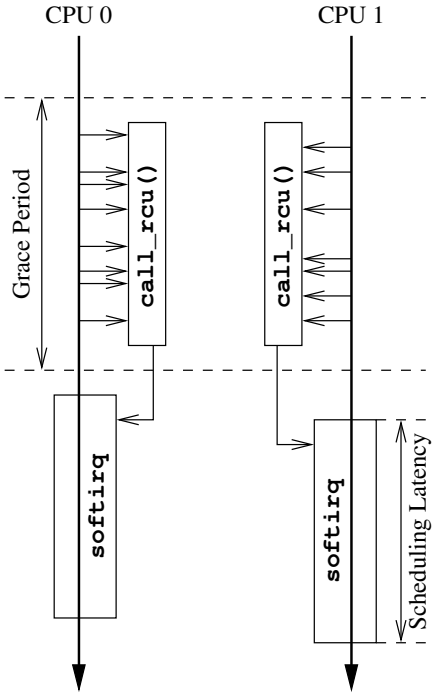


Figure 11: RCU and Realtime Response

The throttling approach seems most attractive currently, but additional testing will be needed after other realtime degradations are resolved. The implementation of each approach and performance results are presented elsewhere [Sarma04b].

4.2.4 System Size

The realtime response of the Linux 2.6 kernel depends on the hardware and software configuration. For example, the current VGA driver degrades realtime response, with multimillisecond scheduling delays due to screen blanking.

In addition, if there are any non-O(1) oper-

ations in the kernel, then increased configuration sizes will result in increased realtime scheduling degradations. For example, in SMP systems, the duration of the worst-case locking "train wreck" increases with the number of CPUs. Once this train-wreck duration exceeds the minimum time between release and later acquisition of the lock in question, the worst-case scheduling delay becomes unbounded. Other examples include the number of tasks and the duration of the tasklist walk resulting from `ls /proc`, the number of processes mapping a given file and the time required to truncate that file, and so on.

In the near term, it seems likely that realtime-scheduling guarantees would only apply to a restricted configuration of the Linux kernel, running a restricted workload.

4.3 Linux Realtime Options

The Linux community can choose from the following options when charting its course through the world of realtime computing:

1. "Just say no" to realtime. It may well be advisable for Linux to limit how much realtime support will be provided, but given recent measurements showing soft-realtime scheduling latencies of a few hundred *micro*seconds, it seems clear that Linux has a bright future in the world of realtime computing.

2. Realtime applications run only on UP kernels. In the past, realtime systems have overwhelmingly been single-CPU systems, it is much easier to provide realtime scheduling guarantees on UP systems. However, the advent of cheap SMP hardware in the form of hyperthreading and multi-CPU cores makes it quite likely that the realtime community will choose to support SMP sooner rather than later. One possibility would be to provide tighter guarantees on UP systems, and, should Linux provide hard realtime support, to provide this support only on UP systems. Another possibility would be to dedicate a single CPU of an SMP system to hard realtime.

3. Realtime applications run only on small hardware configurations with small numbers of tasks, mappings, open files, and so on. This seems to be an eminently reasonable position, especially given that dirt-cheap communications hardware is available, allowing a small system (perhaps on a PCI card) to handle the realtime processing, with a large system doing non-realtime tasks requiring larger configurations.

4. Realtime applications use only those devices whose drivers are set up to provide realtime response. This also seems to be an eminently reasonable restriction, as open-source drivers can be rewritten to offer realtime response, if desired.

5. Realtime applications use only those services able to provide the needed response-time guarantees. For example, an application that needs to respond in 500 microseconds is not going to be doing any disk I/O, since disks cannot respond this quickly. Any data needed by such an application must be obtained from much faster devices or must be preloaded into main memory.

It is not clear that Linux will be able to address each and every realtime requirement, nor is it clear that this would even be desirable. However, it was not all that long ago that common wisdom held that it was not feasible to address both desktop and high-end server requirements with a single kernel source base. Linux is well

on its way to proving this common wisdom to be quite wrong.

It will therefore be quite interesting to see what realtime common wisdom can be overturned in the next few years.

5 Future Plans

With the 2.6 kernel behind us, a number of new scalability issues are currently being investigated. In this section, we outline a few of them and the implications they might have.

5.1 Parallel Directory Entry Cache Updates

In the 2.4 kernel, the directory entry cache was protected by a single global lock `dcache_lock`. In the 2.6 kernel, the look-ups into the cache were made lock-free by using RCU [Linder02a]. We also showed in [Linder02a] that for several benchmarks, only 25% of acquisitions of `dcache_lock` is for updating the cache. This allowed us to achieve significant performance improve by avoiding the lock during look-up while keeping the updates serialized using `dcache_lock`. However recent benchmarking on large SMP systems have shown that `dcache_lock` acquisitions are proving to be costly. Profile for a mutli-user benchmark on a 16-CPU Pentium IV Xeon with HT indicates this:

Function	Profile Counts
`.text.lock.dec_and_lock`	34375.8333
`atomic_dec_and_lock`	1543.3333
`.text.lock.libfs`	800.7429
`.text.lock.dcache`	611.7809
`__down`	138.4956
`__d_lookup`	93.2842
`dcache_readdir`	70.0990
`do_page_fault`	45.0411
`link_path_walk`	9.4866

On further investigation, it is clear that `.text.lock.dec_and_lock` cost is due to frequent `dput()` which uses `atomic_dec_and_test()` to acquire `dcache_lock`. With the multi-user benchmark creating and destroying large number of files in `/proc` filesystem, the cost of corresponding updates to the directory entry cache is hurting us. During the 2.7 kernel development, we need to look at allowing parallel updates to the directory entry cache. We attempted this [Linder02a], but it was far too complex and too late in the 2.5 kernel development effort to permit such a high-risk change.

5.2 Lock-free TCP/UDP Hash Tables

In the Linux kernel, INET family sockets use hash tables to maintain the corresponding `struct socks`. When an incoming packet arrives, this allows efficient lookup of these per-bucket locks. On a large SMP system with tens of thousands on tcp and ip header information. TCP uses `tcp_ehash` for connected sockets, `tcp_listening_hash` for listening sockets and `tcp_bhash` for bound sockets. On a webserver serving large number of simultaneous connections, lookups into `tcp_ehash` table are very frequent. Currently we use a per-bucket reader-writer lock to protect the hash tables and `tcp_ehash` lookups are protected by acquiring the reader side of these per-bucket locks. The hash table makes CPU-CPU collisions on hash chains unlikely and prevents the reader-writer lock from providing any possible performance benefit. Also, on a large SMP system with tens of thousands of simultaneous connection, the cost of atomic operation during `read_lock()` and `read_unlock()` as well as the bouncing of cache line containing the lock becomes a factor. By using RCU to protect the hash tables, the lookups can be done without acquiring the per-bucket lock. This will benefit bot low-end and high-end SMP systems. That said, issues similar to the ones discussed in Section 3 will need to be addressed. RCU can be stressed us-

ing a DoS flood that opens and closes a lot of connections. If the DoS flood prevents user-mode execution, it can also prevent RCU grace periods from happening frequently, in which case, a large number of `sock` structures can be pending in RCU waiting to be free leading to potential out-of-memory situations. The *rcu-softirq* patch discussed in Section 3 will be helpful in this case too.

5.3 Balancing Interrupt and Non-Interrupt Loads

In Section 3.2, we discussed user programs getting starved of CPU time under very high network load. In 2001, Ingo Molnar attempted limiting hardware interrupts based on number of such interrupts serviced during one jiffy [Molnar01a]. Around the same time, Jamal Hadi et al. demonstrated the usefulness of limiting interrupts through *NAPI* infrastructure [Jamal01a]. *NAPI* is now a part of 2.6 kernel and it is supported by a number of network drivers. While *NAPI* limits hardware interrupts, it continues to raise softirqs for processing of incoming packets while polling. So, under high network load, we see user processes starved of CPU. This has been seen with *NAPI* (Robert Olsson's lab) as well as without *NAPI* (in our lab). With extremely high network load like DoS stress, softirqs completely starve user processes. Under such situation, a system administrator may find it difficult to take log into a router and take necessary steps to counter the DoS attack. Another potential problem is that network I/O intensive benchmarks like SPECWeb99™ can have user processes stalled due to high softirq load. We need to look for a new framework that allows us to balance CPU usage between softirqs and process context too. One potential idea being considered is to measure softirq processing time and mitigate it for later if it exceeds its tunable quota. Variations of this need to be evaluated during the development of 2.7 kernel.

5.4 Miscellaneous

1. Lock-free dcache Path Walk: Given a file name, the Linux kernel uses a path walking algorithm to look-up the `dentry` corresponding to each component of the file name and traverse down the `dentry` tree to eventually arrive at the `dentry` of the specified file name. In 2.6 kernel, we implemented a mechanism to look-up each path component in dcache without holding the global `dcache_lock` [Linder02a]. However this requires acquiring a per-`dentry` lock when we have a successful look-up in dcache. The common case of paths starting at the root directory results in contention on the root `dentry` on large SMP systems. Also, the per-`dentry` lock acquisition happens in the fast path (`__d_lookup()`) and avoiding this will likely provide nice performance benefits.

2. Lock-free Tasklist Walk: The system-wide list of tasks in the Linux kernel is protected by a reader-writer lock `tasklist_lock`. There are a number of occasions when the list of tasks need to be traversed while holding the reader side of `tasklist_lock`. In systems with very large number of tasks, the readers traversing the task list can starve out writers. One approach to solving this is to use RCU to allow lock-free walking of the task list under limited circumstances. [McKenney03a] describes one such experiment.

3. Cache Thrashing Measurements and Minimization: As we run Linux on larger SMP and NUMA systems, the effect of cache thrashing becomes more prominent. It would prudent to analyze cache behavior of performance critical code in the Linux kernel using various performance

monitoring tools. Once we identify code showing non-optimal cache behavior, redesigning some of it would help improve performance.

4. Real-time Work—Fix Excessively Long Code Paths: With Linux increasingly becoming preferred OS for many soft-realtime systems, we can further improve its usefulness by identifying excessively long code paths and fixing them.

6 Conclusions

In 2.6 kernel, we have solved a number of scalability problems without significantly sacrificing performance in small systems. A single code base supporting so many different workloads and architectures is an important advantages of the Linux kernel has over many other operating systems. Through this analysis, we have continued the process of evaluating scalability enhancements from many possible angles. This will allow us to run Linux better on many different types of system—large SMP to small TCP/IP routers.

We are continuing to work on some of the core issues discussed in the paper including locking overheads, RCU DoS attack prevention and softirq balancing. We expect to do some of this work in the 2.7 kernel timeframe.

7 Acknowledgments

We owe thanks to Andrew Morton for drawing attention to locking issues. Robert Olsson was the first to show us the impact of denial-of-service attacks on route cache and without his endless testing with our patches, we would have gone nowhere. Ravikiran Thirumalai helped a lot with our lab setup, revived some of the lock-free patches and did some of the measurements. We are thankful to him for this. We would also like to thank a number of Linux kernel hackers, including Dave Miller, Andrea Arcangeli, Andi Kleen and Robert Love, all of whom advised us in many different situations. Martin Bligh constantly egged us on and often complained with large SMP benchmarking results, and for this, he deserves our thanks. We are indebted to Tom Hanrahan, Vijay Sukthankar, Dan Frye and Jai Menon for being so supportive of this effort.

References

[Blanchard02a] A. Blanchard *some RCU dcache and ratcache results*, Linux-Kernel Mailing List, March 2002. `http://marc.theaimsgroup.com/?l=linux-kernel&m=101637107412972&w=2`,

[Jamal01a] Jamal Hadi Salim, R. Olsson and A. Kuznetsov *Beyond Softnet*, Proceedings of USENIX 5th Annual Linux Showcase, pp. 165–172, November 2001.

[Linder02a] H. Linder, D. Sarma, and Maneesh Soni. *Scalability of the Directory Entry Cache*, Ottawa Linux Symposium, June 2002.

[Love04a] Robert Love *Kernel Korner: Kernel Locking Techniques*, *Linux Journal*, Issue 100, August 2002. `http://www.linuxjournal.com/article.php?sid=5833`,

[McK98a] P. E. McKenney and J. D. Slingwine. *Read-copy update: using execution history to solve concurrency problems*, Parallel and Distributed Computing and Systems, October 1998. (revised version available at `http://www.rdrop.com/users/paulmck/rclockpdcsproof.pdf`),

[McK01a] P. E. McKenney and D. Sarma. *Read-Copy Update Mutual Exclusion in Linux*, http://lse.sourceforge.net/locking/rcu/rcupdate_doc.html, February 2001.

[McK01b] P. E. McKenney, J. Appavoo, A. Kleen, O. Krieger, R. Russell, D. Sarma, M. Soni. *Read-Copy Update*, Ottawa Linux Symposium, July 2001. (revised version available at http://www.rdrop.com/users/paulmck/rclock/rclock_OLS.2001.05.01c.sc.pdf),

[McK02a] P. E. McKenney, D. Sarma, A. Arcangeli, A. Kleen, O. Krieger *Read-Copy Update*, Ottawa Linux Symposium, June 2002. http://www.linux.org.uk/~ajh/ols2002_proceedings.pdf.gz

[McKenney03a] MCKENNEY, P.E. Using RCU in the Linux 2.5 kernel. *Linux Journal 1*, 114 (October 2003), 18–26.

[Molnar01a] Ingo Molnar *Subject: [announce] [patch] limiting IRQ load, irq-rewrite-2.4.11-B5*, Linux Kernel Mailing List, Oct 2001. http://www.uwsg.iu.edu/hypermail/linux/kernel/0110.0/0169.html,

[Morton03a] Andrew Morton *Subject: smp overhead, and rwlocks considered harmful*, Linux Kernel Mailing List, March 2003. http://www.uwsg.iu.edu/hypermail/linux/kernel/0303.2/1883.html

[Rich94] B. Rich *Skunk Works*, Back Bay Books, Boston, 1994.

[Sarma02a] D. Sarma *Subject: Some dcache_rcu benchmark numbers*, Linux-Kernel Mailing List, October 2002. http://marc.theaimsgroup.com/?l=linux-kernel&m=103462075416638&w=2,

[Sarma04a] D. Sarma *Subject: Re: route cache DoS testing and softirqs*, Linux-Kernel Mailing List, April 2004. http://marc.theaimsgroup.com/?l=linux-kernel&m=108077057910562&w=2,

[Sarma04b] D. Sarma and P. McKenney *Making RCU Safe for Deep Sub-Millisecond Response Realtime Applications*, USENIX'04 Annual Technical Conference (UseLinux Track), Boston, June 2004.

8 Legal Statement

Achieving CAPP/EAL3+ Security Certification for Linux

Kittur (Doc) S. Shankar
IBM Linux Technology Center
dshankar@us.ibm.com

Olaf Kirch
SUSE Linux AG
okir@suse.de

Emily Ratliff
IBM Linux Technology Center
emilyr@us.ibm.com

Abstract

As far as we know, no Open Source program has been certified for security—until now. Although some people believed that it was not possible for an Open Source program to receive a security certification, we have proven otherwise by obtaining a Common Criteria security certification for SuSE SLES 8 SP3. With the increasing use of Open Source in general and Linux in particular within government and commercial environments, security of Open Source products is of increasing importance and as a result the demand for the security evaluation of Linux is evident. It is also generally believed that security certifications are time consuming and can take years to accomplish. We were able to obtain the Common Criteria certification of Linux in a few months. The presentation will cover our experience and the technical challenges associated with this Linux evaluation. In particular, we will discuss the enhancements we made to SLES 8 SP3 including the Linux kernel to support CAPP audit requirements. In addition the business advantages of the evaluation for Open Source software will be covered.

1 Introduction

In promoting Linux to IBM's enterprise and government customers, the requirement for Common Criteria certification emerged as a barrier to entry. All of Linux's commercial competitors have the required level of certification. As Linux continues to be adopted by the enterprise market, many customers, especially those from the government sector, have raised concerns regarding Linux security and questioned whether Linux was capable of achieving certification. These customers view security certification as table stakes for proving a minimal level of operating system security. In order to increase Linux adoption by these customers, certification is required. The expense of achieving certification makes certification unobtainable by community projects without corporate or government sponsorship. For these reasons, and after a careful analysis, IBM decided to sponsor a Common Criteria (CC) security certification for Linux. SUSE agreed to partner with IBM to evaluate SUSE LINUX Enterprise Server 8 (SLES 8).

In this paper, we will begin with a brief overview of the Common Criteria standard. We will then describe our approach and expe-

rience during this certification effort. We will describe in detail the additional functionality that was needed in kernel and user space to fulfill the requirements of the certification. We will also describe the level of documentation and test needed to obtain the certification.

Throughout the paper, we use the pronoun 'we.' By 'we,' we mean individuals or subteams from the large team of people who contributed time and effort in achieving this evaluation, including:

- IBM, the evaluation sponsor
- SUSE, the developer
- atsec information security GmbH, the evaluator
- BSI, the German agency for information security, the evaluation body

2 Common Criteria Overview

Common Criteria (CC) is documented in the ISO standard 15408 for the security analysis of IT products. The governments of 18 nations have officially adopted the Common Criteria, including the United States, Canada, Germany, France, and the UK. The U.S. government has required a Common Criteria evaluation for all IT-products used for the processing of security-critical data since July 1, 2002.[1]

Common Criteria splits the requirements into two sets: functional and assurance. Functional requirements describe the security attributes of the product under evaluation. Assurance requirements describe the activities that must take place to increase the evaluator's confidence that the security attributes are present, effective, and are designed and implemented correctly. Examples of assurance activities include documentation of the developer's search for vulnerabilities and testing.

[1]The requirement is codified by NSTISSP No. 11.

2.1 Functional Requirements

The functional requirements desired by the customer are described in the Protection Profile (PP). Protection Profiles are targeted at specific types of systems. For example, there are unique protection profiles for operating systems, firewalls, databases and other complex or security sensitive products. Protection Profiles are often created by the product developer, standards bodies, or government agencies, rather than by the customer. To be officially recognized, the Protection Profile must itself be evaluated. Protection Profiles are intended to be reusable and thus typically define standard sets of security attributes that can be used to compare different implementations of a product type. The name of the Protection Profile is therefore often used as shorthand to describe the functional level of the evaluation.

The product being evaluated is known as the Target of Evaluation (TOE). The security policy used by the TOE is known as the TOE Security Policy (TSP) and the functionality that enforces the TSP is known as the TOE Security Functions (TSF). The TSP may be enforced by software, hardware or firmware, but no matter what the enforcement mechanism is, the enforcement functionality is included in the TSF. The TOE does not exist in a vacuum; external forces that act on the TOE are known as the TOE (security) environment. The TOE environment may consist of elements such as non-privileged processes running in an operating system and the network to which a system is attached. The main purpose of an evaluation is to determine whether or not the TSP is correctly enforced.

2.2 Assurance Requirements

Evaluated Assurance Levels (EALs) are defined on a scale of increasingly rigorous development methodologies. The Common Criteria defines multiple classes of assurance components with multiple levels of difficulty for each component. The assurance levels are then composed from these components. These components include items such as level of documentation and testing. The assurance components used for this evaluation are described in more detail in the EAL3 Overview Section. Each higher assurance level requires more proof that security was a fundamental element of the development process; therefore, each higher level is more difficult to achieve than the previous level. There are seven ordered EALs[2]:

- EAL1 – Functionally Tested
- EAL2 – Structurally Tested
- EAL3 – Methodically tested and checked
- EAL4 – Methodically designed, tested and reviewed
- EAL5 – Semiformally designed and tested
- EAL6 – Semiformally verified, designed and tested
- EAL7 – Formally verified, designed and tested

EAL1 is the entry level assurance level. EAL4 is the highest assurance level that any product is expected to be able to achieve without significant expense and rework if it had not been specifically developed with Common Criteria evaluation in mind.

[2]Common Criteria Part 3 available from `http://csrc.nist.gov/cc/Documents/CC%20v2.1%20-%20HTML/CCCOVER.HTM`

2.3 Evaluation Approach

When the developer has decided on a Target of Evaluation and a Protection Profile, the first step towards evaluation is writing a Security Target (ST) which describes the security objectives of the TOE and how they meet the security requirements defined in the chosen PP. It is possible for an ST to claim conformance to multiple PPs or no PP at all. The claims in the security target determine the scope of the evaluation. Every facet of the evaluation is directly impacted by what is claimed in the Security Target. After the evaluation is completed, the Security Target is always made available for customer scrutiny so that the customer can understand exactly what was evaluated.

3 Description of the Evaluated TOE

Our target of evaluation (TOE) was the SUSE LINUX Enterprise Server 8 operating system with Service Pack 3 and the certification-sles-eal3.rpm package.

The SLES evaluation covers a distributed, but isolated, network of IBM® xSeries®, pSeries®, iSeries®, and zSeries® servers running the evaluated version of SLES. The hardware platforms selected for the evaluation consisted of commercially available machines from across the IBM product line.

The TOE Security Functions (TSF) consist of Linux kernel functions plus some trusted processes. These functions enforce the security policy as defined in the Security Target. The TOE includes standard networking applications, such as ftp, ssh, ssl, and xinetd. System administration tools include standard admin commands. Yast2 and several yast2 modules were also included in the package list that formed the TOE. The X Window System was

not included in the evaluated configuration.

The hardware and the system firmware are not considered to be part of the TOE but rather are a part of the TOE environment. The TOE environment also includes applications that are not evaluated, but are used as unprivileged tools to access public system services. For example, an HTTP server using a port above 1024 (e.g., on port 8080) can be used as a normal application running without root privileges on top of the TOE. The Security Guide provides guidance on how to set up a http server on the TOE without violating the evaluated configuration.

4 ST Description

The Security Target specifies that the evaluation covers the Controlled Access Protection Profile (CAPP) functionality at the EAL3 augmented assurance level.[3] The primary security features and assurance documentation are described below, along with how the requirements were satisfied. The key features are supported by domain separation and reference mediation, which ensure that the features are always invoked and cannot be bypassed. Most of the security and assurance features are included in the vanilla kernel (e.g., object reuse) or are standard to most Linux distributions (e.g., PAM, OpenSSH, OpenSSL) and were thus already present in SLES 8. A few, most notably audit, had to be added for the evaluation.

4.1 EAL3 Overview

EAL3 provides assurance by an analysis of the security functions, using its functional and interface specifications, guidance documentation, and the high-level design of the TOE to understand the security behavior. The EAL3 assurance requirements fall into the following seven categories:

- Configuration Management
- Delivery and Operations
- Development
- Guidance Documents
- Life Cycle Support
- Security Testing
- Vulnerability Assessment.

Many of the documents created to support the assurance requirements can be reviewed at `http://oss.software.ibm.com/linux/pubs/?topic_id=5`

4.1.1 Configuration Management

The Configuration Management assurance class specifies the means for establishing that the integrity of the TOE is preserved during development. The Configuration Management process must provide a mechanism for tracking changes and ensuring that all the changes are authorized.

Configuration management procedures within SUSE are highly automated using a process supported by the AutoBuild tool. Source code, generated binaries, documentation, test plan, test cases and test results are maintained under configuration management. Because of this, SUSE already exceeded the requirements for this evaluation, so we just had to document existing procedures to fulfill this requirement.

This assurance requirement is the one that was commonly expected to be the source of difficulty in achieving certification of code developed via the open source methodology. The

[3] The augmentation is the flaw remediation procedure.

key to meeting this assurance requirement is that every line of new code that comes into the SUSE AutoBuild environment is assigned to an owner within SUSE who becomes responsible for its integrity.

4.1.2 Delivery and Operations

The Delivery and Operations class provides requirements for the assurance that the TOE is not corrupted between the time the developer releases it and the customer fires it up.

SLES is delivered on CD/DVD in shrink-wrapped package to the customer. SUSE verifies the integrity of the production CDs and DVDs by checking a production sample. Service Pack 3, the certification-sles-eal3.rpm package, as well as other packages that contain fixes must be downloaded from the SUSE maintenance Web site. Because those packages are digitally signed, the user is both able to and required to verify the integrity and authenticity of those packages. Guidance for installation and system configuration is provided in the Security Guide.

Again, existing SUSE processes met the requirements for the EAL3 assurance level, so documenting existing procedures was sufficient for this evaluation.

4.1.3 Development

The Development class encompasses requirements for documenting the TSF at various levels of abstraction, from the functional interface to the implementation representation. For EAL3, we needed a functional specification and a high-level design. In addition, the correspondence between the security functionality, the functional specification, and the high level design had to be documented.

The functional specification for SLES consists of the man pages that describe the system calls, the trusted commands, and a description of the security-relevant configuration files. A spreadsheet tracks all system calls, trusted commands, and security-relevant configuration files with a mapping (correspondence) to their description in the high-level design and man page(s). The high-level design of the security functions of SLES provides an overview of the implementation of the security functions within the subsystems of SLES, and points to other existing documents for further details where appropriate.

To fulfill this requirement, the functional specification spreadsheet, correspondence, and high-level design were written. Additionally, several new man pages were created for undocumented system calls, PAM modules and utilities, and many man pages required minor corrections.

4.1.4 Guidance Documents

The Guidance Documents class provides the requirements for user and administrator guidance documentation. A security guide is also necessary to fulfill the requirements of this class at EAL3.

SLES 8 already shipped with User and Administrator Guides. The Security Guide and a special README file were created that contain the specifics for the secure administration and usage of the evaluated configuration. The Security Guide explicitly documents setting up and maintaining the system in an evaluated configuration.

4.1.5 Life Cycle Support

The Life Cycle Support assurance requirement includes requirements for processes that deal with vulnerabilities found after release of the product, as well as the physical security of the developer's lab.

The SUSE security procedures are defined and described in documents in the SUSE intranet. The defect handling procedure SUSE has in place for the development of SLES requires the description of the defect with its effects, security implications, fixes and required verification steps.

Again, existing (and previously planned updates to) SUSE procedures met the requirements for this class and were merely required to be documented to fulfill the assurance requirements.

4.1.6 Security Testing

The emphasis of the Security Testing class is on the confirmation that the TSF operates according to its specification. This testing provides assurance that the TOE satisfies the security functionality requirements. Coverage (completeness) and depth (level of detail) are separated for flexibility.

A detailed test plan was produced to test the functions of SLES on each evaluated platform. The test plan includes an analysis of the test coverage, an analysis of the functional interfaces tested, and an analysis of the testing against the high level design. Test coverage of internal interfaces was defined and described in the test plan documents and the test case descriptions. The tests were executed on every platform. The test results are documented so that the tests can be repeated and the results independently confirmed.

Although, SUSE has an excellent test infrastructure for regression testing already in place, additional tests were required to test new functionality, such as audit, and ensure coverage of security relevant events. The Linux Test Project provided an excellent base for the test suite needed for EAL3. It already contained almost all of the necessary test cases for every system call. In some cases, we had to add tests of expected failure cases to ensure that the security was being correctly enforced. We added some test cases for security-relevant programs, such as su, cron, at, and ssh. We created tests to ensure that the system was configured in the evaluated manner. We also created many tests for correct ACL behavior. Many of the system call and security-relevant program test cases were created during the course of the EAL2 evaluation and then reused during the EAL3 evaluation. The largest class of new test cases for EAL3 was tests of the new audit system. Testing the audit subsystem required showing that all security-relevant system calls are logged correctly, all trusted programs (including PAM) correctly logged security-relevant events, the audit userspace tools contained correct functionality, and that audit exhibits Controlled Access Protection File (CAPP)-compliant behavior during threshold and failure events (for example, low disk space). Gcov was used to show test coverage of the kernel internal interfaces. Writing, documenting, and running these test cases on all of the evaluated platforms was a significant portion of the evaluation effort.

4.1.7 Vulnerability Assessment

The Vulnerability Assessment class defines requirements for evidence that the developer looked for vulnerabilities that might arise during development and use of the TOE.

Our search for vulnerabilities was documented

in the Vulnerability Assessment document. This assessment included TOE misuse analysis and a password strength of function analysis. The analysis also describes the approach used to identify vulnerabilities of SLES and the results of the findings.

The Vulnerability Assessment was performed and written as part of this evaluation.

4.2 CAPP Overview

The Controlled Access Protection Profile (CAPP) is based on the C2 class of the "Department of Defense Trusted Computer Systems Evaluation Criteria" (DoD 5200.28 – STD) colloquially known as the "Orange Book." CAPP requires that the operating system implement the Discretionary Access Control (DAC) security policy. DAC allows the information owner to control who is allowed to access the information.

The CAPP functional requirements fall in the following five broad categories:

- Identification and Authentication
- User Data Protection
- Security Management
- Protection of the TSF
- Security Audit.

4.2.1 Identification and Authentication

Identification and Authentication include the functionality required to uniquely identify the user.

SLES provides identification and authentication using pluggable authentication modules (PAM) based upon user passwords. Other authentication methods (e.g., Kerberos authentication, token based authentication) that are supported by SLES as pluggable authentication modules are not part of the evaluated configuration. PAM was configured to ensure medium password strength, to ensure password quality to limit the use of the su command, and to restrict root login to specific terminals.

Meeting the CAPP requirements for Identification and Authentication involved changing the default PAM configuration for SLES 8. The new configuration is documented by the Security Guide.

4.2.2 User Data Protection

User Data Protection specifies the functionality that protects data from unauthorized access and modification—the enforcement of the Discretionary Access Control policy. In addition, deleted information must not be accessible and newly created objects must not contain residual information.

The Discretionary Access Control policy restricts access to file system objects based on Access Control Lists (ACLs) that include the standard UNIX® permissions for user, group, and others. Access control mechanisms also protect IPC objects from unauthorized access.

The evaluated configuration used the ACL support in the ext3 file system. The vanilla kernel already clears file system, memory and IPC objects before they can be reused by a process belonging to a different user. Thus, the User Data Protection functionality requirements were already being met by SLES 8.

4.2.3 Security Management

The Security Management class specifies how security attributes, security data and security functions are managed by the TOE. Security Management includes management of groups and roles, separation of capability, and management of audit data.

Management of the security critical parameters of the TOE is performed by administrative users. Commands that require root privileges, such as useradd and groupdel, are used for system management. Security parameters are stored in specific files that are protected by the access control mechanisms of the TOE against unauthorized access by non-administrative users.

Other than the audit data management commands (which are described in the Security Audit section below) all security management functionality was provided by standard functionality already included in SLES 8.

4.2.4 TSF Protection

Protection of the TSF specifies the requirements for maintaining the integrity of the TSF and its data, particularly the protection of configuration data. The TSF will need to perform the appropriate testing to demonstrate the security assumptions about the underlying abstract machine upon which the TSF relies. In addition, the TSF must be demonstrated to be complete and tamperproof.

While in operation, the kernel software and data are protected by the hardware memory protection mechanisms. The memory and process management components of the kernel ensure that user processes cannot access kernel storage or storage belonging to other processes. Non-kernel TSF software and data are protected by DAC and process isolation mechanisms. In the evaluated configuration, the root user owns the directories and files that define the TSF configuration. Files and directories containing internal TSF data (e.g., configuration files, batch job queues) are also protected by DAC permissions.

The TOE and the hardware and firmware components are required to be physically protected from unauthorized access. The system kernel mediates all access to the hardware mechanisms themselves, other than program visible CPU instruction functions.

4.2.5 Abstract Machine Test Utility (AMTU)

To completely fulfill the TSF Protection requirement, we had to produce a tool to test the underlying abstract machine: "The TSF shall run a suite of tests [selection: during initial start-up, periodically during normal operation, or at the request of an authorized administrator] to demonstrate the correct operation of the security assumptions provided by the abstract machine that underlies the TSF."[4] This requirement is sometimes fulfilled by Power-On Self Test (POST) procedures, but given the diversity of platforms that were included in the certification, we decided that a userspace administrative tool, AMTU, would be the simpler approach. AMTU can be run by an administrator at any time and ensures that the hardware enforced security protection is still in effect. To this end, the tool runs a simple check for memory errors, checks for enforcement of memory separation, checks the correct operation of network and disk I/O controllers, and verifies

[4] Controlled Access Protection Profile available from `http://www.radium.csc.mil/tpep/library/protection_profiles/CAPP-1.d.pdf`

that privileged instructions cannot be executed when the hardware is in user mode.

The source code for AMTU is available at `http://www-124.ibm.com/developerworks/projects/amtu`.

4.2.6 Security Audit

Auditing systems collect information about events related to security-relevant activities. Security-relevant activities are defined as those events that are governed by the security policy. The resulting audit records can be examined to determine which security-relevant activity took place and which user is responsible for them. No fully CAPP-compliant audit subsystem was available for Linux, so we implemented this feature to achieve the certification. The audit subsystem developed for the evaluation is called Linux Audit System or LAuS.[5]

LAuS Conceptual Overview The Linux Audit System (LAuS) consists of three primary components: a kernel module responsible for intercepting system calls and recording relevant events, an audit daemon (auditd) that retrieves the records generated by the kernel and writes them to disk, and a number of command line utilities for displaying, querying and archiving the audit trail. See Figure 1.

The interface between kernel and user space uses a character device named /dev/audit. The audit daemon uses I/O Control operations (ioctls) on this device to configure the audit module, and it retrieves audit records from it using the read() system call.

To improve performance, filtering of audit events is performed at the kernel level. Unlike some existing implementations, the audit daemon does not perform any filtering itself. This eliminates a serious performance bottleneck.

The set of filter primitives provided by LAuS is fairly rich, and primitives can be combined using boolean operations. For instance, it is possible to audit open(2) calls made by a setuid application, while ignoring all other open(2) calls, or to restrict auditing to certain files. The eal3-certification RPM contains the evaluated audit configuration files.

At startup, auditd reads its configuration and the set of filter expressions from one or more files, loads the filters to the kernel, and starts auditing.

Auditd then proceeds to listen for audit events generated by the kernel. It retrieves and writes all records directly to disk. Because of the CAPP requirement that audit records must never be lost, this process is more complex than it might seem. auditd constantly monitors disk usage and can be configured to respond in different ways if free disk space drops below certain thresholds. Possible reactions to low disk space include notifying the administrator, suspending all audited processes, or shutting down the system immediately. Both the thresholds and auditd's reactions can be configured by the administrator.

LAuS supports different output modes to provide a flexible way to configure data collection. The simplest approach simply writes the audit trail to a single file in append mode, similar to the way syslogd works.

In "bin mode," audit writes data to a number of fixed sized files (bins), switches to the next file when the current one fills up, and invokes an external command to archive the full bin. Finally, there is a so-called "stream mode" that lets you pipe the audit trail directly into an ex-

[5] The LAuS Design Document is available at `ftp://ftp.suse.com/pub/projects/security/laus/doc/LAuS-Design.pdf`

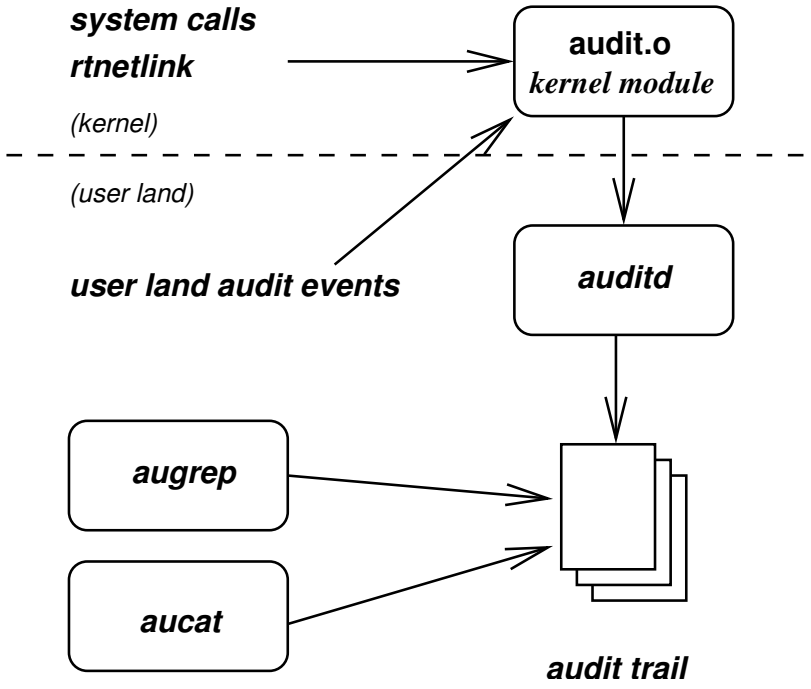


Figure 1: LAuS Conceptual Overview

ternal command; this can be useful if you want to forward the trail to a central storage server.

Auditing can be enabled globally or on a per-process basis; in the latter case, all the child processes are audited as well. The only processes always exempt from auditing are init and the audit daemon itself.

User land utilities were created to parse and read the audit log files. aucat 'cats' the file, transforming all of the audit records to a human readable format. augrep 'greps' the audit records and allows the administrator to selectively review the records. augrep allows the administrator to select audit records based on type, time (range), user, syscall, program (by name or PID) that generated the event, or any combination of these attributes.

Even though the user land utilities are far from trivial, the kernel portion of LAuS proved far more complex; in fact, the kernel portion of the LAuS is a lot more complex than we had initially anticipated. The rest of this section deals with the questions surrounding the audit kernel module.

Additional Design Constraints In addition to making our audit implementation compliant with the CAPP requirements, we had to deal with several constraints which are worth noting.

One was to minimize performance overhead. In the case where auditing was compiled into the kernel, but not configured by the administrator, we wanted it to have zero performance impact if possible. Our kernel developers spent quite a lot of time on additional kernel tuning, making sure the kernel performed and scaled well. Breaking this was not an option.

We also wanted to have a performance overhead as small as possible for the audited case, even though this wasn't as high on our agenda. This definitely took second place to correctness

and CAPP compliance.

A third objective was entirely non-technical, but played a crucial role in choosing an approach to intercepting system calls. We wanted our modifications to the core kernel as small as possible; most of the code should be inside a loadable module.

The rationale behind this was to minimize the probability of introducing bugs (except, of course, bugs in the audit code itself), and to ease maintenance.

The latter point was a fairly important item in the context of the SLES 8 kernel, which includes well above 1,500 additional patches applied on top of the mainline kernel. Updating SLES 8 to a new mainline kernel version was a bit of an adventure, so we wanted to avoid adding audit patches to the kernel that changed lots of files all over the place.

Where to intercept system calls There are basically three ways to intercept system calls on a 2.4 Linux kernel.

The first approach is to create wrappers for those system calls you wish to track, and replace the original function pointers in the system call table with those of the new wrapper functions. This sounds simple enough, and would also satisfy our requirements for zero performance impact in the non-audit case, and a minimally intrusive kernel patch. Unfortunately, this approach doesn't work on all architectures.

The next approach is to add hooks to all kernel functions that must be audited. The major drawback to this approach is that the kernel patch would touch lots of files in the kernel, which we wanted to avoid.

The third approach, which we chose, was to hook into the code path that intercepts system calls for ptrace. This intercept happens very early in the platform-specific assembler code, before the system-call function itself is invoked. The assembly code retrieves a set of flags associated with the calling process, and checks the `PT_TRACESYSbit`. If that bit is set, it jumps to a separate code branch dealing with ptracing. The same test is performed when returning from a system call.

In our audit implementation, we simply defined an additional task flag named PT_AUDITED, and extended the bit test in the system-call entry and exit code to test for both bits at the same time. This gave us system-call intercept with zero performance overhead in the normal, non-audited code path.

See Figure 2 for a picture showing the flow of control when auditing a system call.

Defining which system calls to audit By far, the most important part of auditing concerns system calls. As mentioned above, CAPP requires auditing all security-relevant system calls. We needed to determine which system calls are security relevant and which aren't.

The obvious ones are those that change the state of a process, the file system, or other system resources. These includes calls such as setuid, open, close, and setting the system's host name or clock. An audit implementation also needs to cover less obvious operations, such as binding a socket to a port, attaching shared memory segments, and performing ioctls.

Most system calls are fairly straightforward to handle, and much of the information on system calls and the arguments they take can be encoded statically in tables. Some calls, such as msgrcv, which comes in two versions on the i386 platform for historical reasons, were difficult to handle. 64-bit platforms usually require an additional table as they support a 32-bit sys-

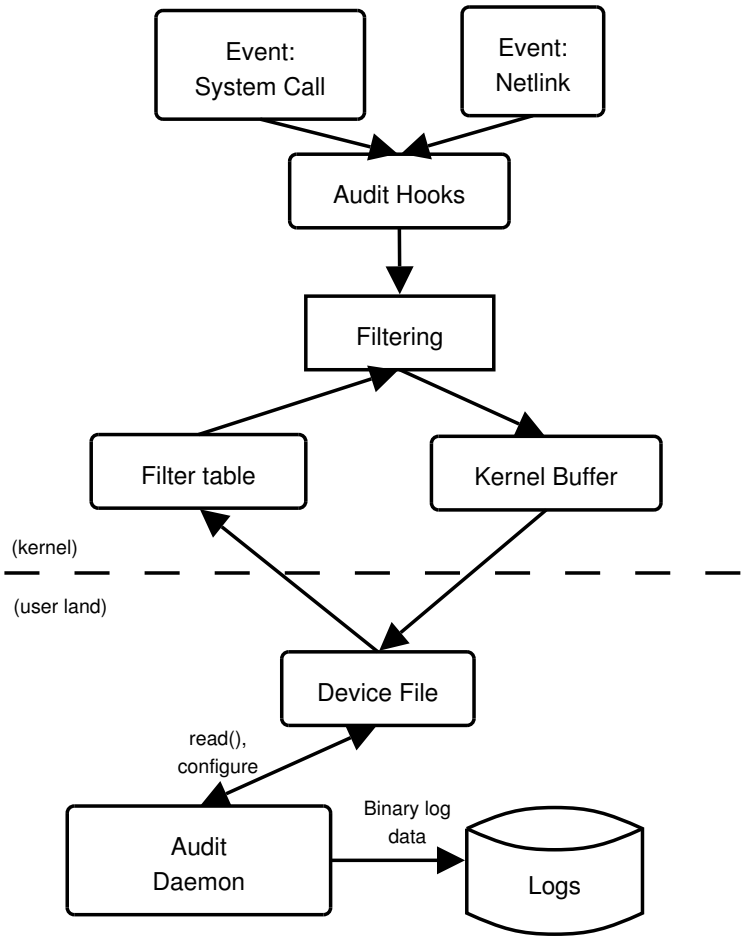


Figure 2: Auditing a System Call

tem call interface in addition to their native 64-bit interface.

However, some of the operations we wanted to audit proved a little more elusive; these were the ioctl system call and network configuration changes.

Auditing ioctls The ioctl system call is the dirty little back alley of UNIX-like operating systems. If a driver for a piece of hardware, a network protocol or a file system needs to expose some driver-specific mechanism or tunable parameters to user-space applications, the most common method for doing so is to define one or more ioctls.

The ioctl system call takes an open file descriptor, which must refer to something controlled by the driver (for example, a terminal, a device file, or a socket), an integer number specifying the request, and an opaque pointer to some chunk of memory. Exactly what to do with this piece of memory depends on the driver that is being talked to, and the integer passed as the request ID.

Unfortunately for us, the Linux kernel supports well over a thousand ioctls, and while many of them are rather obscure, they do change the system's state and are thus subject to auditing. It is obvious that compiling and maintaining a list of 1000 ioctls and their arguments was not an option.

Most ioctl numbers nowadays encode sufficient information on whether the operation passes data into the kernel, retrieves data, or both, and the size of the argument. Therefore, writing audit records for these is straightforward. However, there is still a fair number of ioctls that do not follow this convention.

What is far worse is that ioctl numbers are not unique—frequent users of strace will probably know that the TCGETS ioctl uses the same number as some obscure sound card operation. But this is not the only conflict.

However, the most difficult aspect of auditing ioctls is that it isn't sufficient to simply generate audit records for these calls; you must also be able to display the information of each audit record to the user.

The way we solved this problem was entirely non-technical. Our target of evaluation clearly stated that the super user account remains special. The super user can do everything, from loading unsupported modules not covered by the certification, to disabling the audit subsystem altogether.

Instead of trying to handle each and every ioctl in the audit module, we went through all ioctls available in our to-be-certified configuration and categorized them. The ones we needed to audit were those that were security relevant in some way, but did not require administrative privilege; the list we came up with this way was much smaller than the original list and more manageable.

Auditing network configuration Another aspect that proved to be a challenge was tracking network configuration changes, because only a fraction of those are done through ioctl calls. Most network configuration changes are performed by passing data to a netlink socket. These changes can be audited by simply recording all sendmsg and recvmsg calls on netlink sockets, but that is far from optimal. On the one hand, a send or receive operation on a netlink socket can include more than one request. On the other hand, the outcome of a netlink call is not returned through the system-call return value, but in a separate netlink message generated by the kernel and queued to that socket. Simply logging the raw netlink data sent and received would require quite a bit of built-in intelligence on the part of the user land applications that are supposed to display this data.

So instead, we decided to tap into the netlink layer directly, where a data blob sent to a netlink socket is broken up into separate requests, and each request is processed in return. This allowed us to record each netlink request separately, and place the outcome of the operation into the same audit record as the original request.

The Login User ID An aspect of auditing that is worth mentioning is how to deal with the CAPP requirement that each record identifies the user performing the operation.

The obvious solution (which would be to use the real user ID associated with the calling process) is not sufficient, as setuid applications can change these IDs at will. Tracking all uid changes, and thereby allowing the audit utilities to piece together the original user ID from this mosaic, is not practical either. It is not uncommon for some processes on a dedicated server to run for hundreds of days, so the amount of data to look at would be prohibitive.

The only viable solution in this case is to attach a "login uid" to each process. The login uid remained constant across all other changes of real, effective, and saved user IDs, and was inherited by all child processes. Of course, this required changes to PAM so that this uid would

be set on login.

Nightmare on Audit Street There is one major problem with the approach to system call intercept that we chose and which in hindsight made it a less than optimal choice. The problem is that our approach requires data to be copied twice. To understand why this is a problem, let's look at the open(2) system call, which takes a path name as an argument. This path name is passed into the kernel function as a pointer to a string (essentially a chunk of memory) in the address space of the user process. In order to operate on this string, the kernel must copy it to a buffer in the kernel address space, possibly paging in memory as it goes.

When entering the kernel, the audit module retrieves the path name from user space, and decides whether to create an audit record for this call. If it does decide to create an audit record, it sets up an audit record containing the system call number and a copy of all arguments, including the path.

The system call proceeds as normal, and the kernel function `sys_open` retrieves the path name from user space a second time, and carries out the requested operation based on this data.

The problem is that the memory in user space may have changed in the meantime, so that the record written by the audit module does not correspond to what was actually performed by the operating system.

There are several ways this can happen. Of course, the calling process itself cannot modify this memory, as it is currently executing the system call. However, memory can be shared between processes in a variety of ways. Threads can share the entire address space; processes can attach to the same shared memory segment; memory can be mapped from a file, which can be mapped by other processes as well.

Such an attack on the audit module is not really practical, because proper timing is probably quite hard, and any attempt to perform this attack would most likely leave a trail in the audit file. But even the theoretical possibility of circumventing the audit subsystem is unacceptable in terms of CAPP compliance.

The cases described above can be detected and dealt with by the audit module. Dealing with these problems, however, incurs additional complexity and performance loss (especially in the case of multithreaded applications). Needless to say, the added complexity engendered a considerable number of bugs. For this reason, these additional checks can be turned off by the administrator. These checks are turned off in the evaluated configuration of audit and the associated risk, considered minimal, is documented in the Vulnerability Assessment.

SUSE Linux Server 9 SUSE Linux Server 9 will include an updated version of the LAuS kernel patches. In many respects, the updated LAuS module will work in the same way as the SLES 8 version did, with the major exception being the way system calls are intercepted.

When planning audit for SUSE Linux Server 9, we considered two options.

The first option was a solution we had already looked at for SLES 8 and abandoned, namely adding hooks to all system call functions relevant for CAPP. This is the approach we chose for SUSE Linux Server 9, mainly in order to avoid having to jump through all those extra hoops in an attempt to prevent the race conditions described in the previous section. One pleasant side effect of this approach is that it also eliminated a lot of platform-specific code.

The second option we considered was to add audit support as an LSM module, or extending an existing LSM module such as SELinux. The security framework in the 2.6 kernel goes a long way toward intercepting all security-relevant operations. Adding audit hooks in this place is appealing, because it would mean no additional performance cost (the security hooks do come with a certain performance penalty already) and no additional maintenance problems (because the audit patch would not have to touch multiple kernel files).

The main reason why we did not choose this approach was that the security hooks provide a more abstracted view than we had chosen for LAuS in SLES 8. Security hooks do not correspond directly to system calls, but rather represent the security check necessary to validate whether an operation is permitted. There is a fine distinction between "user X attempted to perform operation Y, and the outcome was Z" and "user X attempted an operation on object A that caused us to perform security check B, and the outcome of this check was C." In particular, we are neither aware of the operation that triggered the security check, nor of its final outcome, because the operation can still fail even if security clearance is given.

Moreover, a single system call may require several security checks, such as renaming a file, where we need permission to remove the file from the source directory and permission to add it to the destination directory.

Changing LAuS to use the security hooks would have meant rewriting much more code than we wanted to, including the filtering code and much of the user-land applications. We also would have had to modify considerable parts of the documentation required for recertification.

Future Directions This is not to say that it is not possible to write an audit implementation leveraging some features of the LSM framework. In fact, we hope to have a common audit implementation in the mainstream kernel one day. It would greatly help acceptance by the kernel community if that solution did not add another set of hooks into many performance-critical functions.

5 Evaluation Roadmap

Performing a security evaluation should never be a one-time accomplishment. To maintain the security level achieved, the security certificate must be maintained. In the case of Linux, the intent is to go a step further: to increase, step-by-step, the assurance level and the security functionality until Linux achieves the highest assurance level of any commercial operating system product, while offering the richest set of security functions. The first step was accomplished in July 2003, when we obtained an EAL2+ evaluation for SUSE SLES 8 as-is. This paper documented the results of the second step, where we obtained a CAPP/EAL3+ certification for SLES 8 SP3 in January 2004. Linux, like its commercial competitors, has now been successfully evaluated for compliance with the requirements of the US government-defined CAPP. As a further step, Linux is currently in evaluation for compliance with the requirements of the EAL4 level. This includes the development of a low-level design of the Linux kernel (the evaluation will be based on the 2.6 version of the kernel) as well as a more sophisticated vulnerability analysis being performed. The experience gathered in the EAL2 and the EAL3 evaluations have given us the confidence that compliance with EAL4 can be achieved in fairly short order.

6 Value of Certification

The value of certification can be considered from two perspectives: business and technical.

In order for Linux to be adopted by the commercial and government markets, it faces stiff competition from entrenched incumbents. All of the incumbent products have been evaluated using the Common Criteria. In addition, the U.S government instituted a national security community policy against procuring unevaluated products (NSTISSP No. 11). There is a high probability that other governments and commercial entities will do the same.

While there is much skepticism surrounding the technical value of certification, certification is very much in line with the "many eyes" philosophy. For commercial products, certification is often the only time the code is reviewed by people outside of the development team. The assurance requirements of Common Criteria add to the number of trained eyes looking at the design and source of a project using defined and rigorous procedures. During the course of the EAL3 evaluation, we found and fixed several bugs, created lots of documentation, and shipped an integrated CAPP-compliant audit system. We noticed an anomaly on the iSeries platform while testing the Abstract Machine Testing Utility. Analysis of this anomaly by the ppc64 development team led to the discovery of a memory separation bug on the iSeries platform.[6] Many PAM module bugs were identified and fixed in SLES 8, including a double free bug in `pam_pwcheck`.[7] Man pages were created for several undocumented system calls, PAM modules and admin utilities, including `io_setup`, `readahead`, `set_thread_area`, `pam_wheel`, `pam_securetty`, and others.

[6] Paul Mackerras fix to "Make kernel RAM user-inaccessible on iSeries" `http://www.kernel.org/diff/diffview.cgi?file=/pub/linux/kernel/v2.4/patch-2.4.23.bz2;z=290`

[7] `http://www.atsec.com/01/index.php?id=03-0002-01&news=28` Patches are available from klaus@atsec.com

7 Conclusion

Achieving the EAL2 and EAL3/CAPP certifications was significant because it proved that Linux is indeed certifiable. The certification opened the market up to include U.S. government agencies and commercial entities that require certification. Future evaluations of Linux distributions can be made easier by Linux adoption of a CAPP-compliant audit subsystem.

8 Legal Statement

Improving Linux resource control using CKRM

Shailabh Nagar
IBM T.J. Watson Research Center
nagar@watson.ibm.com

Rik van Riel
Red Hat, Inc.
riel@redhat.com

Hubertus Franke
IBM T.J. Watson Research Center
frankeh@watson.ibm.com

Chandra Seetharaman
IBM Linux Technology Center
chandra.sekharan@us.ibm.com

Vivek Kashyap
IBM Linux Technology Center
vivk@us.ibm.com

Haoqiang Zheng
Columbia University
hzheng@cs.columbia.edu

Abstract

One of the next challenges faced in Linux kernel development is providing support for workload management. Workloads with diverse and dynamically changing resource demands are being consolidated on larger symmetric multiprocessors. At the same time, it is desirable to reduce the complexity and manual involvement in workload management. We argue that the goal-oriented workload managers that can satisfy these conflicting objectives require the Linux kernel to provide class-based differentiated service for all the resources that it manages. We discuss an extensible framework for class-based kernel resource management (CKRM) that provides policy-driven classification and differentiated service of CPU, memory, I/O and network bandwidth. The paper describes the design and implementation of the framework in the Linux 2.6 kernel. It shows how CKRM is useful in various scenarios including the desktop. It also presents preliminary performance evaluation results that demonstrate the viability of the approach.

1 Introduction

Workload management is an increasingly important requirement of modern enterprise computing systems. There are two trends driving the development of enterprise workload management middleware. One is the consolidation of multiple workloads onto large symmetric multiprocessors (SMPs) and mainframes. Their diverse and dynamic resource demands require workload managers (WLMs) to provide efficient differentiated service at finer time scales to maintain high utilization of expensive hardware. The second trend is the move towards specification of workload performance in terms of the business importance of the workload rather than in terms of low-level system resource usage. This has led to the increasing use of goal-oriented workload managers, described shortly, which are more tightly integrated into the business processes of an enterprise.

Traditional system administration tools have been built with two layers. The lower, OS specific layer deals with modifying and monitoring operating system parameters. The upper

layer(s) provide an OS independent API, generally through a graphical user interface, allowing a multi-tier or clustered system to be managed through a unified API despite containing heterogenous operating systems. While such tools provide a convenient administrative interface to heterogeneous operating systems they do little to address the complexity of managing workloads that span multiple tiers. The burden of translating business goals into workload resource requirements and the latter into OS specific tuning parameters remains on the system administrators. Increasing workload consolidation only adds more complexity to an already onerous problem.

As described in [7], the first stage in improving workload management are *entitlement-based* workload managers (WLMs) such as [9, 5, 11] which enforce entitlements or shares on resources consumed by groups of processes, users, etc. This allows the more important groupings to see improved response times and higher bandwidth due to preferential access to the server hardware. As importantly, it allows expensive SMP servers to have higher utilizations since system administrators can afford to load them more without fear of penalizing the important groupings.

However, the complexity of determining the right entitlements (henceforth called shares) for a grouping remains on the human system administrator. Not only does s/he need to map the importance of a workload to its entitlements, s/he also needs to adjust these shares dynamically when the demand and/or importance of *any* workload changes. Such dynamic share changes have become increasingly difficult to compute in a timely manner when manual involvement is part of the adaptive feedback loop.

To address the complexity of share specifications, *goal-oriented workload managers* have been developed [1, 10] which allow a system to be more self-managed. Such WLMs allow the human system administrator to specify high level performance objectives in the form of policies, closely aligned with the business importance of the workload. The WLM middleware then uses adaptive feedback control over OS tuning parameters to realize the given objectives.

In mainstream operating systems, including Linux, the control of key resources such as memory, CPU time, disk I/O bandwidth and network bandwidth is typically strongly tied to processes, tasks and address spaces and are highly tuned to maximize system utilization. This introduces additional complexity to the WLM which needs to translate the QoS requirements into these low level per task requirements, tough typically QoS is enforced at work class level. Hence, in order to isolate the autonomic goal oriented layers of the system management from the intricacies of the operating system, we introduce the class concept into the operating system kernel and require the OS to provide differentiated service for all major resources at a class granularity defined by the WLM.

In this paper, we discuss a framework called class-based kernel resource management (CKRM) that implements this support under Linux. In CKRM, a class is defined as a dynamic grouping of OS objects of a particular type (classtype) and defined through policies provided by the WLM. Each class has an associated share of each of its resources. For instance, CKRM tasks classes provides resource management for four principal physical resources managed by the kernel namely CPU time, physical memory pages, disk I/O and bandwidth. Sockets classes provide inbound network bandwidth resource control. The Linux resource schedulers are modified to provide differentiated service at a class granu-

larity based on the assigned shares. The WLM can dynamically modify the composition of a class and its share in order to meet higher level business goals. We evaluate the performance of the CKRM using simple benchmarks that demonstrate the efficacy of its approach.

This work makes several contributions that distinguish it from previous related work such as resource containers [2] and cluster reserves [4]. First, it describes the design of a flexible kernel framework for class-based management that can be used to manage both physical and virtual resources (such as number of open files). The framework allows the various resource schedulers and classification engine to be developed and deployed independent of each other. Second, it shows how incremental modifications to existing Linux resource schedulers can make them provide differentiated service effectively at a class granularity. To our knowledge, this is the first open-source resource management package that attempts to provide control over all the major physical resources—i.e., CPU, memory, I/O, and network. Third, it provides a policy-driven classification engine that eases the development of new higher level WLMs and enables better coordination between multiple WLMs through policy exchange. Thirdly, through the resource class filesystem the WLM goals can be manipulated by normal users, making it useful on the desktop. Finally, it develops a tagging mechanism that allows server applications to participate in their resource management in conjunction with the WLM.

The rest of the paper is organized as follows. Section 2 gives an overview of CKRM and its core bits. Sections 3 briefly describes the classification engine. Section 4 presents the facilities provided by CKRM for monitoring. The inbound network controller, the first major controller ported to CKRM's new interface, is described in Section 5. Section 6 describes the filesystem interface which replaces the system call interface used in CKRM's earlier design presented in OLS 2003 [13]. Section 7 describes how CKRM might be used, both on a desktop system and on some server workloads. Section 8 concludes with directions for future work in the project.

2 Framework

A typical WLM defines a workload to be any system work with a distinct business goal. From a Linux operating system's viewpoint, a workload is a set of kernel tasks executing over some duration. Some of these tasks are dedicated to this workload. Other tasks, running server applications such as database or web servers, perform work for multiple workloads. Such tasks can be viewed as executing in phases with each phase dedicated to one workload. Server tasks can explicitly inform the WLM of its phase by setting an application tag. A WLM can also infer the phase by monitoring significant system events such as forks, execs, setuid, etc. and classifying the server task as best as possible.

In this scenario, a WLM translates a high level business goal of a workload (say response time) into system goals for the set of tasks executing the workload. The system goals are a set of delays seen by the workload in waiting for individual resources such as CPU ticks, memory pages, etc. The WLM monitors the business goals, possibly using application assistance, and the system usage of its resources. If the business goal is not being met, it identifies the system resource(s) which form a performance bottleneck for the workload and adjusts the workload's share of the resource appropriately. The CKRM framework enables a WLM to regulate workloads through a number of components, as shown in Fig. 1:

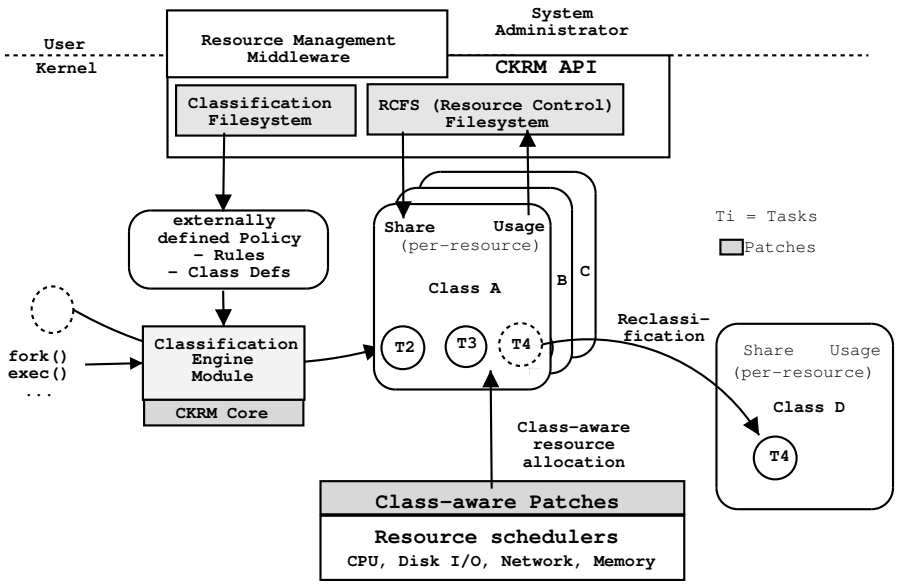


Figure 1: CKRM lifecycle

Core: The core defines the basic entities used by CKRM and serves as the link between all the other components. A class is a group of kernel objects with an associated set of constraints for resource controllers operating on those kernel objects—e.g., a class could consist of a group of tasks which have a joint share of cpu time and resident page frames. Each class has an associated classtype which identifies the kernel object being grouped. CKRM currently defines two classtypes called `task_class` and `socket_class` for grouping tasks and sockets. For brevity, the term taskclass and socketclass will be used to denote a class of classytpe `task_class` and `socket_class` respectively. Classtypes can be enabled selectively and independent of each other. A user not interested in network regulation could choose to disable `socket_classes`. Classes in CKRM are hierarchical. Children classes can be defined to subdivide the resources allocated to the parent.

Classification engine (CE): This optional components assists in the association of kernel objects to classes of its associated classtype. Each kernel object managed by CKRM is always associated with some class. If no classes are defined by the user, all objects belong to the default class for the classtype. At significant kernel events such as fork, exec, setuid, listen, when the attributes of a kernel object are changed, the Core queries the CE, if one is present, to get the class into which the object should be placed. CE's are free to use any logic to return the classification. CKRM provides a rule-based classification engine (RBCE) which allows privileged users to define rules which use attribute matching to return the class. RBCE is expected to meet the needs of most users though they can define their own CE's or choose not to have any and rely upon manual classification of each kernel object through CKRM's rcfs user interface (described later).

Resource Controllers: Each classtype has a set of associated resource controllers, typically one for each resource associated with the classtype—e.g., taskclasses have cpu, memory, and I/O controllers to regulate the cpu ticks, resident page frames and per-disk I/O bandwidth consumed by it while socketclasses have an accept queue controller to regulate the num-

ber of TCP connections accepted by member sockets. Resource requests by a kernel object in a class are regulated by the corresponding resource controller, if one exists and is enabled. The resource controllers are deployed independent of each other so a user interested only in controlling CPU time for taskclasses could choose to disable the memory and I/O controllers (as well as the socketclass classtype and all its resource controllers).

Resource Control File System (RCFS): It forms the main user-kernel interface for CKRM. Once RCFS is mounted, it provides a hierarchy of directories and files which can be manipulated using well-known file operations such as open, close, read, write, mkdir, rmdir and unlink. Directories of rcfs correspond to classes. User-kernel communication of commands and responses is done through reads/writes to virtual files in the directories. Writes to the virtual files trigger CKRM Core functions and responses are available through reads of the same virtual file.

The CKRM architecture outlined above achieves three major objectives:

- Efficient, class-based differentiation of resource allocation and monitoring for dynamic workloads: Regulate and monitor kernel resource allocation by classes which are defined by the privileged user and not only in terms of tasks. The differentiation should work in the face of relatively rapid changes in class membership and over roughly the same time intervals at which process-centric regulation currently works.

- Low overhead for non-users: Users disinterested in CKRM's functionality should see minimum overhead even if CKRM support is compiled into the kernel. Signs of user disinterest include omitting to mount rcfs or not defining any classes. Even for users, CKRM tries to keep overheads proportional to the features used.

- Flexibility and extensibility through minimization of cross-component dependencies: Classification engines should be independent of classtypes and optional, classtypes should be independent of each other and so should resource controllers, even within the same classtype. This goal is achieved through object-oriented interfaces between components. Minimizing dependencies allows kernel developers to selectively include components based on their perception of its utility, performance and stability. It also permits alternative versions of the components to be used depending on the target environment—e.g., embedded Linux distributions could have a different set of taskclass resource controllers (or even classtypes) than server-oriented distributions.

3 Classification

The Classification Engine (CE) is an optional component that enables CKRM to automatically classify kernel objects within the context of its classtype. Since the CE is optional and since we want to main flexibility in its implementation, functionality and deployment, it is supplied as a dynamically loadable module. The CE interacts with CKRM core as follows. The CKRM core defines a set of ckrm events that constitute a point during execution where a kernel object could potentially change its class. A classtype can register a callback at any of these events. As an example, the task class hooks the fork, exec, exit, setuid, setgid calls where as the socket class hooks the listen and accept calls. In these callbacks the classtypes typically invoke the optional CE to obtain a new class. If no CE is registered or the

CE does not determine a class, the object remains in its current class, otherwise the object is moved to the new class and the corresponding resource managers of that class's type are informed about the switch.

For every classtype the CE wants to provide automatic classification for, it registers a classification callback with the classtype and the set of events to which the callback is limited to. The task of CE is then to provide a target class for the kernel objects passed in the context of the classtype. For instance, task classes pass only the task, while socket classes pass the socket kernel object as well as the task object. Though the implementation of the classification engine is completely independent of CKRM, the CKRM project provides a default classification, called RBCE, that is based on classification rules. Rules consist of a set of rule terms and a target class. A rule term specifies one particular kernel object attribute, a comparision operator (=,<,>,!) and a value expression. To speed up the classification process we maintain state with tasks about which rules and rule terms have been examined for a particular task and only reexamine those terms that are indicated by the event. RBCE provides rules based on task parameters ((pid, gid, uid, executable) and socket information (IP info). The rules in conjunction with the defined classes constitute a site policy for workload managment and is dynamically changable (See user interface section) into the RBCE. Hence, this approach ensures the separation of policy and enforcement.

To facilitate the interaction with WLMs to provide event monitoring and tracing, the CE can also register a notification callback with any classtype, that is called when a kernel object is assigned to a new class. Similar so the classification callback, the notification callback can be limited to a set of ckrm events. This facility is utilized in resource monitoring, described next.

4 Monitoring

We now describe the monitoring infrastructure. Strictly speaking, the per-class monitoring components are part of CKRM while the per-process components are not. However, we shall describe them together as they both can be utilized by goal-based WLMs. Furthermore, they are bundled with the classification engine and utilize the CE's notification callback to obtain classification events. The monitoring infrastructure illustrated in Fig. 2 is based on the following design principles:

1. Event-driven: Every significant event in the kernel that affectsthe state of a task is recorded and reported back to the state-agent. The events of importance are aperiodic such as process fork, exit and reclassification as well as periodic events such as sampling. Commands sent by the state-agent are also treated as events by the kernel module.

2. Communication Channel: A single logical communication channel is maintained between the state-agent and the kernel module and is used for transferring all commands and data. Most of the data flow is from the kernel to user space in the form of records resulting from events.

3. Minimal Kernel State: The design minimizes the additional per-process state that needs to be maintained within the kernel. Most of the state needed for high level control purposes is kept within the state agent and updated through the records sent by the kernel.

The state-agent, which can also be integrated within a WLM, maintains state on each existing and exited task in the system and provides it to the WLM. Since the operating system

does not retain the state of exited processes, the stateagent must maintain it for future consumption by the WLM. The state-agent communicates with a kernel module through a single bidirectional communication chan-nel, receiving updates to the process state in the form of records and occasionally sending com-mands. Events in the kernel such as process fork, exit, reclassify (resulting from change in any process attribute such as gid, pid) cause records to be generated through functions provided by the kernel module.

Server tasks can assist the WLM by informing it about the phase in which they are operating (each phase corresponds to a workload). Such tasks invoke CKRM to set a tag associated with their `task_struct` in the kernel. CKRM uses this event to reclassify the task and also records the event (to be transmitted to the WLM through the state-agent). Other kernel events that might cause a task to be reclassified (such as the exec and setuid system calls, etc.) are also noted by CKRM and passed to the WLM through the state-agent. In addition, CKRM performs periodic sampling of each task's state in the kernel to determine the resource it is waiting on (if any), its resource consumption so far and the class to which it belongs. The sample information is transmitted to the state-agent. The WLM can correlate the information with the tag setting to statistically determine the resource consumption and delays of both server and dedicated processes executing a workload. Sampling is done through a kernel module function that is invoked by a selfrestarting kernel timer. Commands sent by the state-agent cause appropriate functions in the kernel module to execute and also return data in the form of records. The kernel components are kept simple and only minimal additional state has to be maintained in the kernel. In particular, the kernel does not have to maintain extra state about exited processes which introduces problems with PID reusage, memory management to name a few. Instead, relevant task information is replicated in user space, is by definition received in the correct time order (see below) and can be kept around until the WLM has consumed the information. Furthermore, the semantics of a reclassification in the kernel, which identifies a new phase in a server process, does not have to be introduced into the kernel space.

The following small changes are required to the linux kernel to track system delays. The `struct delay_info` is added to the `task_struct`. Delay_info contains 32-bit variables to store cpu delay, cpu using, io delay and memory io delay. The counters provide micro second accuracy. The current cpu scheduler records timestamps whenever i) a task becomes runnable and is entered into a runqueue and ii) when a context switch occurs from one task to another. We use these same timestamps to get per-task cpu wait and cpu using times recorded respectively. I/O delays are measured by the difference of timestamps taken when a task blocks waiting for I/O to complete and when it returns. All I/O is normally attributed to the blocking task. Page-fault delays, however, are treated as special I/O delays. On entrance to and exit from the page fault handler the task is marked or unmarked as being in a memory path using flags in `task_struct`. If during the I/O delay, this flag is set, the I/O delay is counted as a memory delay instead of as a pure I/O delay. The per-task delay information is accessible through the file `/proc/<pid>/delay`. Similarly, each class contains a `delay_info` structure.

In contrast to the precise accounting of delays, sampling examines the state of tasks at fixed interval. In particular, we sample at fixed intervals (~1sec) the entire set of tasks in the system and increment per task counters that are integrated into the task private structure attached

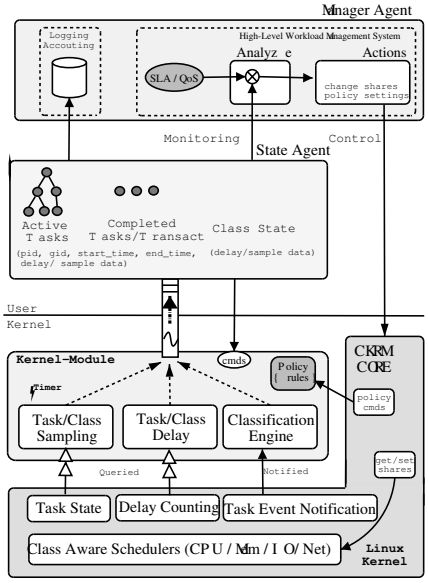


Figure 2: CKRM lifecycle

by the classification engine that builds the core of the kernel module. We increment counters if a task is running, waiting to run, performing I/O or handles a pagefault I/O. Task data (sampled and/or precise) is requested by and sent to the state-agent in coarser intervals. We can send data in continuous aggregate mode or in delta mode, i.e. only if task data has changed do we send a new data record and then reset the local counters. The task transition events are sent at the time they occur. We distinguish the fork, exit, and reclassification events as records. At each reclassification (which could potentially be the end of a phase) we transmit the sample and delay data and reset them locally.

As a communication channel we utilize the linux relayfs pseudo filesystem, a highly efficient mechanism to share data between kernel and user space. The user accesses the shared buffers, called channels, as files, while the kernel writes to them using buffer reservations and memory read/write operations. The content and structure of the buffer is determined by the kernel and user client. Currently the communication channel is self pacing. The underlying relayfs channel buffer will dynamically resize upto a maximum size. If for any reason the relayfs buffer overflows, record sending will automatically stop, an indication is sent and the state-agent will have to drain the channel and request a full state dump from the kernel.

We have measured the data rate during a standard kernel build, which creates a significant amount of task events (fork,exec,exits). For a 2-CPU system with 2 seconds sample collection we observed a data rate of 8KB/second and a total of 190 records/sec, well within a limit that can be processed without creating significant overhead in the system.

5 Inbound Network

Various OS implementations offer well established QoS infrastructure for outbound bandwidth management, policy-based routing and Diffserv [3]. Linux in particular, has an elaborate infrastructure for traffic control [8] that consists of queuing disciplines(qdisc) and filters. A qdisc consists of one or more queues and a packet scheduler. It makes traffic con-

form to a certain profile by shaping or policing. A hierarchy of qdiscs can be constructed jointly with a class hierarchy to make different traffic classes governed by proper traffic profiles. Traffic can be attributed to different classes by the filters that match the packet header fields. The filter matching can be stopped to police traffic above a certain rate limit. A wide range of qdiscs ranging from a simple FIFO to classful CBQ or HTB are provided for outbound bandwidth management, while only one ingress qdisc is provided for inbound traffic filtering and policing. The traffic control mechanims can be used invarious places where bandwidth is the primary resource to control.

Due to the above features, Linux is widely used for routers, gateways, edge servers; in other words, in situtations where network bandwidth is the primary resource to differentiate among classes. When it comes to endservers networking, QoS has not received as much attention since QoS is primarily governed by the systems resources such as memory, CPU and I/O and less by network bandwidth. When we consider end-to-end service quality, we should require networking QoS in the end servers as exemplified in the fair share admission control mechanism proposed in this section.

We present a simple change to the existing TCP accept mechanism to provide differentiated service across priority classes. Recent work in this area has introduced the concept of prioritized accept queues [6] and accept queue schedulers using adaptive proportional shares to self-managed web [14]. In a typical TCP connection, the client initiates a request to connect to a server. This connection request is queued in a global accept queue belonging to the socket associated with the server's port. The server process picks up the next queued connection request and services it. In effect, the incoming connections to a particular TCP socket are serialized and handled in FIFO order. When the incoming connection request load is higher than the level that can be handled by the server requests have to wait in the accept queue until the next can be picked up.

We replace the existing single accept queue per socket with multiple accept queues, one for each priority class. Incoming traffic is mapped into one of the priority classes and queued on the accept queue for that priority. The accept queue implements a weighted fair scheduler such that the rate of acceptance from a particular accept queue is proportional to the weight of the queue. In the first version of the priority accept queue design initially proposed by the CKRM project [13], starvation of certain priority classes was a possibility as the accepting process picked up connection requests in the order of descending priority.

The efficacy of the proportional accept queue mechanism is demonstrated by an experiment. We used Netfilter [12] to MARK options to characterize traffic into two priority classes with respective weights of 3:1. The server process utilises a configurable number of threads to service the requests. The results are shown in Figure 3. When the load is low and there are service threads available no differentiation takes place and all requests are processed as they arrive. Under higher load, requests are queued in the accept queue with class 1 receiving a proportionally higher service rate than class 2. The expriment was repeated, maintaining a constant inbound connection request rate. The proportions of the two classes were then switched to see the service rate for the two classes reverse as seen in Figure 4

6 Resource Control Filesystem

In the Linux kernel development community, filesystems have become very popular as user

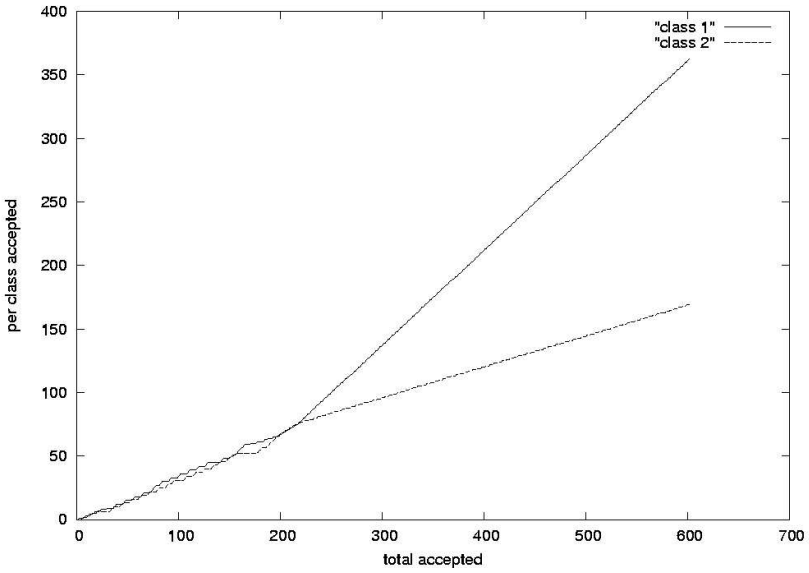


Figure 3: Proportional Accept Queue: Results

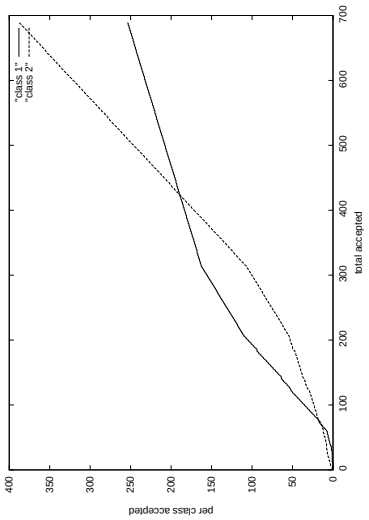


Figure 4: Proportional Accept Queue: Results under change

interfaces to kernel functionality, going well beyond the traditional use for disk-based persistent storage. The Linux kernel's object-oriented Virtual File System (VFS) makes it easy to implement a custom filesystem. Common file operations like open, close, read and write map naturally to initalization, shutdown, kernel-to-user and user-to-kernel communication. For CKRM, the tree structured namespace of a filesystem offers the additional benefit of an intuitive representation of the class hierarchy. Hence CKRM uses the Resource Control Filesystem (RCFS) as its user interface.

The first-level directories in RCFS contain the roots of subtrees associated with classtypes build or loaded into the kernel (`socket_class` and taskclass currently) and the clas-

sification engine (ce). Within the classtype subtrees, directories represent classes. Users can create new classes by creating a directory as long as they have the proper access rights. Within the `task_class` directory, each directory represents a task class. `/rcfs/taskclass`, the root of the `task_class` classtype, represents the default taskclass which is always present when CKRM is enabled in the kernel. Each `task_class` directory contains a set of virtual files that are created automatically when the directory is created. Each virtual file has a specific function as follows:

1. members: Reading it gives the names of the tasks in the taskclass.
2. config: To get/set any configuration parameters specific to the taskclass.
3. target: Writing a task's pid to this file causes the task to be moved to the taskclass, overriding any automatic classification that may have been done by a classification engine.
4. shares: Writing to this file sets new lower and upper bounds of the resource shares for the taskclass for each resource controller. Reading the file returns the current shares. The controller name is specified on a write which makes it possible to set the values for controllers independent of each other.
5. stats: Reading the file returns the statistics maintained for the taskclass by each resource controller in the system. Writing to the file (specifying the controller) resets the stats for that controller.

The `socket_class` directory is somewhat similar. Directories under `/rcfs/socket_class/` represent listen classes and have the same magic files as task_classes. Whereas task_classes use the pid to identify the class member, socket_classes, which group listening sockets, use ip address + port name to identify their members. Within each listen class, there are automatically created directories, one for each accept queue class. The accept queue directories, numbered 1 through 7, have their own shares and stats virtual files similar to those for `task_classes`.

The `/rcfs/ce` directory is the user interface to the optional classification engine. It contains the following virtual files and directory:

1. reclassify: writing a pid or ipadress+port to the file causes the corresponding task or listen socket to be put back under the control of the classification engine. On subsequent significant kernel events, the ce will attempt to reclassify the task/socket to a new taskclass/socketclass if the task/sockets attributes have changed.
2. state: to set/get the state (active or inactive) of the classification engine. To allow a new policy to be loaded atomically, CE's can be set to inactive before loading a set of rules and activated thereafter.
3. Rules: The directory allows privileged users to create files with each file representing one rule. Reading the files, permitted for all, gives the classiication policy which is currently active. The ordering of rules in a policy is determined either by creation time of the corresponding file or by an explicitly specified order number within the file. The rule files contain rule terms consisting of attribute-value pairs and a target class. E.g., the rule *gid=10, cmd = bash, target = /rcfs/taskclass/A* indicates that tasks with gid 10 and running the bash program (shell) should get reclassified to task_class A.

7 Example uses

In this section we will describe a number of uses for CKRM, ranging from the traditional large server workload consolidation, to a university shell server, to the desktop—a novel use of workload management systems, made possible through the resource class filesystem.

7.1 Workload Consolidation

The classical use of a workload management system is workload consolidation, whether it's multiple departmental database servers on one large server, or one small server balancing resources between apache, ftpd, postfix and the interactive users. In either scenario the main objective is to make sure that none of the workloads can, through excessive resource use, cause the machine to become unusable for any of the others.

The simple solution is to start each of the services up in their own resource class and guaranteeing a certain amount of resources (say, 10% of the CPU and 20% of memory) for each of the services. Simultaneously the services can also have resource limits (say, 50% of memory). This combination of guarantees and limits gives the system a certain amount of freedom to balance the actual amount of resources each workload gets, while still putting effective guarantees and limits in place.

7.2 Shell Server

A shell server at a university faces a number of challenges. For example, the staff and postdocs should be protected from the load the students put on the machine and the students should be protected from each other. Similarly, batch jobs will usually have larger resource use limits (e.g. max cpu time used, max memory allocated), but a lower resource priority, as compared to any of the interactive programs. These problems can be solved by starting each class of process in the right process class.

On the other hand, if a staff member sends email to a student, the resources used by the student's mail filter should be accounted against that student's limits. This problem cannot be solved by having programs start out in a certain resource class, since the MTA process needs to transition between resource classes automatically. This can be solved by setting up a classification engine to automatically transfer a process to the *email* resource class when it execs `/usr/sbin/sendmail`. Similarly, when `/usr/bin/procmail` is being execed with a certain UID, the classification engine can move the process to the resource class where that user's interactive processes would normally run.

7.3 Desktop

With the right file and directory ownerships in the resource class filesystem, CKRM can be used in an area where traditional resource management systems tend to be cumbersome: on the desktop. A typical desktop configuration would have as its main goals that the system remains responsive to the user, no matter the background load, and would look something like the following.

The X server would get a good resource guarantee, e.g. 20% of CPU time and 20% of RAM. This makes sure that no matter what other processes run on the system, X can run smoothly and react to the console user with acceptably low latency.

At login time a PAM module would make sure that the rest of the user's processes get a good resource guarantee, too. An acceptable guarantee would be 50% of CPU time and 50% of RAM. This leaves enough resources free so that other things in the system can run (e.g. dis-

tro updates, updatedb, mail delivery), yet keeps most of the system dedicated to the user. The resource class created for the console user, e.g. `/rcfs/taskclass/console`, is set up to be writable for the console user. This way the user's processes can set resource guarantees and limits to certain classes of applications.

The user's GUI menu would take care of this subdividing of the resources guaranteed to the user. For example, the web browser could be restricted to 40% of RAM, so as to not put much pressure on the user's other processes. Multimedia processes could get part of the user's resource guarantees, e.g. 30% of the CPU and 10% of RAM guaranteed for the multimedia applications. This way the playback of multimedia should remain smooth, regardless of what the user's web browser and office suite are doing.

No superuser privileges are needed to configure these resource classes, or to move the user's processes between them. Any GUI framework or individual application will be able to determine the resources allocated to it, leading to more flexibility than possible with resource management systems that can only be configured by the super user. Note that since the user cannot raise the resource limits or guaranteed allocated to his main class, there should be no security risks involved with letting the user processes manipulate their own resource guarantees and limits.

8 Conclusion and Future Work

The consolidation of increasingly dynamic workloads on large server platforms has considerably increased the complexity of systems management. To address this, goal-oriented workload managers are being proposed which seek to automate low-level system administration requiring human intervention only for defining high level policies that reflect business goals.

In an earlier paper [13], we had argued that goal-oriented WLMs require support from the operating system kernel for class-based differentiated service where a class is a dynamic policy-driven grouping of OS processes. We had introduced a framework, called class-based kernel resource management, for classifying tasks and incoming network packets into classes, monitoring their usage of physical resources and controlling the allocation of these resources by the kernel schedulers based on the shares assigned to each class.

In this paper, we have described more details of the evolving design. In particular, CKRM has become more generic and supports groups of any kernel object involved in resource management, not just tasks. It has a new filesystem-based user API. Finally, the design introduces hierarchies into classes which permits greater flexibility for resource managers but also introduces challenges for CKRM controllers. A working prototype which includes an inbound network controller has been developed and made available through [15].

Future work in the project will involve redeveloping controllers for CPU, memory and I/O that are not only class-aware but can handle hierarchies of classes while keeping overheads low. Another important direction is the interactions of the resource schedulers and the impact of these interactions on the shares specified.

References

[1] J. Aman, C.K. Eilert, D. Emmes, P. Yocom, and D. Dillenberger. Adaptive algorithms for Managing a Distributed Data Processing Workload. In *IBM Systems Journal*, volume 36(2), 1997.

[2] G. Banga, P. Druschel, and J.C. Mogul. Resource containers: A new facility for resource management in server systems. In *Operating Systems Design and Implementation*, pages 45–58, 1999.

[3] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss. An Architecture for Differentiated Services. RFC 2475, Dec 1998.

[4] J. Blanquer, J. Bruno, E. Gabber, B. Ozden, and A. Silberschatz. Resource management for qos in eclipse/bsd. In *Proc. FreeBSD 1999 Conference*, Oct 1999.

[5] IBM Corp. AIX 5L Workload Manager. http://www.redbooks.ibm.com/redbooks/SG245977.html.

[6] IBM DeveloperWorks. Inbound connection control home page. http://www-124.ibm.com/pub/qos/paq_index.html.

[7] Inc. D.H.Brown Associates. HP Raises the Bar for UNIX Workload Management. http://h30081.www3.hp.com/products/wlm/docs/hp.raises.bar.wkld.mgmt.pdf.

[8] Bert Hubert. Linux Advanced Routing & Traffic Control. http://www.lartc.org.

[9] Hewlett Packard Inc. HP Process Resource Manager. http://h30081.www3.hp.com/products/prm/.

[10] Hewlett Packard Inc. HP-UX Workload Manager. http://h30081.www3.hp.com/products/wlm/.

[11] Sun Microsystems Inc. Solaris Resource Manager. http://www.sun.com/software/resourcemgr/wp-srm/.

[12] J. Kadlecsik, H. Welte, J. Morris, M. Boucher, and R. Russel. Netfilter: Firewalling, NAT, and packet mangling for Linux 2.4. http://www.netfilter.org.

[13] S. Nagar, H. Franke, J. Choi, M. Kravetz, C. Seetharaman, V. Kashyap, and N. Singhvi. Class-based prioritized resource control in Linux. In *Proc. 2003 Ottawa Linux Symposium, Ottawa*, July 2003. http://ckrm.sf.net/documentation/ckrm-ols03-paper.pdf.

[14] P. Pradhan, R. Tewari, S. Sahu, A. Chandra, and P. Shenoy. An Observation-based Approach Towards Self-Managing Web Servers. In *IWQoS 2002*, 2002.

[15] CKRM Open Source Project. Class-based kernel resource management. http://ckrm.sf.net/.

Linux on a Digital Camera

Porting 2.4 Linux kernel to an existing digital camera

Alain Volmat
Ricoh Company Ltd.
avolmat@src.ricoh.co.jp

Shigeki Ouchi
Ricoh Company Ltd
shigeki@src.ricoh.co.jp

Abstract

The RDC-i700 is one of high specs digital camera of Ricoh. Its relatively big size, large amount of different interfaces, input methods (buttons, or touch panel), have made it a good candidate for prototyping the world first Linux embedded digital camera. This paper presents our experiences of porting the 2.4 linux kernel to an existing digital camera. (the RDC-i700 is originally build on top of VxWorks). Eventhough embedded systems running on Linux are getting more and more popular, the digital camera field remains to be unexplored. The paper introduces how digital cameras differ from any other PC-like devices (PDA, HDD recorder...) and what problems, such as timing or software design issues, have to be (have been) solved in order to get the world first linux digital camera running on the linux 2.4 kernel.

1 The hardware

Ricoh's RDC-i700 [1] is a relatively old digital camera (released late 2000 in japan) running on VxWorks, a famous Real-Time OS (RTOS). Some might be asking the reason why we decided to port Linux OS to the camera. The reason is to make it become a *programmable camera*. Once it becomes a programmable device, many VARs or individual programmers may write a lot of useful software for it. Then it will be a good platform for business imaging use.

Figure 1: The RDC-i700 digital camera

The RDC-i700 is one of high specs digital camera of Ricoh. It integrates all peripherals traditional digital camera has, but also several different interfaces, allowing wide range of application to run on it. The VxWorks version allows user to perform various tasks such as taking picture or movie, recording voice memo, browse the Internet, send email or upload picture to a remote server. Its relatively big size, large amount of different interfaces, input methods (buttons, or touch panel), makes it a good candidate for prototyping the world first Linux embedded digital camera.

The RDC-i700 is a 3.2 million pixels digital camera equipped with a Hitachi SH3

[1] http://www.ricohzone.com/product_rdci700.html

(SH7709A) CPU. The SH7709A is a 32 bit RISC CPU which include MMU and several other peripherals such as serial communication interfaces (SCIs), D/A - A/D converters. Around this CPU, traditional digital camera peripherals (CCD, LCD, buttons, Image Processor) but also 1 PCMCIA and 1 CF socket, touch panel, audio input/output interface, USB device controller and a serial port are available. Figure 2 shows a block diagram of the RDC-i700.

2 Digital camera is not "PDA combined with camera function"

Nowadays, embedded Linux has become a very hot topic in the Linux community. More and more Linux gadget are becoming available and the share of embedded related paper published has literally exploded in the last 3 years. Linux seems to be everywhere, lots of devices that were running on RTOS in the past are now running on Linux. However, one field seems to be still unexplored: digital camera. Some might say that a digital camera is just a PDA combined with a CCD (this kind of combination is actually already available, for example the Zaurus CF Digital Camera option), but this is not that simple.

The quality point of view: PDA combined with a digital camera option can take pictures or even movies and in that sense can be compared to a digital camera. But digital cameras still have some advantages that make them irreplaceable. Indeed optical zoom, but also autofocus or strobe are all precious elements that are currently not available on Linux PDA. For example, the Zaurus camera has a focus but this one is manual. Auto-exposure is also another very important part when taking picture; for that, Zaurus PDA has some-kind of gain control but this cannot have same quality as a traditional digital camera auto-exposure system.

The technical point of view: We will see that having digital camera specific peripherals is a very good plus in term of quality, but it also creates lots of problem that traditional Linux PDA doesn't face. Keeping the Zaurus PDA as an example, only few parameters are configurable and the CPU doesn't actually have to perform much work in order to get an image. On the contrary, in case of a fully configurable digital camera, the OS must orchestrate all devices in order to get a picture.

2.1 Zoom and Focus

Several motors are used inside the camera. Two of them are used for zoom and focus in order to adjust the lens position. Due to high precision requirements, those two motors are stepping motors. As the name says, this kind of motors are controlled step by step (at the difference with traditional motors which only have start/stop command). The CPU has to set ports of the motor at a quite fast frequency in order to make the motor turn. In that case the period between two steps is only few milli-seconds.

2.2 Strobe

In case of strobe, the problem is not doing thing at very high speed, but making perfect synchronization between the moment the strobe is going to flash and the moment the CCD sensor will acquire the picture.(see figure 3) For that purpose, we will need precision of only very few milli-seconds.

2.3 Auto-Exposure / White Balance

So called Auto-Exposure is the algorithm in charge of adjusting the exposure time (that is to say the time period while the CCD is exposed to light) in order to have a good image

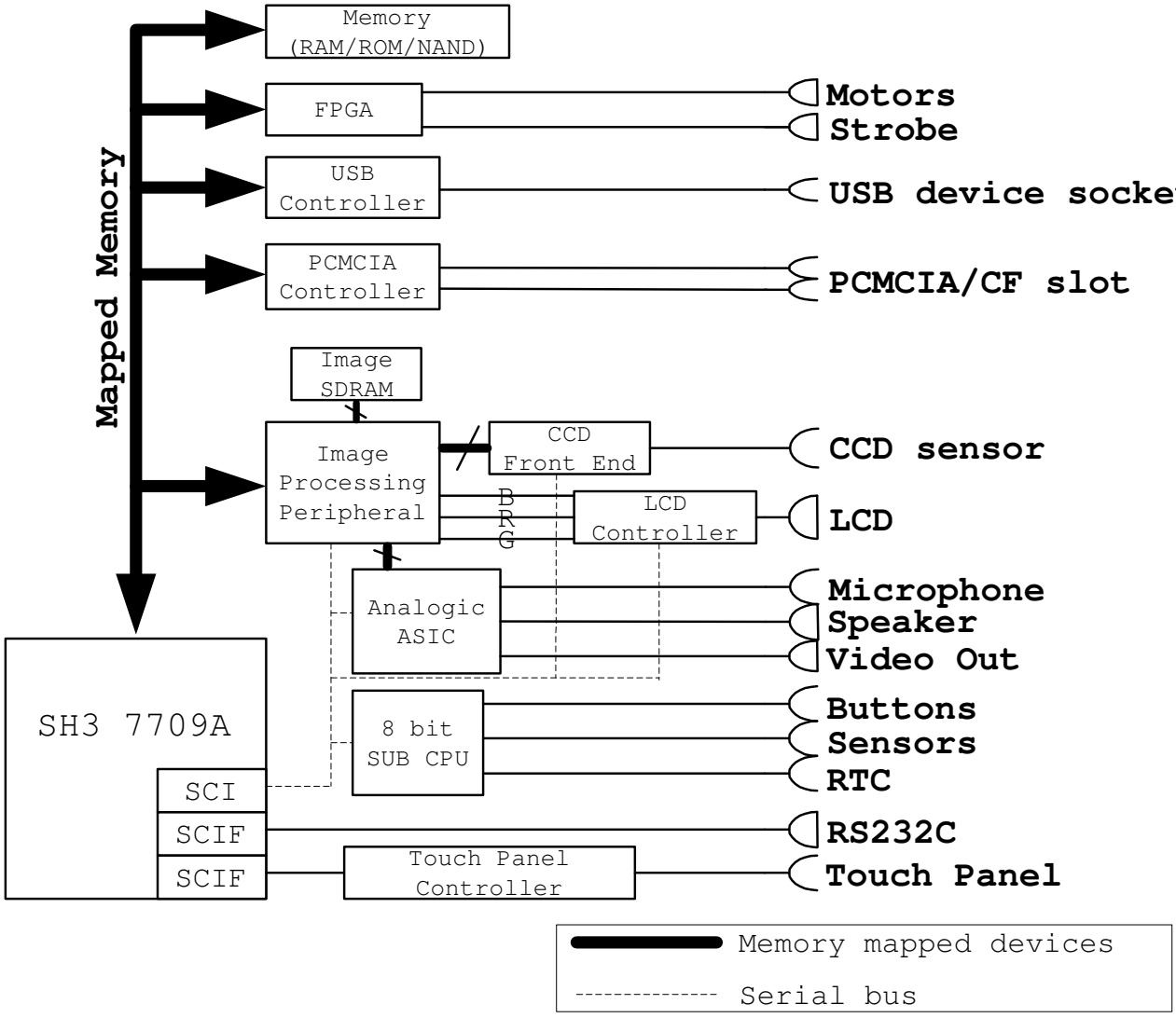


Figure 2: The RDC-i700 peripherals diagram

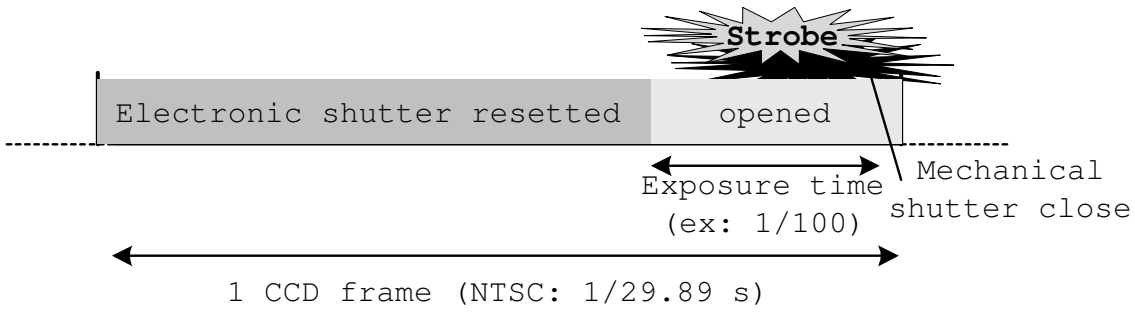


Figure 3: CCD Frame sequence

in both dark and bright conditions. White balance algorithm needs to analyze data coming from the CCD sensor and adjusts the Image Processing Peripheral (IPP) settings in order to have good color matching between the generated image and the reality. Several implementations of those two algorithms can exist (some needs very heavy calculations while others can be very simple), but the main issue is that those algorithms have to be performed very often (in the worst case, every frame of the CCD, that is to say every 33 milli-seconds).

This second part introduced particularity of digital cameras. The next part will discuss how those functions have been implemented into the Linux RDC-i700.

3 Current Support

As the name "Linux on a digital camera" suggests, the RDC-i700 can now run using Linux OS. Although some work remains in some areas, kernel support now exists for most of the hardware and features of the camera. This part explains current status for important features (digital camera related) of the kernel.

3.1 SH-Linux

The RDC-i700 linux kernel is originally based on the work of the SH-Linux [1] team. First tested with the kernel version 2.4.2, the camera is now using the version 2.4.19 of the kernel. SH-Linux kernel already had support for almost all parts of the SH3 7709, but since the RDC-i700 is using the CPU in big endian mode, some modifications were necessary in that field. Source code necessary to run the kernel on this new platform has also been added into `/arch/sh/kernel`.

3.2 RDC-i700 device drivers

RDC-i700 drivers can be separated into two kinds (or two layers). (See figure 4) The lower layer contains so-called **Low level drivers**, or drivers providing control to a specific device (such as focus sensor, IPP ...). All those drivers doesn't have any algorithm included and only provide basic access to the device capabilities. For example in case of the driver controlling motors (MECH driver), only functions provided are to set or get the position of the motor (motors have some predefined positions). RDC-i700 currently has 5 device drivers controlling imaging related devices (we will avoid non-imaging specific drivers here):

- **The CCD F/E (Front End)** which permits to control the CCD parameters (such as exposure time, gain ...)
- **The IPP (Image Processing Peripheral)** which is actually the heart of the camera (almost everything goes through the IPP)
- **The Strobe** driver which allows to charge or flash the strobe
- **The Focus Sensor** which permits to evaluate the distance between the camera and the target
- **Mech** driver which controls all mechanical parts of the camera, that is to say, iris, shutter, zoom and focus.

All those drivers are very system dependent and might change from one camera to another.

On the top of those 5 drivers is what we could called the "Algorithms" layer. This layer contains "intelligent" drivers such as auto focus driver, or auto exposure driver. One more driver, simply called CCD driver, is actually the driver which performs actions such as taking a picture or switching to monitoring mode.

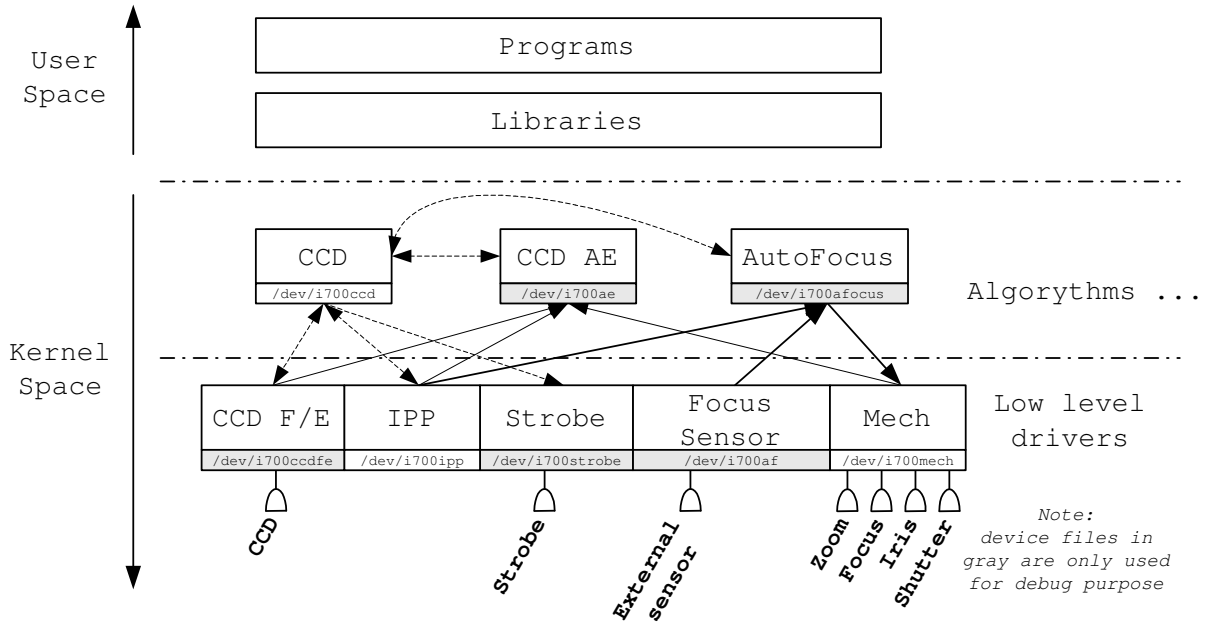


Figure 4: RDC-i700 device drivers

This driver has to access both low level drivers and algorithms drivers. Drivers of the upper layer are "virtually" platform independent. However since the current system lacks of a well defined abstraction layer, upper layer drivers are currently directly accessing lower layer driver which make them unable to work with any other lower driver without having to slightly change the source code. (see Future Work section). Currently device drivers communicate with each others by accessing EXPORTED functions.

All 8 drivers are registered to the kernel as characters drivers and can be accessed from user-level using each device file. Some of those drivers don't actually need to be accessed from user space and in that case device file is only used for debugging purpose. In user space, libraries provide easy access to camera functionalities, avoiding an intensive usage of IOCTL commands.

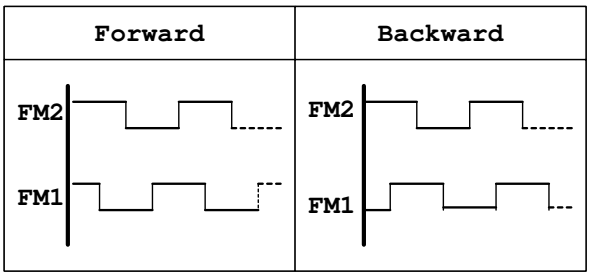


Figure 5: Motors step sequence

3.3 Motors

As the name says, stepping motors are going step by step; the CPU sets 2 I/O ports in order to specify the position of the rotor. Figure 5 shows motor ports state sequence when going forward and backward. Rotation speed is determined by the time between 2 states. Each motor has already predetermined positions, 19 for zoom and 18 for focus; however, if positions for the zoom are fixed, focus position varies depending on the current zoom position. All

those things are handled by the MECH driver and are accessible via IOCTL command such as "Get/Set position." The MECH driver is timer based, that is to say, the delay between 2 steps is performed using a timer (2.8 ms in case of focus, and 1.4 ms in case of zoom); we will see in part 4 the current problems when using this implementation. The driver provides both SYNC and NOSYNC mode; that is to say, in the first one, the ioctl command will hold until the command finish, but in NOSYNC mode, the ioctl will immediately return, allowing to call another ioctl command, even if the motor is still running. This is useful in the case of user's adjustment of the zoom. Since motor needs time to start and stop, it would be inefficient to request each time 1 position change. Instead of this, when the user uses the zoom lever, the first IOCTL command request the motor to go to max position and then when the user release the button, the ioctl STOP command will be requested. In other cases, such as when controlling the FOCUS motor from the auto focus driver, the SYNC mode should be used.

3.4 Auto Exposure

In order to control exposure, the auto exposure driver is accessing 3 different device drivers (IPP, MECH and the CCD front end). The IPP has the ability to divide a CCD image into several block and inform about each block luminance. By looking at those luminance values, the auto exposure algorithm decide how parameters should be modified; it can decide to change iris diameter (MECH driver), or use the strobe (STROBE driver), and in most of the case change parameters of the CCD front end (electronic shutter speed...). The digital camera can work in 2 different modes: the monitoring mode which permits to see in real time what the CCD sensor is targeting, and the still image mode which is used when the user pushes the shutter button to take a still image. The behaviour of the auto exposure module depends on the camera mode:

- **monitoring mode**

in that mode, we adjust CCD F/E parameters every 2 CCD frames. The auto exposure driver starts a kernel thread which needs to be synchronized with the CCD frame (an hardware interrupt is generated by the CCD F/E at every start of frame). Synchronization is achieved by using wait queues. [2] The function which needs to get synchronized creates a wait queue (as follows):

```
struct task_struct *tsk = current;
DECLARE_WAITQUEUE(wait, tsk);
add_wait_queue(&ccd_vd_wq,&wait);
set_current_state(TASK_INTERRUPTIBLE);
schedule();
set_current_state(TASK_RUNNING);
remove_wait_queue(&ccd_vd_wq,&wait);
```

and the wait queue is woken up by the interrupt handler (as follow):

```
wake_up(&ccd_vd_wq);
```

The CCD F/E is controlled using the SCI port of the SH3 which use is shared with some other devices. In some case, it might be necessary to wait for the SCI port availability and, for that reason, the kernel thread implementation has been preferred to some other solutions such as bottom halves (it is not possible to schedule from a bottom halves while kernel thread allows that).

- **still image mode**

in that mode, exposure parameters are only adjusted once before taking the picture. The CCD

[2] This synchronization method is also heavily used by the CCD driver

driver requests information to the auto exposure module which will calculate parameters used to take the picture. After that, the CCD driver will directly control lower driver to set those parameters. Synchronization between all devices is very important and for that reason it is easier to perform everything sequentially from a unique driver. No thread is used and the CCD driver get synchronized with the CCD frame using the same wait queue method as in monitoring mode.

Currently only a very simple algorithm is available for both monitoring and still mode. This algorithm doesn't make use of neither iris nor strobe and the exposure is only controlled using CCD FE's parameters and the mechanical shutter.

3.5 Auto Focus

Compared to the auto exposure, the auto focus driver is quite an easy one. The IPP driver has the ability to determine the "focus level" (the more the focus is correct, the more the value returned by the IPP will be high). In normal mode, the auto-focus driver should get an approximation of the distance to the target by using the focus sensor, then first adjust the focus to this approximation. This permits to perform the "fine focus" (using the IPP capability) to a smaller range. However, the current implementation doesn't use the focus sensor approximation which means that the "fine focus" is performed to the full range of the focus (this is actually the mode which is used in case of MACRO mode). The consequence is that the auto-focus process is much slower than in normal mode. Currently the driver performs the following things:

- check zoom status to calculate focus positions
- retrieve focus level for all focus positions
- go back to the position with the highest focus level

3.6 CCD - IPP

The IPP driver is some kind of library which, except performing initialization of the device, mainly provides a lot of functions, accessible from other drivers and permitting to control the hardware. The driver is quite big since the IPP performs very various things such as

- JPEG compression/decompression
- YUV-RGB conversion
- video output (for the LCD and TV)
- image scaler

The CCD driver is considered as the main driver since almost everything starts from it. It is in charge of coordination between all other drivers. The driver can be controlled using a user land library permitting to control the monitoring mode or to take still image. The driver mainly uses other drivers functions (CCD F/E, IPP, MECH) and performs synchronization using the waitqueue method introduced previously.

3.7 LCD

The RDC-i700 LCD has a fixed resolution of 640x480 pixels. What we could call video card is actually a part of the IPP chip and can control 4 layers of display (1 layer for image/video data, and 3 On Screen Display or OSDs). In the current design the first layer is controlled by the IPP driver and doesn't have direct interface to the user land. Even if 3 OSDs are available, only one is currently used as a framebuffer device. The OSD uses a 8 bit YUV palette (maintained by the IPP device) which

means that `_setcolreg` and `_getcolreg` entry points are used to perform conversion between RGB and YUV color space. This solution allows to use the camera LCD as any traditional Linux console and run any software that usually works on the top of a Linux Framebuffer. One reason why only 1 OSD level is supported is because all OSDs share the same palette which means that it cannot be simply designed as 3 different framebuffer devices. However there is also currently no real need for 3 OSDs so this is not actually a big issue.

3.8 Filesystem

The RDC-i700 has 8 MB of NAND Flash internal memory. The Linux kernel now provides support for this kind of memory by using the Memory Technology Devices (MTD) [2] support. Only a very small layer needs to be written in order to get the camera's NAND work. [3] JFFS2 [4] is usually used on the top of a NAND device, however we will see that in case of digital camera, it might not be the best solution. In our case, internal flash memory is usually exclusively used for storage of compressed data such as JPEG or MPEG. In that case, using JFFS2, which is a compressed filesystem, makes the CPU spend lots of time compressing data which anyway will almost not get compressed more than they are. In such case, YAFFS [5] should be preferred to JFFS2 since it is not a compressed filesystem.(see table 1 for details of tests performed on the RDC-i700)

Time consumption for 1 transaction (secs)		
	YAFFS	JFFS2
JPG (80k)	0.37	0.54
JPG (193k)	0.79	1.32
JPG (547k)	2.21	3.63
MJPG (1463k)	5.6	9.51
*1 transaction = NAND to NAND file copy		

Table 1: YAFFS / JFFS2 tests on RDC-i700

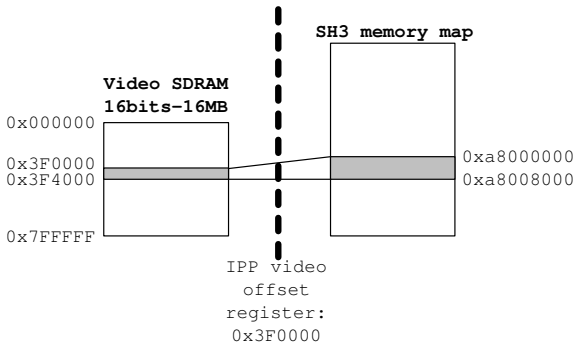


Figure 6: Accessing the video SDRAM

4 Issues

Several problems have been encountered while developing the Linux RDC-i700. Some have been solved but some are still under progress.

4.1 Framebuffer with non-linear memory

The purpose of a framebuffer device is to provide a standard way to access **linearly** video memory from the user space. This is the case of almost (or probably all) video card running on linux. Usually the kernel CPU can directly access any part of the video memory, linearly. However, in case of the RDC-i700, this is not the case. As figure 6 shows, the video SDRAM is not directly accessible from the CPU but is seen through the IPP chip. The IPP provides a 8KB window of memory directly accessible from the SH CPU. By setting a register of the IPP it becomes possible to define which "page" of the SDRAM becomes visible to the CPU. This makes problem with the framebuffer since the video memory is supposed to be linearly accessible, so no method is provided for such kind of system.

Inside a framebuffer device driver, two memory access methods exist:

linux console access mode is performed via function call. 6 functions (read and write for byte, word and long type) are provided. In case of the RDC-i700, the trick is just to overwrite those function in order to set the IPP register to the page needed to be accessed.

mmap access mode is necessary to allow user land application to access the framebuffer memory. The problem in case of mmap access is that the driver doesn't know which part of the memory is being accessed and so it becomes impossible to correctly set the page selector.

To solve this problem, the framebuffer uses a NOPAGE memory handler, combined with the `remap_page_range` function. By always leaving only 1 page mapped at a time, we ensure that the NOPAGE handler will be called everytime the application is trying to access a page which is not mapped. The handler will then unmap the previous page and map the page corresponding to the address to be accessed. One problem remains, which is, if `remap_page_range` allows to map page, it seems there is no function to "unmap" a previously mapped page. The function `zap_page_range` seems to do similar thing and by implementing the nopage handler as follows, the trick seems to work.

```
nopage_handler(...)
{
  calculate pointed addr in SDRAM;
  calculate physical addr;
  if(already_mapped){
    zap_page_range(...);
    flush_tlb_range(...);
  }
  remap_page_range(...);
  already_mapped=1;
}
```

However, some errors occurs time to time when using mmap on the framebuffer and those errors might come from this implementation.

4.2 Timers

We have seen in previous section that several timers are used in various drivers. Some must be very short and for that reason, timer is a very hot topic in our case. Several implementations are possible to perform delay.

busy wait: this solution should be avoided since it would for example almost stop the camera everytime the zoom is adjusted.

kernel timers: kernel timers should be used in order to avoid problems introduced by the busy wait solution. However, in order to achieve such implementation we need first to solve one big problem. While we need about 1ms or less resolution timer, a vanilla 2.4.19 kernel only permits to use timers with resolution of 10ms. In case of stepping motors, this doesn't really make big problem except that motors will just run about 10 times slower than their nominal speed. However, the low accuracy of kernel timers makes problem when used to control other devices such as mechanical shutter, strobe or iris since it can result in low quality image or, even worse, wrong operation performed (narrow instead of large for the iris). In order to solve this problem, several possibilities exist:

- **High Resolution Timer: [7]** is a project hosted on sourceforge in order to add high resolution (nano seconds) capable timer to the Linux kernel. However currently only the i386 architecture is supported, which means that some work is needed in order to get it work on the SH architecture.
- **Hardware Timer:** if such short delay cannot be achieved using software timer,

it still remains the possibility to use hardware timers of the SH3. However, this would make drivers very architecture dependent which should be avoided if possible. Moreover, even if the hardware timer can generate very short period, we need to ensure that the time between the hardware interrupt is generated and the interrupt handler get called is not too long otherwise using hardware timer wouldn't have any meaning. In order to achieve such requirement, it might be necessary to use preemptive kernel.

- **Vanilla kernel with HZ=1000:** changing tick period from 10ms to 1ms allows to use 1ms timer. However, if this solution works fine in some case, we need further experimentation in order to check the accuracy under heavy load condition.

5 Future works

5.1 Design of device driver architecture and user access

Kernel driver layering: the goal is to create a proper abstraction layer permitting to have upper kernel drivers (algorithms) totally independent from the hardware. Currently EXPORTED functions are used to allow function calls between drivers, however it means that upper drivers must understand the behavior of hardware drivers. The abstraction layer needs to define both function prototypes and structures used to access lower driver functionalities. Such interface would permit to easily customize any upper drivers, for example auto exposure algorithm or auto white balance.

Exporting functionalities to user space: currently small libraries are available, permitting to control camera functionalities. However, it doesn't seem reasonable to write new libraries specifically for the camera. Modifying device drivers to make them compatible with some existing standard should be the solution to take advantage of the large amount of existing software. In the camera field, Video For Linux would probably be a good candidate, and especially the second release which is currently under development. The idea would be to provide access to the CCD as a standard video input interface, similar to any USB camera for example. Other functionalities, such as JPEG compression, decompression could be accessed as a CODEC.

5.2 Remaining tasks

Some features still need to be implemented on the camera such as:

Power Management: currently no power management is performed while running Linux on the RDC-i700. This makes the battery life as short as about 25 minutes when using PCMCIA cards. This part should be the next big issue for the linux RDC-i700.

USB controller: the RDC-i700 includes a PDIUSBD12 USB device (slave) controller[3]. The Linux-USB Gadget API [8] allows to easily implement USB device class on the top of controller drivers, however this device controller is currently not supported yet.

6 Conclusion

The linux RDC-i700 has now enough support in order to be used as a digital camera. Most of the constrains due to the architecture and specific hardware have been solved but we still need some more performance testing in order to ensure that everything can run well. But remaining issues are not only technical one. Since we are now preparing for distributing the

[3] http://www.semiconductors.philips.com/pip/PDIUSBD12.html

source codes, we still needs some more coordination in our company. We also have to think of how to make and support a developing community.

References

[1] SH-Linux
`http://linuxsh.sourceforge.net`

[2] Memory Technology Devices (MTD)
`http://www.linux-mtd.infradead.org`

[3] MTD's NAND Flash support
`http://www.linux-mtd.infradead.org/tech/nand.html`

[4] JFFS2 homepage
`http://source.redhat.com/jffs2/`

[5] YAFFS homepage
`http://www.aleph1.co.uk/yaffs/`

[6] Linux Framebuffer Driver Writing HOWTO
`http://linux-fbdev.sourceforge.net/HOWTO/`

[7] High Resolution Timer Project
`http://high-res-timers.sourceforge.net/`

[8] USB-Gadget API
`http://www.linux-usb.org/gadget/`

[9] *Linux Device Drivers*, 2nd Edition
Alessandro Rubini & Jonathan Corbet

`ct_sync`: state replication of `ip_conntrack`

Harald Welte
netfilter core team / Astaro AG / hmw-consulting.de
laforge@gnumonks.org

Abstract

With traditional, stateless firewalling (such as ipfwadm, ipchains) there is no need for special HA support in the firewalling subsystem. As long as all packet filtering rules and routing table entries are configured in exactly the same way, one can use any available tool for IP-Address takeover to accomplish the goal of failing over from one node to the other.

With Linux 2.4/2.6 netfilter/iptables, the Linux firewalling code moves beyond traditional packet filtering. Netfilter provides a modular connection tracking susbsystem which can be employed for stateful firewalling. The connection tracking subsystem gathers information about the state of all current network flows (connections). Packet filtering decisions and NAT information is associated with this state information.

In a high availability scenario, this connection tracking state needs to be replicated from the currently active firewall node to all standby slave firewall nodes. Only when all connection tracking state is replicated, the slave node will have all necessary state information at the time a failover event occurs.

Due to funding by Astaro AG, the netfilter/iptables project now offers a `ct_sync` kernel module for replicating connection tracking state accross multiple nodes. The presentation will cover the architectural design and implementation of the connection tracking failover sytem.

1 Failover of stateless firewalls

There are no special precautions when installing a highly available stateless packet filter. Since there is no state kept, all information needed for filtering is the ruleset and the individual, separate packets.

Building a set of highly available stateless packet filters can thus be achieved by using any traditional means of IP-address takeover, such as Heartbeat or VRRPd.

The only remaining issue is to make sure the firewalling ruleset is exactly the same on both machines. This should be ensured by the firewall administrator every time he updates the ruleset and can be optionally managed by some scripts utilizing scp or rsync.

If this is not applicable, because a very dynamic ruleset is employed, one can build a very easy solution using iptables-supplied tools iptables-save and iptables-restore. The output of iptables-save can be piped over ssh to iptables-restore on a different host.

Limitations

- no state tracking
- not possible in combination with iptables stateful NAT

- no counter consistency of per-rule packet/byte counters

2 Failover of stateful firewalls

Modern firewalls implement state tracking (a.k.a. connection tracking) in order to keep some state about the currently active sessions. The amount of per-connection state kept at the firewall depends on the particular configuration and networking protocols used.

As soon as `any` state is kept at the packet filter, this state information needs to be replicated to the slave/backup nodes within the failover setup.

Since Linux 2.4.x, all relevant state is kept within the *connection tracking subsystem*. In order to understand how this state could possibly be replicated, we need to understand the architecture of this conntrack subsystem.

2.1 Architecture of the Linux Connection Tracking Subsystem

Connection tracking within Linux is implemented as a netfilter module, called `ip_conntrack.o` (`ip_conntrack.ko` in 2.6.x kernels).

Before describing the connection tracking subsystem, we need to describe a couple of definitions and primitives used throughout the conntrack code.

A connection is represented within the conntrack subsystem using `struct ip_conntrack`, also called *connection tracking entry*.

Connection tracking is utilizing *conntrack tuples*, which are tuples consisting of

- source IP address
- source port (or icmp type/code, gre key, ...)
- destination IP address
- destination port
- layer 4 protocol number

A connection is uniquely identified by two tuples: The tuple in the original direction (`IP_CT_DIR_ORIGINAL`) and the tuple for the reply direction (`IP_CT_DIR_REPLY`).

Connection tracking itself does not drop packets[1] or impose any policy. It just associates every packet with a connection tracking entry, which in turn has a particular state. All other kernel code can use this state information[2].

2.1.1 Integration of conntrack with netfilter

If the `ip_conntrack.[k]o` module is registered with netfilter, it attaches to the `NF_IP_PRE_ROUTING`, `NF_IP_POST_ROUTING`, `NF_IP_LOCAL_IN`, and `NF_IP_LOCAL_OUT` hooks.

Because forwarded packets are the most common case on firewalls, I will only describe how connection tracking works for forwarded packets. The two relevant hooks for forwarded packets are `NF_IP_PRE_ROUTING` and `NF_IP_POST_ROUTING`.

Every time a packet arrives at the `NF_IP_PRE_ROUTING` hook, connection tracking creates a conntrack tuple from the packet. It then compares this tuple to the original and re-

[1] well, in some rare cases in combination with NAT it needs to drop. But don't tell anyone, this is secret.

[2] State information is referenced via the `struct sk_buff.nfct` structure member of a packet.

ply tuples of all already-seen connections[3] to find out if this just-arrived packet belongs to any existing connection. If there is no match, a new conntrack table entry (`struct ip_conntrack`) is created.

Let's assume the case where we have already existing connections but are starting from scratch.

The first packet comes in, we derive the tuple from the packet headers, look up the conntrack hash table, don't find any matching entry. As a result, we create a new `struct ip_conntrack`. This `struct ip_conntrack` is filled with all necessarry data, like the original and reply tuple of the connection. How do we know the reply tuple? By inverting the source and destination parts of the original tuple.[4] Please note that this new `struct ip_conntrack` is **not** yet placed into the conntrack hash table.

The packet is now passed on to other callback functions which have registered with a lower priority at `NF_IP_PRE_ROUTING`. It then continues traversal of the network stack as usual, including all respective netfilter hooks.

If the packet survives (i.e., is not dropped by the routing code, network stack, firewall ruleset, ...), it re-appears at `NF_IP_POST_ROUTING`. In this case, we can now safely assume that this packet will be sent off on the outgoing interface, and thus put the connection tracking entry which we created at `NF_IP_PRE_ROUTING` into the conntrack hash table. This process is called *confirming the conntrack*.

The connection tracking code itself is not monolithic, but consists of a couple of separate modules[5]. Besides the conntrack core, there are two important kind of modules: Protocol helpers and application helpers.

Protocol helpers implement the layer-4-protocol specific parts. They currently exist for TCP, UDP, and ICMP (an experimental helper for GRE exists).

2.1.2 TCP connection tracking

As TCP is a connection oriented protocol, it is not very difficult to imagine how conntection tracking for this protocol could work. There are well-defined state transitions possible, and conntrack can decide which state transitions are valid within the TCP specification. In reality it's not all that easy, since we cannot assume that all packets that pass the packet filter actually arrive at the receiving end...

It is noteworthy that the standard connection tracking code does **not** do TCP sequence number and window tracking. A well-maintained patch to add this feature has existed for almost as long as connection tracking itself. It will be integrated with the 2.5.x kernel. The problem with window tracking is its bad interaction with connection pickup. The TCP conntrack code is able to pick up already existing connections, e.g. in case your firewall was rebooted. However, connection pickup is conflicting with TCP window tracking: The TCP window scaling option is only transferred at connection setup time, and we don't know about it in case of pickup...

[3] Of course this is not implemented as a linear search over all existing connections.

[4] So why do we need two tuples, if they can be derived from each other? Wait until we discuss NAT.

[5] They don't actually have to be separate kernel modules; e.g. TCP, UDP, and ICMP tracking modules are all part of the linux kernel module `ip_conntrack.o`.

2.1.3 ICMP tracking

ICMP is not really a connection oriented protocol. So how is it possible to do connection tracking for ICMP?

The ICMP protocol can be split in two groups of messages:

- ICMP error messages, which sort-of belong to a different connection ICMP error messages are associated *RELATED* to a different connection. (`ICMP_DEST_UNREACH`, `ICMP_SOURCE_QUENCH`, `ICMP_TIME_EXCEEDED`, `ICMP_PARAMETERPROB`, `ICMP_REDIRECT`).
- ICMP queries, which have a `request-reply` character. So what the conntrack code does, is let the request have a state of *NEW*, and the reply *ESTABLISHED*. The reply closes the connection immediately. (`ICMP_ECHO`, `ICMP_TIMESTAMP`, `ICMP_INFO_REQUEST`, `ICMP_ADDRESS`)

2.1.4 UDP connection tracking

UDP is designed as a connectionless datagram protocol. But most common protocols using UDP as layer 4 protocol have bi-directional UDP communication. Imagine a DNS query, where the client sends an UDP frame to port 53 of the nameserver, and the nameserver sends back a DNS reply packet from its UDP port 53 to the client.

Netfilter treats this as a connection. The first packet (the DNS request) is assigned a state of *NEW*, because the packet is expected to create a new 'connection.' The DNS server's reply packet is marked as *ESTABLISHED*.

2.1.5 conntrack application helpers

More complex application protocols involving multiple connections need special support by a so-called "conntrack application helper module." Modules in the stock kernel come for FTP, IRC (DCC), TFTP, and Amanda. Netfilter CVS currently contains patches for PPTP, H.323, Eggdrop botnet, mms, DirectX, RTSP, and talk/ntalk. We're still lacking a lot of protocols (e.g. SIP, SMB/CIFS)—but they are unlikely to appear until somebody really needs them and either develops them on his own or funds development.

2.1.6 Integration of connection tracking with iptables

As stated earlier, conntrack doesn't impose any policy on packets. It just determines the relation of a packet to already existing connections. To base packet filtering decision on this state information, the iptables *state* match can be used. Every packet is within one of the following categories:

- **NEW**: packet would create a new connection, if it survives
- **ESTABLISHED**: packet is part of an already established connection (either direction)
- **RELATED**: packet is in some way related to an already established connection, e.g. ICMP errors or FTP data sessions
- **INVALID**: conntrack is unable to derive conntrack information from this packet. Please note that all multicast or broadcast packets fall in this category.

2.2 Poor man's conntrack failover

When thinking about failover of stateful firewalls, one usually thinks about replication of state. This presumes that the state is gathered at one firewalling node (the currently active node), and replicated to several other passive standby nodes. There is, however, a very different approach to replication: concurrent state tracking on all firewalling nodes.

While this scheme has not been implemented within `ct_sync`, the author still thinks it is worth an explanation in this paper.

The basic assumption of this approach is: In a setup where all firewalling nodes receive exactly the same traffic, all nodes will deduct the same state information.

The implementability of this approach is totally dependent on fulfillment of this assumption.

- *All packets need to be seen by all nodes*. This is not always true, but can be achieved by using shared media like traditional ethernet (no switches!!) and promiscuous mode on all ethernet interfaces.
- *All nodes need to be able to process all packets*. This cannot be universally guaranteed. Even if the hardware (CPU, RAM, Chipset, NICs) and software (Linux kernel) are exactly the same, they might behave different, especially under high load. To avoid those effects, the hardware should be able to deal with way more traffic than seen during operation. Also, there should be no userspace processes (like proxies, etc.) running on the firewalling nodes at all. WARNING: Nobody guarantees this behaviour. However, the poor man is usually not interested in scientific proof but in usability in his particular practical setup.

However, even if those conditions are fulfilled, there are remaining issues:

- *No resynchronization after reboot*. If a node is rebooted (because of a hardware fault, software bug, software update, etc.) it will lose all state information until the event of the reboot. This means, the state information of this node after reboot will not contain any old state, gathered before the reboot. The effects depend on the traffic. Generally, it is only assured that state information about all connections initiated after the reboot will be present. If there are short-lived connections (like http), the state information on the just rebooted node will approximate the state information of an older node. Only after all sessions active at the time of reboot have terminated, state information is guaranteed to be resynchronized.
- *Only possible with shared medium*. The practical implication is that no switched ethernet (and thus no full duplex) can be used.

The major advantage of the poor man's approach is implementation simplicity. No state transfer mechanism needs to be developed. Only very little changes to the existing conntrack code would be needed in order to be able to do tracking based on packets received from promiscuous interfaces. The active node would have packet forwarding turned on, the passive nodes, off.

I'm not proposing this as a real solution to the failover problem. It's hackish, buggy, and likely to break very easily. But considering it can be implemented in very little programming

time, it could be an option for very small installations with low reliability criteria.

2.3 Conntrack state replication

The preferred solution to the failover problem is, without any doubt, replication of the connection tracking state.

The proposed conntrack state replication soltution consists of several parts:

- A connection tracking state replication protocol
- An event interface generating event messages as soon as state information changes on the active node
- An interface for explicit generation of connection tracking table entries on the standby slaves
- Some code (preferrably a kernel thread) running on the active node, receiving state updates by the event interface and generating conntrack state replication protocol messages
- Some code (preferrably a kernel thread) running on the slave node(s), receiving conntrack state replication protocol messages and updating the local conntrack table accordingly

Flow of events in chronological order:

- *on active node, inside the network RX softirq*
  - `ip_conntrack` analyzes a forwarded packet
  - `ip_conntrack` gathers some new state information
  - `ip_conntrack` updates conntrack hash table
  - `ip_conntrack` calls event API
  - function registered to event API builds and enqueues message to send ring
- *on active node, inside the conntrack-sync sender kernel thread*
  - `ct_sync_send` aggregates multiple messages into one packet
  - `ct_sync_send` dequeues packet from ring
  - `ct_sync_send` sends packet via in-kernel sockets API
- *on slave node(s), inside network RX softirq*
  - `ip_conntrack` ignores packets coming from the `ct_sync` interface via NOTRACK mechanism
  - UDP stack appends packet to socket receive queue of `ct_sync_recv` kernel thread
- *on slave node(s), inside conntrack-sync receive kernel thread*
  - `ct_sync_recv` thread receives state replication packet
  - `ct_sync_recv` thread parses packet into individual messages
  - `ct_sync_recv` thread creates/updates local `ip_conntrack` entry

2.3.1 Connection tracking state replication protocol

In order to be able to replicate the state between two or more firewalls, a state replication protocol is needed. This protocol is used

over a private network segment shared by all nodes for state replication. It is designed to work over IP unicast and IP multicast transport. IP unicast will be used for direct point-to-point communication between one active firewall and one standby firewall. IP multicast will be used when the state needs to be replicated to more than one standby firewall.

The principal design criteria of this protocol are:

- **reliable against data loss**, as the underlying UDP layer only provides checksumming against data corruption, but doesn't employ any means against data loss
- **lightweight**, since generating the state update messages is already a very expensive process for the sender, eating additional CPU, memory, and IO bandwith.
- **easy to parse**, to minimize overhead at the receiver(s)

The protocol does not employ any security mechanism like encryption, authentication, or reliability against spoofing attacks. It is assumed that the private conntrack sync network is a secure communications channel, not accessible to any malicious third party.

To achieve the reliability against data loss, an easy sequence numbering scheme is used. All protocol messages are prefixed by a sequence number, determined by the sender. If the slave detects packet loss by discontinuous sequence numbers, it can request the retransmission of the missing packets by stating the missing sequence number(s). Since there is no acknowledgement for sucessfully received packets, the sender has to keep a reasonably-sized[6] backlog of recently-sent packets in order to be able to fulfill retransmission requests.

[6] *reasonable size* must be large enough for the round-trip time between master and slowest slave.

The different state replication protocol packet types are:

- **`CT_SYNC_PKT_MASTER_ANNOUNCE`**: A new master announces itself. Any still existing master will downgrade itself to slave upon reception of this packet.
- **`CT_SYNC_PKT_SLAVE_INITSYNC`**: A slave requests initial synchronization from the master (after reboot or loss of sync).
- **`CT_SYNC_PKT_SYNC`**: A packet containing synchronization data from master to slaves
- **`CT_SYNC_PKT_NACK`**: A slave indicates packet loss of a particular sequence number

The messages within a `CT_SYNC_PKT_SYNC` packet always refer to a particular *resource* (currently `CT_SYNC_RES_CONNTRACK` and `CT_SYNC_RES_EXPECT`, although support for the latter has not been fully implemented yet).

For every resource, there are several message types. So far, only `CT_SYNC_MSG_UPDATE` and `CT_SYNC_MSG_DELETE` have been implemented. This means a new connection as well as state changes to an existing connection will always be encapsulated in a `CT_SYNC_MSG_UDPATE` message and therefore contain the full conntrack entry.

To uniquely identify (and later reference) a conntrack entry, the only unique criteria is used: `ip_conntrack_tuple`.

2.3.2 `ct_sync` sender thread

Maximum care needs to be taken for the implementation of the ctsyncd sender.

The normal workload of the active firewall node is likely to be already very high, so generating and sending the conntrack state replication messages needs to be highly efficient.

It was therefore decided to use a pre-allocated ringbuffer for outbound `ct_sync` packets. New messages are appended to individual buffers in this ring, and pointers into this ring are passed to the in-kernel sockets API to ensure a minimum number of copies and memory allocations.

2.3.3 `ct_sync` initsync sender thread

In order to facilitate ongoing state synchronization at the same time as responding to initial sync requests of an individual slave, the sender has a separate kernel thread for initial state synchronization (and `ct_sync_initsync`).

At the moment it iterates over the state table and transmits packets with a fixed rate of about 1000 packets per second, resulting in about 4000 connections per second, averaging to about 1.5 Mbps of bandwith consumed.

The speed of this initial sync should be configurable by the system administrator, especially since there is no flow control mechanism, and the slave node(s) will have to deal with the packets or otherwise lose sync again.

This is certainly an area of future improvement and development—but first we want to see practical problems with this primitive scheme.

2.3.4 `ct_sync` receiver thread

Implementation of the receiver is very straightforward.

For performance reasons, and to facilitate code-reuse, the receiver uses the same pre-allocated ring buffer structure as the sender. Incoming packets are written into ring members and then successively parsed into their individual messages.

Apart from dealing with lost packets, it just needs to call the respective conntrack add/modify/delete functions.

2.3.5 Necessary changes within netfilter conntrack core

To be able to achieve the described conntrack state replication mechanism, the following changes to the conntrack core were implemented:

- Ability to exclude certain packets from being tracked. This was a long-wanted feature on the TODO list of the netfilter project and is implemented by having a “raw” table in combination with a “NOTRACK” target.
- Ability to register callback functions to be called every time a new conntrack entry is created or an existing entry modified. This is part of the nfnetlink-ctnetlink patch, since the ctnetlink event interface also uses this API.
- Export an API to externally add, modify, and remove conntrack entries.

Since the number of changes is very low, their inclusion into the mainline kernel is not a problem and can happen during the 2.6.x stable kernel series.

2.3.6 Layer 2 dropping and `ct_sync`

In most cases, netfilter/iptables-based firewalls will not only function as packet filter but also

run local processes such as proxies, dns relays, smtp relays, etc.

In order to minimize failover time, it is helpful if the full startup and configuration of all network interfaces and all of those userspace processes can happen at system bootup time rather then in the instance of a failover.

l2drop provides a convenient way for this goal: It hooks into layer 2 netfilter hooks (immediately attached to `netif_rx()` and `dev_queue_xmit`) and blocks all incoming and outgoing network packets at this very low layer. Even kernel-generated messages such as ARP replies, IPv6 neighbour discovery, IGMP, ... are blocked this way.

Of course there has to be an exemption for the state synchronization messages themselves. In order to still facilitate remote administration via SSH and other communication between the cluster nodes, the whole network interface used for synchronization is subject to this exemption from l2drop.

As soon as a node is propagated to master state, l2drop is disabled and the system becomes visible to the network.

2.3.7 Configuration

All configuration happens via module parameters.

- `syncdev`: Name of the multicast-capable network device used for state synchronization among the nodes
- `state`: Initial state of the node (0=slave, 1=master)
- `id`: Unique Node ID (0..255)
- `l2drop`: Enable (1) or disable (0) the l2drop functionality

2.3.8 Interfacing with the cluster manager

As indicated in the beginning of this paper, `ct_sync` itself does not provide any mechanism to determine outage of the master node within a cluster. This job is left to a cluster manager software running in userspace.

Once an outage of the master is detected, the cluster manager needs to elect one of the remaining (slave) nodes to become new master. On this elected node, the cluster manager will write the ascii character `1` into the `/proc/net/ct_sync` file. Reading from this file will return the current state of the local node.

3 Acknowledgements

The author would like to thank his fellow netfilter developers for their help. Particularly important to `ct_sync` is Krisztian KOVACS `<hidden@balabit.hu>`, who did a proof-of-concept implementation based on my first paper on `ct_sync` at OLS2002.

Without the financial support of Astaro AG, I would not have been able to spend any time on `ct_sync` at all.

Increasing the Appeal of Open Source Projects

Experiences from the LSB Project

Mats Wichmann
Intel Corporation / LSB Project
mats.d.wichmann@intel.com

Abstract

It is often said that open source projects will "win" or "lose" based purely on technical merit. Experiences from the LSB Project's interface standardization efforts indicate there are some concrete steps an open-source project producing interface libraries for general use can take to make the project more usable for a wider audience, leading to greater chance of widespread acceptance. Such projects have a reasonable chance of becoming standards, whether de-facto or by inclusion in formal specifications such as the LSB.

The evidence is that projects ready for large-scale use typically meet most of a set of criteria that include: demand; stable, well-documented interfaces; comprehensive interface and regression tests; an easily-deployed (portable) working implementation; and an appropriate choice of license. With the exception of demand, most of these criteria can be consciously worked towards. The paper will present some case studies of libraries that have successfully been incorporated into the LSB specification. It will also discuss some tools the LSB has developed that may help in describing public interfaces and developing tests, and discuss some ways in which portability of the code base can be improved.

1 Introduction

The Free Software and Open Source Software models present some unique concepts which seem to work best when the software is widely used and there's an active feedback loop to debug and improve the software. In order for this to be possible, it's important that some core requirements that apply to all software are attended to in this space as well: consistency and compatibility, documentation, and ease of use. If the software is to hard to deploy or make use of, the user base will remain small and the synergy which is so important to these projects will be harder to achieve.

While ease of use is a concept that is hard to measure for the developer as it means different things to different users, for an individual user it's pretty easy to tell when an application or library is not easy enough to use—it's painful to install, get running, or program to, making it hard to use it to solve the problem at hand. Where money did not change hands to obtain the software, the likely response will be to give up and look for a different solution, while what we as developers would rather have is feedback about the problems and suggestions for improvement. Often lacking a "marketing department" to drive requirements (whatever one may think of such a situation), this feedback is crucial to the open source process.

The Linux Standard Base (LSB) project

(`http://www.linuxbase.org`) aims to drive the creation of a consistent runtime environment for applications. Drawing from the experiences of the LSB project, we will examine a pair of issues, one on either side of the "runtime environment" boundary: building better libraries, and making applications (and libraries) easier to deploy.

2 Building Better Libraries

Libraries are an effective mechanism to provide for code reuse.

In Linux, libraries are normally provided as shared objects, although they may also be provided as static archives. Some libraries are foundational in that they are expected to be used by a broad variety of applications, such as the GNU C library, which is used by all programs; or the GNOME glib, which is used (directly or indirectly) by all graphical applications written to GNOME. Other libraries may export a programming interface specific to one application family such as libMagick for ImageMagick.

If a project produces libraries which are to be usable by others there are some particular issues that apply.

2.1 Stable Interfaces

A library provides certain programming interfaces which are available to programs to use (external), and probably also contains interfaces which are not intended to be used outside the library (internal). The set of external interfaces provides the Application Programming Interface (API). As programmers become familiar with the library, they will want the API to provide some stability so that they don't always have to recode their programs when the library is revised.

When a program is linked with a shared library, it will contain references to library interfaces which are resolved at runtime by the dynamic linker. The runtime instantiation of the library interface set provides the Application Binary Interface (ABI), and programmers will want the ABI to remain stable as well, or their programs may work incorrectly run against a different version of the shared library then it was originally linked against.

The dilemma for the library developer is that it's hard to get it (completely) right the first time. Bugs will be found, often the design will be found to be limiting or even incorrect, or the library may simply need to evolve to meet new needs. It would be terribly limiting to never be able to evolve the library just because users and developers demand stability. Fortunately, there are some techniques that can be used to make life a little easier.

A useful step is to identify the intended API and make sure that is all the library exports to programmers. If the API is designed as an abstraction layer distinct from the internal implementation, considerable freedom will be available to modify the library "under the covers" while still keeping the ABI stable. It's worth taking the time to design the API in this manner. It is also very useful if programmers cannot reach the internal routines which may need to change—experience has shown that if an interface can be found, someone will find a way to use it. A linker script can be used to export the desired symbols, hiding the others:

```
{
    global:
        lsbfoo;
    local:
        foo*;
};
```

A linker script is used when build-

ing a shared library by including the `-version-script=`*scriptname* directive in the gcc link line.

It's quite possible, however, that some interface in the ABI will need to change in an incompatible way. To provide for this, the symbols making up the ABI can be assigned versions, leaving the possibility of changing the version. The following example shows the use of a linker script which exports two routines and assigns them version `LSBLIB_1.0`:

```
LSBLIB_1.0 {
    global:
        lsbfoo;
        lsbbar;
    local:
        foo*;
};
```

If the symbol version is changed, old binaries won't run against the new library as the symbol version in those binaries will not be found; while binaries compiled against the new library will pick up the new symbol version. It is also possible—and may be desirable—to provide both the old and the new version of the interface in the newer library, this way old binaries can continue to run, while new binaries will be linked against the newer version of the interface by default, but could also be explicitly linked against the old version. The following example shows creating a new symbol version set which is inclusive of the previous one, only the `lsbfoo` interface will get the version tag `LSBLIB_1.1`.

```
LSBLIB_1.0 {
    lsbfoo;
    lsbbar;
};
LSBLIB_1.1 {
    lsbfoo;
} LSBLIB_1.0;
```

To make this work, the GNU linker is needed, and some special directives (`__asm__(".symver realname, alias, version");` are needed in the code, so that the old routine can be bound to the old version and the new code to the new version. The GNU linker documentation has more details on this.

If a lot of interfaces need to change incompatibly, it is better to change the major version of the library. The library version will be bound into binaries compiled against it. With major changes, multiple versions of the library can be provided, giving compatibility for old and new code.

In the LSB project, symbol versioning is used for those libraries which are already normally built that way, essentially the GNU libc set. Adding symbol versioning is a nice way to avoid breaking compatibility if a small number of interfaces have to be changed in incompatible ways. The LSB specification calls out specific library versions which must be provided by a conforming runtime, and where the symbols are versioned, the specific symbol versions. As conforming runtimes may have evolved the interfaces in the manner described, a trick is used for linking LSB conforming applications: a set of stub libraries has been constructed which contains only the LSB interfaces, with the versions required by the spec, and these are used for link-time symbol resolution.

2.2 API Documentation

A factor in how useful a library is is the quality of api documentation The documentation must describe in detail the programming interfaces available, with function calling and return conventions, boundaries, and error con-

ditions. This is the kind of information traditionally captured in the "manpage." The best measure of the quality of API documentation seems to be whether assertion-based tests (see next section) can be developed completely from the documentation, or whether the source code must be referred to fill in the details.

It is especially useful to use a tool to automate a part of this process. There are a number of tools that understand how to produce documentation from commented source code, one example would be *doxygen* (`http://www.doxygen.org`) although documentation generators seem to be more commonly used with higher-level languages (e.g. Javadoc for Java, Pydoc for Python, etc.)

The advantages of a generator approach is that the interface descriptions in the documentation don't depend on human transcription to get them right in the first place, and then don't go out of skew if the interfaces in the code ever change. It's particularly galling to try to code to an interface that does not work as documented.

The LSB specification has to date included mostly libraries which are already standardized at the API level—for example, the GNU C library is designed to be compatible with POSIX specification, so the LSB specification for the C library is able to reference this existing specification for almost all of the functional descriptions. As the LSB seeks to expand the base to other important libraries found on Linux systems, the API documentation will have to be imported by copy or by reference into the specification, so the existence of such documentation has become an LSB selection criteria.

The LSB itself has a slightly different documentation problem, as it has to capture an ABI description to describe the binary interface programs will see. A single API prototype or structure definition has been captured the way it will be seen on each of the (currently seven) architectures the LSB supports, based on things like data model (sizes of integers and pointers, for example). The symbol versions matching the interfaces must also be captured. All of this information is represented in a MySQL database which is browsable on the web (`http://www.linuxbase.org/dbadmin`) but which is also used to generate LSB header files, the stub libraries mentioned in the previous section, and the portion of the LSB specification that contains library listings, interface listings, and data definitions.

The database is also used to generate test code. Of particular note, the LSB generates two test programs, one to test the presence of the libraries and interfaces on a runtime, and another to test that an application uses only the libraries and interfaces in the specification. The data for these two programs is generated directly out of the specification database.

The LSB database schema and tools to extract data and build code (essentially a set of Perl scripts) are freely available for use by other projects, although they are probably mostly applicable to projects that support a large number of libraries and want to build similar test tools. They can be browsed from the LSB CVS tree (`cvs.gforge.freestandards.org`).

The summary is that while there's no magic to producing good documentation, it's important in producing a stable library that can be widely used. It's worth the time to see if some level of automation can help with the tasks, particularly if there are several areas that need to be kept in sync.

2.3 Interface Tests

Another area for consideration is detailed interface testing. Good tests allow checking

that interfaces perform as intended. The POSIX testing standard calls for such tests to be assertion-based, which means a written description of an intended behavior is produced, this is then used to develop the test case. The following example of an assertion is taken from the Open POSIX Test Suite (`http://sourceforge.net/projects/posixtest`:

mmap assertion 9 When MAP_FIXED is set in the flags argument, the implementation is informed that the value of pa shall be addr, exactly. If MAP_FIXED is set, mmap( ) may return MAP_FAILED and set errno to [EINVAL]. If a MAP_FIXED request is successful, the mapping established by mmap( ) replaces any previous mappings for the process' pages in the range [pa,pa+len].

Tests intended to operate at the source code level can be built and executed as part of the product build and are an effective way to catch regressions introduced during regular maintenance and development activity.

Binary level tests operate against an already built library, and are a way to test that a particular library is compatible with a particular API definition. Such tests increase the confidence of developers in the stability of the library.

In the LSB project, interface testing is the most important way of measuring a runtime against the LSB specification. However, the process of writing assertions and developing tests is not easy. It depends on a quality interface specification, good choice of testing methodologies, etc. There is little doubt that the most effective place for this work to take place is within the project itself. The source code file describing an interface can contain the interface, documentation, test assertions, and test code. All can be developed together without the kind of extra overhead incurred if each of the four items is developed separately by separate persons. The author is not aware of an existing toolkit which could automatically generate all of the necessary pieces from a single source file so endowed, but this would certainly make an interesting open source project of its own!

2.4 License Choice

The choice of license under which to release a library makes a considerable difference in who can use the libraries and how. This paper does not attempt a license recommendation as only the developer can know their own targets, needs and desires, which will guide the choice of license.

A Free Software license along the lines of the well-known GPL effectively restricts usage to programs under the same or compatible licenses. Such code cannot be used in closed source programs, even through dynamic linking, and also cannot be used by code under certain open source licenses that are not considered compatible, perhaps because they place some restriction on the user (one example might be a license that restricts usage to academic or personal use and disallows commercial use). The related LGPL license allows the use of the library by code of any sort through dynamic linking, but makes no similar provision for static linking. There are a variety of other licenses which grant greater or lesser freedoms in the ways the code may be used.

Some applications release code under dual licenses, for example a GPL-like license for those who can use it, and a separate license with commercial terms for those who cannot. It is also possible to release a package consisting of program code and library code with separate licenses for each.

As noted above, some licenses have compatibility clauses relating to how to code may be mingled with code under certain other licenses.

Various potential users of the code may have their own selection criteria that includes license choice. For example, the Debian project has a particular definition of "free" and consigns code which does not meet these criteria to the "nonfree" area.

Continuing with the use of LSB project experiences to illustrate, the LSB is concerned with functional interface descriptions, not with specific implementations. So the license of an *implementation* is not crucial—unless it's effectively the only implementation available, in which case it becomes a determining factor in practical use of the interface set.

An example may help clarify: the popular Qt toolkit was for a while the subject of some controversy in the open source community over its license terms, and a project was started to create an open source reimplementation of the Qt interface specification. When Qt licensing was changed to a dual license (one GPL-like, with a separate license for commercial developers) the open source reimplementation project was dropped as the problem people had with the previous license was resolved. However, the LSB project favors a "no strings attached" selection policy which suggests *against* the inclusion of a library where the only implementation doesn't allow a certain class of developers to just make use of the library in their code without arranging a commercial license.

The upshot is that choice of license needs to be considered very carefully.

3 Software Packaging and Deployment

The other major consideration this paper will examine is improving the accessibility of the software through producing a package that is easy to put into use. This discussion applies to both libraries and to complete applications.

The most common way to install software on Linux must be to install a distribution-specific package that has already been prepared. This has many advantages, as it's configured, compiled, and tested for that distribution, and the package will be tagged with dependencies so the user can determine what else needs to be installed to make it work. It will normally have security update patches made available should such become necessary.

Of course, not every package can be chosen for distribution packaging, and it's quite possible that an interested user for your software may find that a package is not available at all, or just not available for her distribution of choice. This should pose no problem since by definition the source code is available, and the software can simply be built from source. Unfortunately, in many cases the *simply* is a misnomer since there may be dependencies on other software, toolchain versions, etc. that may prove to be impediments.

3.1 How Not to Install Software

Although probably everyone reading this paper has had some negative experiences of their own with software installation, by way of example here is a condensed version of a situation that befell the author, and indirectly provided the motivation for recording these thoughts here:

At one point, I became interested in doing some transpositions on a piece of music, and I thought there must be a piece of software that would help with this. There are certainly commercial PC-centric applications that do this very well but there must be something open source as well. Some searching turned up a promising application named *noteedit*. Surprisingly, **rpmfind** told me that the one distribution for which a current version was pack-

aged was Mandrake, luckily my distribution of choice. The package did indicate Cooker, which is Mandrake's early-access build tree, but since it was only a couple of weeks after that last release, I assumed the Cooker could not have migrated too far and it would probably work.

After obtaining and installing the package, plus an attendant library package as well as another library (libtse) also needed, I installed and tried to run the package. Alas, it had been linked against a different C++ library version and so had references to some symbols that were not in my C++ library and thus was not runnable.

My next effort was to download the noteedit and tse library tarballs and attempt to build them from source. This was not a great success either, as the configuration scripts kept reporting fatal problems due to missing build headers and libraries, of course I had to correlate these back to the packages they would be installed by and install those. After several cycles I abandoned this approach and went to the third try, going back to rpmfind and pulling down the source, rather than binary, rpms and trying to build from source that way. This ultimately yielded a runnable binary although not without some further pain which involved tweaking the rpm specfiles. And this success still came because some Mandrake user contributed a build to the Cooker, which although it was for the wrong version (from my point of view) could be adjusted at the source level to work. What if I were running something different?

3.2 Binary Software Distribution

A project can certainly make their software easier to check out if there's a binary package available. Even if packaged by some distributions (and for many projects even this does not happen, especially early on), there's still the question of reaching users of other distributions.

The difficulty with a project building binary packages is deciding what to build for: there are an endless number of combinations of distributions and versions, and only a small fraction could be targeted. Further, this potentially puts the project into a "distro support" mode, that is worrying about oddities on the particular distro/version they have chosen to build for. A better solution seems to be to build a portable (distro-neutral) version.

Producing a portable binary package as an example has many advantages for a project:

- One package works on multiple kinds of systems
- Users interested in the software can get it running quickly
- Bugreports don't have to worry about the user's build environment
- Bugreports will be against a known set of configure and build options

There's still plenty of use for users building from source as well, including trying out combinations the developers have not tried, but the opportunity to come up quickly should broaden the base of potential users since not everybody wants to go through building from source.

Of course a really good build procedure from source—which clearly identifies dependencies, is also very valuable. Configure scripts have the unfortunate habit of quitting on the first "fatal error," which means after you satisfy that build dependency you try again and occasionally run into another, and then another. In frustration, the author once coded a configure script which issues warnings (AC_MSG_WARN) instead of errors (AC_MSG_FAIL), setting a flag which

is used to signal a fatal error at the end of the script. The author is not sure this hack is a "really good build procedure" however!

3.3 Using the LSB to Build Binary Packages

If a portable binary package is a target, the LSB provides a good model. The LSB specification describes a runtime platform, and also describes some things about how the package is delivered.

To build a portable binary, a relatively short set of rules needs to be followed:

- Link with the LSB runtime linker
- Use only LSB-specified libraries with the correct version
- Use only LSB-specified interfaces and symbol versions from those libraries
- All other interfaces must be supplied with the application

The runtime linker has a distinct name for LSB programs. For example, on the IA32 architecture, `ld-lsb.so.1` is used instead of `ld-linux.so.2`. This allows an implementation to do something different for LSB programs, such as resolving against libraries in a different directory. This capability is rarely used: most runtimes simply make the LSB linker name a symbolic link to the regular linker.

An application may only count on LSB libraries to be present on a conforming runtime, thus the restriction to link only with those libraries. If other libraries are needed, they can be statically linked, or provided in an application-supplied shared library. It is also possible to depend on *another* LSB-conforming package which supplies a shared library. Any such libraries must be constructed LSB conforming, which in practice means they need to watch their own dependencies on other libraries.

Some libraries may have more public interfaces than are described in the LSB specification. The most notable example is GNU libc. Even though these interfaces are likely to be present on every conforming system's version of those libraries, this is not required by the specification, and thus a conforming runtime may not count on them. For libraries which are symbol versioned, the binary must be linked against the symbol versions described in the specification.

While these rules are not terribly complex, it would be painful to modify build trees with many makefiles to apply them, so the LSB project supplies a compiler wrapper program `lsbcc` (as well as `lsbc++` for C++ programs) which applies the rules by fiddling with the compiler line before handing it off to the regular compiler, usually `gcc`.

If we get lucky, an LSB build can be as simple as:

```
CC=lsbcc ./configure
make
```

Of course it's not always this easy, and usually the problem is the use of libraries which are not in the LSB. The wrapper will actually turn references to non-LSB libraries into static links (the tool can be told to warn about this behavior as it's often useful to know what's happening behind your back). Sometimes static linking is a reasonable solution, sometimes packaging up the missing library in LSB mode is workable, and sometimes nothing will help but to lobby the LSB project to add the library—which will undoubtedly result in a polite request for help! The LSB still has quite a bit of evolving to do

and it's hoped that exposing it here will help identify the features which need to be added to future versions.

The other helpful aspect the LSB covers has to do with delivery of the software. Again, there are several areas:

- Portable format for the package
- Rules for where the package may place files
- Rules about names of packages to avoid clashes
- Special features such an an installer for startup scripts

The package format called out in the LSB specification is that used by the rpm package manager. This is a relatively portable format in that tools such as `alien` can convert these packages into other formats which can be handled by a system's package manager. There's no requirement that a runtime be rpm-based itself, and the only thing a package needs to (or is allowed to) depend on are provides for LSB modules (currently `lsb-core` and `lsb-graphics`) or other LSB packages.

It's also possible to deliver a package in other formats; in this case the rule is that the installer must be an LSB-conforming binary or an LSB-required command. A combination of a shell script and a tarball actually meet this requirement as both commands are required by the LSB specification. The use of other than the LSB package format is discouraged, however, as it makes it hard for system administrators to keep a view of what has been installed as would be the case if all software used the same package manager.

The File Hierarchy Standard (FHS) is imported into the LSB by reference and describes where an application may place files. To state these rules imprecisely, the package name serves as a tag, and it may install files into `/opt/`*tag*, `/etc/opt/`*tag*, and `/var/opt/`*tag*. This avoids clashes with distribution-provided packages and locally added software.

The naming of the package is also described by the LSB; essentially the rule is to register either a single package name, or a provider name, with the Linux Assigned Names and Numbers Authority or LANANA (`http://www.lanana.org`).

Finally, there are some provisions for things which don't fit into the above picture. For example, startup ("init") scripts and cron entries have to go in specific places. The LSB describes a special installer which may be invoked to create the links in the `/etc/rcX.d` directories.

With the specified behavior and tools, the LSB makes possible the creation of portable binary packages.

4 Summary

There are many considerations towards making software projects more popular. This paper has concentrated on only a small portion of those.

We have examined some issues towards making shared libraries useful. The assertion is that as a library becomes more Standard, whether that be a self-published standard or one promoted by a larger group or even a standards organization, it becomes easier for a wider audience to depend on it, software that uses it can be free of compatibility fears, and the larger community will lead to more and better feedback to continue to improve. Some steps that could help move a project towards such a state include developing solid interface specifications; stabilizing the interfaces as seen by

software through versioning, which leaves the freedom to continue to innovate while provide backward compatibility; and through comprehensive interface tests. We also looked at how choice of license plays into the usability of a library.

Another consideration towards usable software is lending the ability for potential users to get "on the air" with the software quickly, so they can evaluate it and see if it suits their needs without going through a lot of trouble. To that end, we looked at some benefits of projects delivering binary package in addition to source packages. The components of the LSB project which help in producing portable binary packages were also covered, to show how a project might be able to build a single binary package which helps the software become more accessible.

5 Disclaimer

The opinions expressed in this paper are those of the author and do not necessarily represent the position of Intel Corporation.

Repository-based System Management Using Conary

Michael K. Johnson, Erik W. Troan, Matthew S. Wilson
Specifix, Inc.
ols2004@specifixinc.com

Abstract

There have been few advances in software packaging systems since the creation of dpkg and RPM. Conary is being developed to provide a fresh approach to Open Source Software management and provisioning, one that applies new ideas from distributed software version control tools such as GNU arch and Monotone. Rather than concentrating on package files, Conary provides an architecture built around distributed repositories and change sets, and includes features designed to make branching and tracking Linux distributions simple operations.

The rise of distributions such as Fedora and Gentoo has moved the development of Linux distributions from small, tightly-connected groups to widely-dispersed groups of informal collaborators. These changes have brought to light many shortcomings of the dominant packaging metaphor. By providing version trees distributed across Internet-based software repositories, Conary allows these casual groupings of contributors to work together much more effectively than they can today.

1 Packaging Limitations

Traditional package management systems (such as RPM and dpkg) provided a major improvement over the previous regime of installing from source or binary tar archives. However, they suffer from a few shortcomings, and some of these shortcomings are felt more acutely as the Internet and the Open Source communities have developed and expanded. The authors' experience with the shortcomings of current package management systems strongly motivated Conary's design.

1.1 Branching

Traditional package management systems use simple version numbers to allow the different package versions to be sorted into "older" and "newer" packages, adding concepts such as **epochs** to work around version numbers that do not follow the packaging system's ideas of how they are ordered. While the concepts of "newer" and "older" seem simple, they break down when multiple streams of development are maintained simultaneously using the package model. For example, a single version of a set of sources can yield different binary packages for different versions of a Linux distribution. A simple linear sorting of version numbers cannot represent this situation, as neither of those binary packages is newer than the other; the packages simply apply to different contexts.

1.2 Package Repository Limitations

Traditional package management systems provide no facilities for coordinating work between independent repositories.

- Repositories have version clashes; the same version-release string means different things in different repositories. Repositories can even have name clashes—the same name in two different repositories might not mean the same thing.
- There is no way to identify which distribution, let alone which version of the distribution, a package is intended and built for.

For example, is the `aalib-1.4.0-5.1fc2.fr` package newer than the `aalib-1.4.0-0-fdr.0.8.rc5.2` package? One is from the freshrpms repository, and the other is from the fedora.us repository. Which package should users apply to their systems? Does it depend on which version of which distribution they have? How are the two packages related? Are they related at all?

This is not really a problem in a disconnected world. However, when you install packages from multiple sources, it can be hard to tell how to update them—or even what it means to update a package. You have to rely on your memory of where you fetched a package from in order even to look in the right repository. Once you look there, it is not necessarily obvious which packages are intended for the particular version of the distribution you have installed. Automated tools for fetching packages from multiple repositories have increased the number of independent package repositories over the past few years, making the confusion more and more evident.

The automated tools helped exacerbate this problem (although they did not create it); they have not been able to solve it because the packages do not carry enough information to allow the automated tools to do so.

1.3 Source Disconnected from Binaries

Traditional package management does not closely associate source code with the packages created from it. The binary package may include a hint about a filename to search for to find the source code that was used to build the package, but there is no formal link contained in the packages to the actual code used to build the packages.

Many repositories carry only the most recent versions of packages. Therefore, even if you know which repository you got a package from, you may not be able to access the source for the binary packages you have downloaded because it may have been removed when the repository was upgraded to a new version. (Some tools help ameliorate this problem by offering to download the source code with binaries from repositories that carry the source code in a related directory, but this is only a convention and is limited.)

1.4 Namespace Arbitrary and Unmanaged

Traditional package management does not provide a globally unique mechanism for avoiding package name, version, and release number collisions; all collision-avoidance is done by convention and is generally successful only when the scope is sufficiently limited. Package dependencies (as opposed to file dependencies) suffer from this; they are generally valid only within the closed scope of a single distribution; they generally have no global validity.

It can also be difficult for users to find the right packages for their systems. Both SUSE and Fedora provide RPMs for version 1.2.8 of the iptables utility; if a user found release 101 from

SUSE and thought it was a good idea to apply it to Fedora Core 2, they would quite likely break their systems.

1.5 Build Configuration

Traditional packaging systems have a granular definition of architecture, not reflecting the true variety of architectures available. They try to reduce the possibilities to common cases (`i386`, `i486`, `i586`, `i686`, `x86_64`, etc.) when, in reality, there are many more variables. But to build packages for many combinations means storing a new version of the entire package for every combination built, and then requires the ability to differentiate between the packages and choose the right one. While some conventions have been loosely established in some user communities, most of the time customization has required individual users to rebuild from source code, whether they want to or not.

In addition, most packaging systems build their source code in an inflexible way; it is not easy to keep local modifications to the source code while still tracking changes made to the distribution (Gentoo is the most prominent exception to this rule).

1.6 Fragile Scripts

Traditional package management systems allow the packager to attach arbitrary shell scripts to packages as metadata. These scripts are run in response to package actions such as installation and removal. This approach creates several problems.

- Bugs in scripts are often catastrophic and require complicated workarounds in newer versions of packages. This can arbitrarily limit the ability to revert to old versions of packages.
- Most of the scripts are boilerplate that is copied from package to package. This increases the potential for error, both from faulty transcription (introducing new errors while copying) and from transcription of faults (preserving old errors while copying).
- Triggers (scripts contained in one package but run in response to an action done to a different package) introduce levels of complexity that defy reasonable QA efforts.
- Scripts cannot be customized to handle local system needs.
- Scripts embedded in traditional packages often fail when a package written for one distribution is installed on another distribution.

2 Introduction to Conary

Conary provides a fresh approach to open source software management and provisioning, one that applies new ideas from distributed configuration management tools such as GNU arch and monotone. Rather than concentrating on separate package files as RPM and dpkg do, Conary uses networked repositories containing a structured version hierarchy of all the files and organized sets of files in a distribution.

This new approach gives us exciting new features:

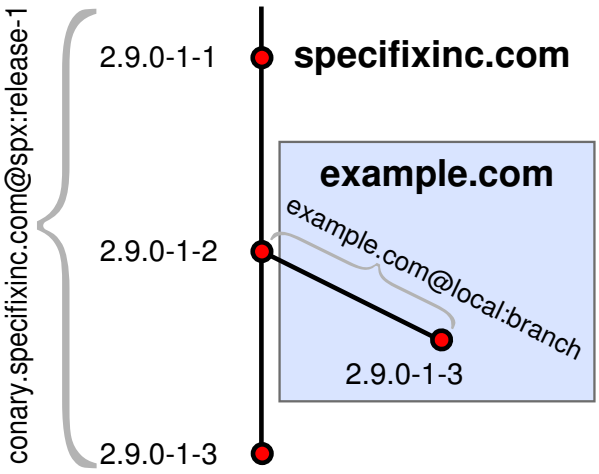
- Conary allows you to maintain and publish changes, both by allowing you to create new branches of development, and by helping track changes to existing branches of development while maintaining local changes.
- Conary intelligently preserves local changes on installed systems. An update

will not blindly obliterate changes that you have made on your local system.

- Conary can duplicate local changes made on one machine, installing those changes systematically on other machines, thereby easing provisioning of large sets of similar or identical systems.

3 Distributed Version Tree

Conary keeps track of versions in a tree structure, much like a source code control system. The difference between Conary and many source code control systems is that Conary does not need all the branches of a tree to be kept in a single place. For example, if Specifix maintains a kernel at `specifixinc.com`, and you, working for `example.com`, want to maintain a branch from that kernel, your branch could be stored on your machines, with the root of that branch connected to the tree stored on Specifix's machines.



3.1 Repository

Conary stores everything in a **distributed repository**, instead of in package files. The repository is a network-accessible database that contains files for multiple packages, and multiple versions of these packages, on multiple development branches. Nothing is ever removed from the repository once it has been added. In simple terms, Conary is like a source control system married to a package system.

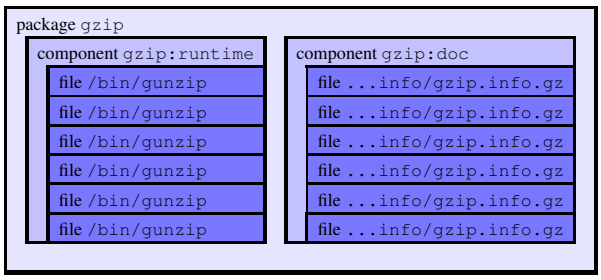
3.2 Files

When Conary stores a file in the repository, it tracks it by a unique file identifier rather than by name. Among other things, this allows Conary to track changes to file names—the file name is merely one piece of metadata associated with the file, just like the ownership, permission, timestamp, and contents. If you think of the repository as a filesystem, the file identifier is like an inode number.

3.3 Troves, Packages, and Components

When you build software with Conary, it collects the files into **components**, and then collects the components into one or more **packages**. Components and packages are both called **troves**. A trove is (generically) a collection of files or other troves.

A package does not directly contain files; a package references components, and the components reference files. Every component's name is constructed from the name of its container package, a `:` character, and a suffix describing the component. Conary has several standard component suffixes: `:source`, `:runtime`, `:devel`, `:docs`, and so forth. Conary automatically assigns files to components during the build process, but you can overrule its assignments and create arbitrary component suffixes as appropriate.



One component, with the suffix `:source`, holds all source files (archives, patches, and build instructions); the other components hold files to be installed. The `:source` component is not included in any package, since several different packages can be built from the same source component. For example, the `mozilla:source` component builds the packages `mozilla`, `mozilla-mail`, `mozilla-chat`, and so forth. The version structure in Conary's repositories always tells exactly which source component was used to build any other component.

3.4 Labels and Versions

Conary uses strongly descriptive strings to compose the version and branch structure. The amount of description makes them quite long, so Conary hides as much of the string as possible for normal use. Conary version strings act somewhat like domain names, in that for normal use you need only a short portion. For example, the version `/conary.specifixinc.com@spx:trunk/2.2.3-4-2` can usually be referred to and displayed as `2.2.3-4-2`. The entire version string uniquely identifies both the source of a package and its intended context. These longer names are globally unique, preventing any confusion.

Let's dissect the version string `/conary.specifixinc.com@spx:trunk/2.2.3-4-2`. The first part, `conary.specifixinc.com@spx:trunk`, is a **label**. It holds three pieces of information:

- **The repository host name:** `conary.specifixinc.com`
- **Namespace:** `spx` A high-level context specifier that allows branch names to be reused by independent groups. Specifix will maintain a registry of namespace identifiers to prevent conflicts. Use `local` for branches that will never need to be shared with other organizations.
- **Branch name:** `trunk` This is the only portion of the label that is essentially arbitrary; and will be defined by the owner of the namespace it is part of.

The next part, `2.2.3-4-2`, contains the more traditional version information.

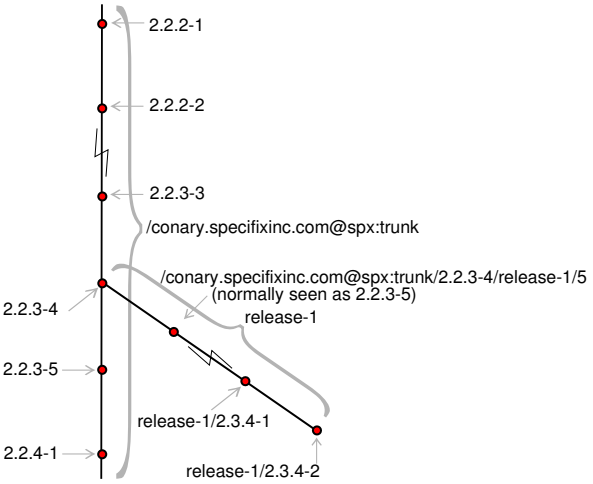
- **Upstream version string:** `2.2.3` This is the version number or string assigned by the upstream maintainer: Conary never interprets this string in any way; the only check it does is whether it is the same or different. It is there primarily to present useful information to the user. Conary never tries to determine whether one upstream version is "newer" or "older" than another. It makes these decisions based on the ordering specified by the repository's version tree.
- **Conary revision:** `4-2` This pair is composed from:
  - **Source build serial number:** `4` Incremented each time a version of the sources with the same upstream version string is checked in. It is similar to the release number used by traditional packaging systems.
  - **Binary build serial number:** `2` How many times this particular source package has been built. This number is not provided for source

packages, because it is meaningless in that context.

Conary describes branch structure by appending version strings, separated by a `/` character. The first step to make a release is to create a branch that specifies what is in the release. Let's create the `release-1` branch off the trunk: `/conary.specifixinc.com@spx:trunk/2.2.3-4/release-1` (note that because we are branching the source, there is no binary build number).

In this branch, `release-1` is a label. The label inherits the repository and namespace of the node it branches from; in this case, the full label is `conary.specifixinc.com@spx:release-1`

The first change that is committed to this branch can be specified in somewhat shortened form as `/conary.specifixinc.com@spx:trunk/2.2.3-4/release-1/5` Because the upstream version is the same as the node from which the branch descends, the upstream version may be omitted, and only the Conary version provided. Users will normally see this version expressed as `2.2.3-5`, so this string, still long even when it has been shortened by elision, will not degrade the user experience.



Labels also have an unusual property: a single label can reference *multiple* branches. To demonstrate why this is useful, let's look at the glib library. Like many other libraries, glib is designed to allow more than one version to be installed on the system at once. Older programs require glib 1.2; newer programs require glib 2. All new releases of glib 1.2 are compatible with programs written and compiled for older versions of glib 1.2; all new releases of glib 2 are compatible with programs written and compiled for older versions of glib 2. They are not, however, compatible with each other; a program compiled for glib 1.2 will certainly not run with glib 2. Therefore, a complete system requires that glib 1.2 and glib 2 both be installed.

Packaging systems often solve this problem by naming the packages differently, putting part of the version number into the name of the package (i.e. `glib` and `glib2`). This works, but it dilutes the revision history that the repository model provides.

By contrast, Conary solves this problem by allowing labels to apply to more than one branch. To see how, we will start by "going back in time" and looking at the version string for glib on the trunk with only glib 1.2 packaged: `/conary.specifixinc.com@spx:trunk/1.2.10-19-3`

Now, we want to add glib 2 to the repository. We want to have a branch for continuing maintenance of maintain glib 1.2, though, so let's create that first: `/conary.specifixinc.com@spx:trunk/1.2.10-19-3/glib1.2`

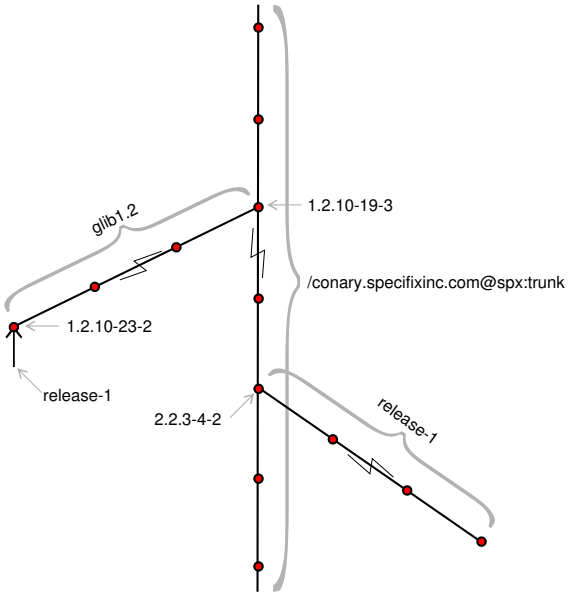
Now, we upgrade the trunk to glib 2: `/conary.specifixinc.com@spx:trunk/2.2.3-1-1`

Having maintained both glib 1.2 and glib 2 for a while, we decide that we want to make

our first release. We will label every package in the release, including two versions of glib: `/conary.specifixinc.com@spx:trunk/2.2.3-4-2/release-1/4-2` and `/conary.specifixinc.com@spx:trunk/1.2.10-19-3/glib1.2/23-2/release-1/23-2`

The label `conary.specifixinc.com@spx:release-1` now specifies *both* versions of glib. Therefore, if you install `glib conary.specifixinc.com@spx:release-1`, you will get both versions of glib.

Normally, the label to install will be set by installation scripts, and Conary will automatically install both versions of glib. Of course, updates will be applied only when there is a change; an update to glib 1.2 does not affect glib 2. In other words, it "just works" without you having to worry about it.



3.5 Shadows

The most powerful way to manage local changes is (of course) to build changes from source code. Conary makes this possible in two ways. One way is a simple branch, just like you would do with any source code control software. Unfortunately, this is not always the best solution.

Imagine a stock 2.6 Linux kernel packaged in Conary, being maintained on the `/linux26` branch (we have omitted the repository host name and namespace identifier from the label for brevity) of the `kernel:source` package, currently at version `2.6.5-1` (note that because it is a source package, there is no binary build number). You have one patch that you want to add relative to that version, and then you wish to track that maintenance branch, keeping your own change up to date with the maintenance branch, and building new versions as you go.

If you create a new branch from `/linux26/2.6.5-1`, say `/linux26/2.6.5-1/mybranch`, all the work you do is relative to that one version. Creating a new branch does not help you, because the new branch goes off in its own direction from one point in development, rather than tracking changes. Therefore, when the new version `/linux26/2.6.6-1` is committed to the repository, the only way to represent that version in your branch would be to manually compare the changes and apply them all, bring your patch up to date, and commit your changes to your branch. This is time-consuming, and the branch structure does not represent what is really happening in that case.

Conary introduces a new concept: a **shadow**. A shadow acts primarily as a repository for local changes to a tree. A shadow tracks changes relative to a particular upstream version string and source build serial number. Therefore, you cannot change the upstream version of the package—though you can apply any patch you like. (In order to change the upstream version of the package, you would need to create a branch rather than a shadow.) The

name of a shadow is the name of the branch with `//shadowname` appended; for example, `/branch//shadow`. The whole branch is shadowed, so if `/branch/1.2.3-3` and `/branch//shadow` exist, then so does `/branch//shadow/1.2.3-3`, regardless of whether `/branch/1.2.3-3` existed at the time the shadow was created. Similarly, if `/branch/1.2.3-3/rel1/1.2.3-3` exists, then so does `/branch//shadow/1.2.3-3/rel1/1.2.3-3`.

Both `/branch/1.2.3-3` and `/branch//shadow/1.2.3-3` refer to exactly the same contents. Changes are represented with a dotted source build serial number, so the first change to `/branch/1.2.3-3` that you check in on the `/branch//shadow` shadow will be called `/branch//shadow/1.2.3-3.1`.

So, to track changes to the `/linux26` branch of the `kernel:source` package, you create the `mypatch` shadow of the `/linux26` branch, `/linux26//mypatch`, and therefore `/linux26//mypatch/2.6.5-1` now exists. Commit a patch to the shadow, and `/linux26//mypatch/2.6.5-1.1` exists. Later, when the `linux26` branch is updated to version `2.6.6-1`, you merely need to update your shadow, modify the patch to apply to the new kernel source code if necessary, and commit the your new changes to the shadow, where they will be named `/linux26//mypatch/2.6.6-1.1`. You can use the shadow branch name `/linux26//mypatch` just like you can use the branch name `/linux26`; you can install that branch, and `conary update` will use the same rules to find the latest version on the shadow that it uses to find the latest version on the branch.

3.6 Flavors

Conary has a unified approach to handling multiple architectures and modified configurations. It has a very fine-grained view of architecture and configuration. Architectures are viewed as an instruction set, including settings for optional capabilities. Configuration is set with system-wide flags. Each separate architecture/configuration combination built is called a **flavor.**

Using flavors, the same source package can be built multiple times with different architecture and configuration settings. For example, it could be built once for `x86` with `i686` and `SSE2` enabled, and once for `x86` with `i686` enabled but `SSE2` disabled. Each of those architecture builds could be done twice, once with `PAM` enabled, and once with `PAM` disabled. All these versions, built from exactly the same sources, are stored together in the repository.

At install time, Conary picks the most appropriate flavor of a component to install for the local machine and configuration (unless you override Conary's choice, of course). Furthermore, if two flavors of a component do not have overlapping files, and both are compatible with the local machine and configuration, both can be installed. For example, library files for the `i386` family are kept in `/lib` and `/usr/lib`, but for `x86_64` they are kept in `/lib64` and `/usr/lib64`, so there is no reason that they should not both be installed, and since the AMD64 platform can run both, it is convenient to have them both installed.

4 Changesets

Just as source code control systems use patch files to describe the differences between two versions of a file, Conary uses **changesets** to

describe the differences between versions of troves and files. These changesets include information on how files have changed, as well as how the troves that reference those files have changed.

These changesets are often transient objects; they are created as part of an operation and disappear when that operation has completed. They can also be stored in files, however, which allows them to be distributed like the packages produced by a classical package management system.

Applying changesets rather than installing new versions of packages allows Conary to update only the parts of a package that have changed, rather than blindly reinstalling every file in the package.

Besides saving space and bandwidth, representing updates as changes has another advantage: it allows merging. Conary intelligently merges changes not only to file contents, but also to file metadata such as permissions.

This capability is very useful if you wish to maintain a branch or shadow of a package—for example, keeping current with vendor maintenance of a package, while adding a couple of patches to meet local needs.

Conary also keeps track of local changes in essentially the same way, preserving them. When, for example, you add a few lines to a configuration file on an installed system, and then a new version of a package is released with changes to that configuration file, Conary can merge the two unless there is a direct conflict (unusual but possible). If you change a file's permission bits, those changes will be preserved across upgrades.

Conary supports two types of change sets:

- The differences between two versions in a repository
- The complete contents of a version in a repository (logically, this is the difference between nothing at all and that version)

In the first case, where Conary is calculating the differences between two different versions, the result is a **relative changeset**. In the second case, where Conary is encoding the entire content of the version, the result is an **absolute changeset**. (If you use an absolute changeset to upgrade to the version provided in the absolute changeset, Conary internally converts the changeset to a relative changeset, thereby preserving your local changes.) Absolute changesets are convenient ways of distributing versions of troves and files to users who have various versions of those items already installed on their systems. In practice, they can be distributed just like package files created by traditional package management systems.

Conary can do two things with one of these changesets. It can update a system, either directly from a changeset file, or by asking the repository to provide a changeset and then applying that changeset. It can also store existing changesets in a repository. This capability will be used in the future to provide repository mirroring, and it can also be used to move changes from one repository to a branch in a different repository.

4.1 Representing Local Changes

Conary can also generate a **local changeset** that is a relative changeset showing the difference between the repository and the local system for the version of a trove that is installed. You can distribute a local changeset to another machine in two ways:

- You can distribute it to other machines

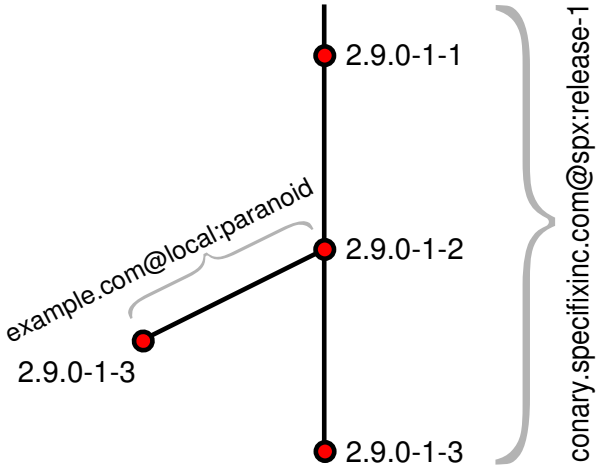
with the same version of the trove in question installed.

- You can commit the local changeset to a branch of a repository, and then update to that branch on target machines.

There is an important distinction between the two cases. In the first case, the machine that applies the changeset will act as if those changes had been made by the system's administrator; since those changes are not in a repository they are not versioned. In the second case, however, the machine gets those changes by updating the trove to the branch that contains those changes, and it can continue to track changes from that branch.

For example, assume that you have machines with troves from branches labeled `conary.specifixinc.com@spx:rel1` installed, and you have some local changes that you want to distribute to a group of machines. Let's say that after updating to version `2.9.0-1-2` of `tmpwatch`, you want to change the permissions of the /usr/sbin/tmpwatch binary because you are paranoid: `chmod 100 /usr/sbin/tmpwatch` Now, you record that change in a local changeset; that changeset is relative to `2.9.0-1-2`, and describes your local changes.

You then commit your local changeset to the `conary.example.com@local:paranoid` branch in your local repository. Now, on all the machines in the group, you can update `tmpwatch conary.example.com@local:paranoid`. Each machine will now look in the `conary.example.com` repository on the `paranoid` branch if you simply run `conary update tmpwatch`. This means that if you make further changes to the `tmpwatch` package, you can commit those changes to the `paranoid` branch on the `conary.example.com` repository, and each of the machines will update to the latest version you have committed to that branch. Every time a new version of `tmpwatch` is released on the `conary.specifixinc.com@spx:rel1` branch, you will have to apply the changeset to the `conary.example.com@local:paranoid` branch before the machines with your `paranoid` branch installed will update their copies of `tmpwatch`.



If rather than maintaining a branch, you merely want to distribute some changes that are local to the group of machines, you do not want to commit the local changeset to the repository. Instead, you want to copy the changeset file (let's call it paranoid.ccs) to each machine and run `conary localcommit paranoid.ccs` on each machine. Now, your change to permissions applies to each system, but `conary update tmpwatch` will still look at `conary.specifixinc.com@spx:rel1` and Conary will apply updates to `tmpwatch` from `conary.specifixinc.com@spx:rel1` without additional work required on your part, and it will preserve the change to the permissions of the /usr/sbin/tmpwatch binary on each machine.

Both ways of managing local change are useful. Committing local changesets to a repository is best for systems with entirely centralized management policy, where all sys-

tem changes must be cleared by some central agency, whereas distributing local changesets is best when individual systems are expected to autonomously update themselves asynchronously.

4.2 Merging

When Conary updates a system, it does not blindly obliterate all changes that have been made on the local system. Instead, it does a three-way merge between the currently installed version of a a file as originally installed, that file on the local system, and the version of the file being installed. If an attribute of the file was not changed on the local system, that attribute's value is set from the new version of the package. Similarly, if the attribute did not change between versions of the package, the attribute from the local system is preserved. The only time conflicts occur is if both the new value and the local value of the attribute have changed; in that case a warning is given and the administrator needs to resolve the conflict.

For configuration files, Conary creates and applies context diffs. This preserves changes using the the widely-understood diff/patch process.

4.3 Efficiency

Conary is more efficient than traditional packaging systems in several ways.

- By utilizing relative changesets whenever possible, Conary uses less bandwidth.
- By modifying only changed files on updates, Conary uses less time to do updates, particularly for large packages with small changes.
- By using a versioned repository, Conary saves space because unchanged files are stored once for the whole repository, instead of once in each version of each package.
- By enabling distributed repositories, Conary
  - saves the time it takes to maintain a modified copy of an entire repository, and
  - saves the space it takes to store complete copies of an entire repository.

4.4 Rollbacks

Because Conary updates systems by applying changesets, and because it is able to follow changes on the local system intrinsically, it easily supports **rollbacks**. If requested, Conary can store an inverse changeset that represents each **transaction** (a set of trove updates that maintains system consistency, including any dependencies) that it commits to the local system. If the update creates or causes problems, the administrator can ask Conary to install the changeset that represents the rollback.

Because rollbacks can affect each other, they are strictly stacked; you can (in effect) go backward through time, but you cannot browse. You have to apply the most recent rollback before you apply the next most recent rollback, and so forth.

This might seem like a great inconvenience, but it is not. Because Conary maintains local changes vigorously, including merging changes to configuration files, and because all the old versions you might have installed before are still in the repositories they came from, you can "update" to older versions of troves and get practically the same effect as rolling back your upgrade from that older version.

Applying rollbacks can be more convenient when you know that you want to roll back the

previous few transactions and restore the system to the state it was in, say, two hours ago. However, if you want to be selective, "upgrading" to an older version is actually more convenient than it would be to try to select a rollback transaction that contains the change you have in mind.

5 Other Concepts

5.1 Dynamic Tags

In place of the fragile script metadata provided by traditional package management systems, Conary introduces a concept called **dynamic tags**. Files managed by Conary can have sets of arbitrary text tags that describe them. Some of these tags are defined by Conary (for example, `shlib` is reserved to describe shared library files that cause Conary to update /etc/ld.so.conf and run `ldconfig`), and others can be more arbitrary. (In order to allow tag semantics to be shared between repositories, it is likely that Specifix will host a global tag registry in the future.)

By convention, a tag is a noun or noun phrase describing the file; it is not a description of what to do to the file. That is, *file* is-a *tag*. For example, a shared library is tagged as `shlib` instead of as `ldconfig`. Similarly, an info file is tagged as `info-file`, not as `install-info`.

Conary can be explicitly directed to apply a tag to a file, and it can also automatically apply tags to files based on a **tag description** file. A tag description file provides the name of the tag, a set of regular expressions that determine which files the tag applies to, the path of the **tag handler** program that Conary runs to process changes involving tagged files, and a list of actions that the handler cares about. Conary then calls the handler at appropriate times to handle the changes involving the tagged files.

Actions include changes involving either the tagged files or the tag handlers. Conary will pass in lists of affected files whenever it makes sense, and will coalesce actions rather than running all possible actions once for every file or component installed.

The current list of possible actions is:

- Tagged files have been installed or updated; Conary provides a list of all installed or updated tagged files.
- Tagged files are going to be removed; Conary provides a list of all tagged files to be removed.
- Tagged files have been removed; Conary provides a list of filenames that were removed.
- The tag handler itself has been installed or updated; Conary provides a list of all tagged files already installed on the system.
- The tag handler itself will be removed; Conary provides a list of all the tagged files already installed on the system to facilitate cleanup.

Because the tag description files list the actions they handle, the tag handler API can be expanded easily while maintaining backward compatibility with old handlers.

Avoiding duplication between packages by writing scripts once instead of many times avoids bugs in scripts. Practically speaking, it avoids whole classes of common bugs that cause package upgrades to break installed software, and even more importantly from a provisioning standpoint, bugs that would cause rollbacks to fail. It makes it much easier to fix

bugs when they do occur, without any need for "trigger" scripts that are often needed to work around script bugs in traditional package management. It also allows components to be installed across distributions—as long as they agree on the semantics for the tags, the actions taken for any particular tag will be correct for the distribution on which the package is being installed.

Calling tag handlers when they have been updated makes recovery from bugs in older versions of tag handlers relatively benign; Conary needs to install only a single new tag handler with the capability to recover from the effects of the bug. Older versions of packages with tagged files will use the new, fixed tag handler, which allows you to revert those packages to older versions as desired, without fear of re-introducing bugs created by old versions of scripts.

Furthermore, storing the scripts as files in the filesystem instead of as metadata in a package database means:

- they can be modified to suit local system peculiarities, and those modifications will be tracked just like other configuration file modifications;
- they are easier for system administrators to inspect; and
- they are more readily available for system administrators to use for custom tasks.

5.2 Groups and Filesets

There are two other kinds of troves that we did not discuss when we introduced the trove concept: groups and filesets.

Filesets are troves that contain only files, but those files come from components in the repository. They allow custom re-arrangements of any set of files in the repository. (They have no analog at all in the classical package model.) Each fileset's name is prefixed with `fileset-`, and that prefix is reserved for filesets only.

Filesets are useful primarily for creating small embedded systems. With traditional packaging systems, you are essentially limited to installing a system, then creating an archive containing only the files you want; this limits the options for upgrading the system. With Conary, you can instead create a fileset that references the files, and you can update that fileset whenever the components on which it is based are updated, and use Conary to update even very thin embedded images.

The desire to be able to create working filesets was a large motive for using file-specific metadata instead of trove-specific metadata wherever possible. For example, files in filesets maintain their tags, which means that exactly the right actions will be taken for the fileset. If Conary had package scripts like traditional package managers, it would be impossible to automatically determine which parts (if any) of the script should be included in the fileset. (As already discussed, scripts have other problems that tags solve; this is just another one of the architectural reasons that tags are preferable to scripts.)

Groups are troves that contain any other kind of trove, and the troves are found in the repository. (The task lists used by apt are similar to groups, as are the components used by anaconda, the Red Hat installation program.) Each group's name is prefixed with `group-`, and that prefix is reserved for groups only.

Groups are useful for any situation in which you want to create a group of components that should be versioned and managed together. Groups are versioned like any trove, including packages and components. Also, a group ref-

erences only specific versions of troves. Therefore, if you install a precise version of a group, you know exactly which versions of the included components are installed; if you update a group, you know exactly which versions of the included components have been updated.

If you have a group installed and you then erase a component of the group without changing the group itself, the local changeset for the group will show the removal of that component from the group. This makes groups a powerful mechanism administrators can use to easily browse the state of installed systems.

The relationship between all four kinds of troves is illustrated as follows:

Troves		built from	
		source	repository
contain	files	component	fileset
	troves	package*	group

*packages contain only components

Groups and filesets are built from `:source` components just like packages. The contents of a group or fileset is specified as plain text in a source file; then the group or fileset is built just like a package.

This means that groups and filesets can be branched and shadowed just like packages can. So if you have a local branch with only one modified package on it, and then you want to create a branch of the whole distribution containing your package, you can branch the group that represents the whole distribution, changing only one line to point to your locally changed file. You do not have to have a full local branch of any of the other packages or components.

Furthermore, when the distribution from which you have branched is updated, your modification to the group can easily follow the updates, so you can keep your distribution in sync without having to copy all the packages and components.

6 Further Work

An alpha release of Conary is now available from `http://www.specifixinc.com`, along with a Linux distribution built with Conary. While these releases allow users and developers to begin making use of Conary's features, there is significant work remaining.

The shadow design discussed in this paper has not yet been implemented.

Conary does not yet resolve dependencies. Although some dependency information is already generated and tracked on a per-file basis, no effort is made to ensure that those dependencies are resolved when components are installed.

As Conary and Conary-based distributions become more popular, there will be a need for both repository caches and repository mirrors. While some preliminary design work has been done for each of these, no implementation work has begun.

The implementation of flavors is preliminary, especially in regards to configuration settings. While limited testing has been done with troves built for varying architectures and Specifix's build scripts implement some configuration settings, Conary does not yet properly select the flavor to install on a system.

Conclusion

Conary was designed to address many of the limitations of the traditional packaging

metaphor. The enormous growth in the Linux developer base over the past decade has shown that packaging systems do not scale well to multiple repositories with conflicting content, and can make it difficult for large numbers of developers to coordinate package releases.

Conary provides flexible branching, which enables it to find both binaries and sources anywhere on the Internet, and allows the local administrator to preserve local changes and create local development branches of those packages. By providing a name space separator as part of the branch names, Conary allows many groups to use the same tool while building a single distributed version tree, without any formal collaboration between the groups.

Innovations such as shadows and versioning groups of packages and files (allowing those container objects themselves to be branched and shadowed) significantly reduce the difficulty of maintaining customized Linux distributions. Instead of being forced to accept complete responsibility for all aspects of the distribution, developers can now concentrate on maintaining just their changes. Those changes are represented in a concise way that can track upstream changes to the entire distribution.

Conary is designed to enable a loosely-coupled, Internet-based collaborative approach to building Linux distributions. By making branching and shadowing inexpensive operations that can change almost any aspect of a Linux system, we hope members of the Linux community will be able to build the Linux distribution they want, rather than use one that is merely close enough.

On-demand Linux for Power-aware Embedded Sensors

Carl Worth, Mike Bajura, Jaroslav Flidr, and Brian Schott
USC/Information Sciences Institute
{cworth,mbajura,jflidr,bschott}@east.isi.edu

Abstract

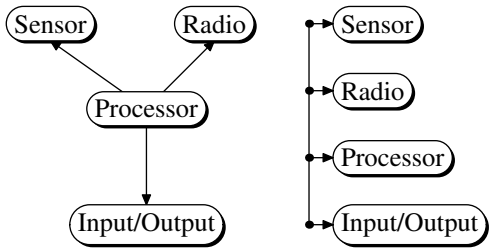
We introduce a distributed sensor architecture which enables high-performance 32-bit Linux capabilities to be embedded in a sensor which operates at the average power overhead of a small microcontroller. Adapting Linux to this architecture places increased emphasis on the performance of the Linux power-up/shutdown and suspend/resume cycles.

Our reference hardware implementation is described in detail. An acoustic beamforming application demonstrates a 4X power improvement over a centralized architecture.

1 Introduction

Traditional sensor platform architectures are based on a hub-and-spoke model with peripherals clustered around a central processor as shown in Figure 1(a). In this model, the lower-bound of total system power is set by the lowest active mode of the central processor which must be continually active to broker peripheral operations.

System power is typically reduced by using less-capable processors or microcontrollers in place of the central processor. Although sensor activity is mostly infrequent and bursty with low average computational requirements, peak processing requirements can still be quite high.



(a) Hub-and-spoke (b) Distributed

Figure 1: Alternate sensor node architectures

Within a system design, this creates tension between the desire for the high-performance processing capability of a larger processor and the low-power operation of a smaller one.

This tension is further complicated by the observation that while many large processors require significantly more power than small ones when inactive, they also often provide significantly more power-efficient computation when active. Another tradeoff in this design space weighs the strength of development and debugging tools, such as Linux, available for larger processors versus the constrained programming environments available for small ones.

Replacing the hub-and-spoke architecture with a distributed model as shown in Figure 1(b) can decouple processing from peripheral operation and create a system that combines the strengths of both large and small processors.

In this model, processor and peripherals become autonomous modules that are each powered independently. High-performance processing can be made available when needed, but without increasing the lower-bound of total system power. Low average system power can be achieved by operating for a majority of the time in extremely low-power modes with only essential modules active.

This distributed architecture places Linux in an unconventional role as a peer module rather than as a central processor. This emphasizes the performance of the power-up/shutdown and suspend/resume cycles as keys for achieving low average system power.

The remainder of this paper is organized as follows. Section 2 describes several popular research and commercial sensor platforms. Section 3 recounts the design challenges we faced as we built a reference sensor node with autonomous modules. As usual, real-world issues forced difficult engineering decisions. Section 4 details the modules we have built so far.

Our hardware is in a more complete state than our software. We have identified some aspects of the behavior of Linux that we need to investigate more fully. These issues are discussed in Section 5. Section 6 contains power results we have obtained with a vehicle tracking algorithm. Finally, Section 7 draws conclusion and Section 8 describes areas for future work.

2 Related Work

Applied research in wireless sensor networks has made use of a variety of platforms with varying processing capabilities and power requirements, but almost always with a hub-and-spoke model. Several platforms are described below in order from more-capable, higher-power platforms to less-capable, lower-power platforms.

2.1 PC/104

PC/104 [3] is a well-supported specification for PC-compatible, embedded systems consisting of stacking modules. Cerpa et al [2] chose PC/104 systems for their "high end" sensors in a tiered deployment for habitat monitoring. Their cited reasons for choosing PC/104 include the ability to run PC-compatible software (i.e. Linux) and the wide spectrum of available PC/104 modules.

PC/104 provides great flexibility and the power requirements are lower than that of desktop PCs. However, at 1-2 Watts per module[3], and with most sensors requiring at least two modules, this platform requires too much power for many sensor applications.

2.2 Embedded StrongARM devices/PDAs

Off-the-shelf devices based on low-power, embedded processors such as the Intel SA-1110 or PXA25x offer another convenient platform for sensor network research. Compared to x86-class processors, these processors offer a significant power savings along with a reduction in maximum clock rate and the absence of a hardware floating-point unit.

Representative devices of this class include the HP iPAQ, the CerfCube[4], and the Crossbow Stargate[14]. These devices are extensible via Compact Flash, Bluetooth, etc. but have less flexibility than PC/104. And while the power requirements of these systems can be less than a comparable PC/104 stack, Mainwaring et al[9] found that at 2.5W active power, the power usage of the CerfCube was excessive for long-term use in a sensor network.

Several research sensors have been developed with architectures similar to these off-the-shelf platforms. These include the μAMPS[10] and WINS[1] nodes. For example, the WINS node

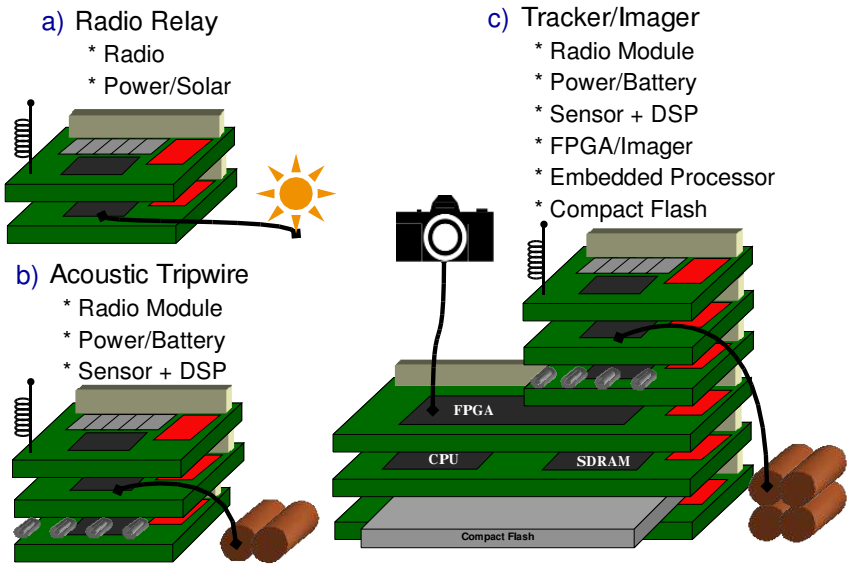


Figure 2: Power-aware sensor concept

has a central 133MHz StrongARM SA-1100 processor along with radio and sensor peripherals. As measured by Raghunathan et al[12] the WINS node operates in the range of 360-1080 mW and can also be placed into a 64 mW sleep mode.

2.3 Motes

An example of a very low-power sensor architecture is that of the Berkeley Motes[6, 7, 8]. Current Motes are based around a central microcontroller (MCU) such as the ATmega 90LS8535, an 8-bit MCU with 128KB Flash and 8KB SRAM. The Mote includes a radio and has serial connections and 10-bit analog ADC ports to various sensors on expansion modules. Typical power consumption for this sensor when active is in the 10-100 mW range. Sleep power is about 60 μW.

Mote-class sensors demonstrate that a wide variety of low-bandwidth sensing applications can be accomplished with very small processors and with very little memory. The limitation of these systems is encountered when an application doesn't fit within the memory and processing footprint of the MCU. High bandwidth sensor processing is beyond the capabilities of these small-scale sensors and there is little room for expansion.

3 Implementation

Our primary system design goal was to construct a family of interchangeable processor, sensor, and communication modules that can be mixed and matched according to the application requirements. Ideally, our architecture would be able scale from simple sensors as shown in Figure 2(a) and Figure 2(b), to complex sensors as shown in Figure 2(c) without having to learn and port to a new platform at every scale.

Other goals were driven by practical experiences of using other platforms in the field. Rapid prototyping is an important concern. The ability to create testbeds using COTS peripherals is a strength of platforms such as PC/104. Availability of Linux device drivers

and a friendly programming environment were strong motivators in our implementation decisions. Data collection is an important step in sensor network algorithm development. We wanted lots of data storage and data networking options in our new platform.

Primary constraints on the system design include size and power. We targeted the size of the Berkeley Mote, while still supporting a Linux-capable processor in the stack. In the end, the size was dictated by the minimum footprint of a Compact Flash socket and our chosen stack connector. Power in our system needs to be able to scale from 1 mW to a few Watts. The low power target limited many of our implementation choices.

An early design decision was how the autonomous modules would communicate. Interfaces such as ethernet were quickly dismissed due to power requirements. In small embedded devices, the most power-efficient communication is available with hardware-supported interfaces such as Serial Peripheral Interface (SPI), Controller Area Network (CAN), Universal Asynchronous Receiver Transmitter (UART), and Inter-Integrated Circuit (IIC or I2C). Each of these interfaces has strengths and weaknesses. I2C supports multi-master operation, but has a limit of 100kbps on most devices. SPI has faster transfers, (up to a few megabits per second), but has limited multi-master support in most devices and little flow control. UART is widely supported, but requires clock agreement on both interfaces. CAN is multi-master, but the bus drivers in the supporting devices are relatively high power, (since CAN is designed for long cables).

We decided to support the three major interface standards (I2C, SPI, and UART). We allocated six 8-bit channels on our connector and specified two preferred channels for each standard interface. In order to prevent bus contention and power leakage when modules are off, all modules have bus isolation switches between themselves and the connector. We also introduced a separate I2C control network for module discovery and coordination of the switches.

A small microcontroller (MCU) is standard on most modules to control the bus isolation switches, a power switch, and any module-specific functions. The MCUs are intended to be always on when the node is operating, (with a power overhead as low as .05 mW). They network with each other over I2C to facilitate module discovery and coordinate access to the channels using a common messaging protocol.

Figure 3 contains a diagram of the features common to each module, as well as an optional processor expansion bus so that high-speed, high-power peripherals, (USB, Compact Flash, LCD, and AC97 audio), can be used in the stack or removed for low-power operation.

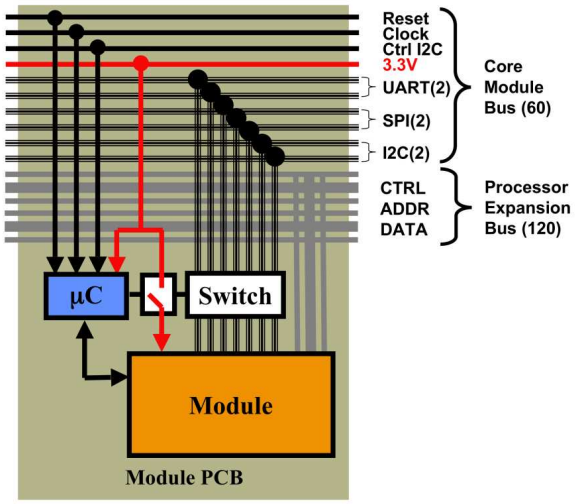


Figure 3: Power-aware module diagram

4 Available Modules

Our hardware modules are small boards approximately 6.5×4.5cm (2.5×1.75"), with a

180-pin connector on either side. So far, we have designed, built, and tested 4 modules, 3 of which are shown in Figures 4 and 5.

PXA A module including an Intel PXA255 XScale processor, 64MB of SDRAM and 32MB of Flash. This board supports dynamic voltage scaling, an active clock rate range of 100-400MHz, and a 33MHz idle mode. An SA-1111 coprocessor provides support for USB master and two Compact Flash cards. All interface lines are routed to the stack connector.

ADC A four-channel, 12-bit analog-to-digital converter module. The MCU has sufficient memory for a dedicated 7kB sample buffer. Basic signal processing can be performed by the local MCU or samples can be efficiently transmitted over an SPI interface to the PXA module for advanced processing. An important point is that the PXA255 processor can be off or suspended during data sampling.

IOB The power & I/O board is the only required module in the stack. It provides the primary power supply and contains most of the digital I/O connectors, (USB master and slave, SPI, I2C, and UART).

FPGA This module was developed as an emulation board for a low-power DSP being developed at MIT[15], but is interesting for other high-performance applications. It contains a Xilinx Virtex-II 3000 FPGA, 2MB of synchronous SRAM, and 32MB of SDRAM. This module operates as a coprocessor on the memory bus of the PXA255 and supports the two SPI channels on the stack.

We also have a Compact Flash (CF) board, two of which can be placed into the stack. This board connects to the processor expansion

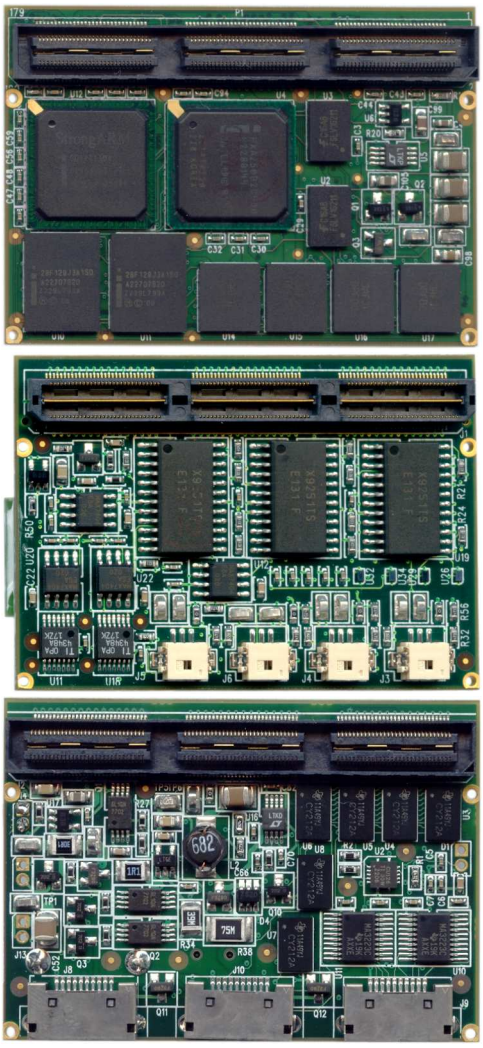


Figure 4: PXA, ADC, and IOB modules at actual size

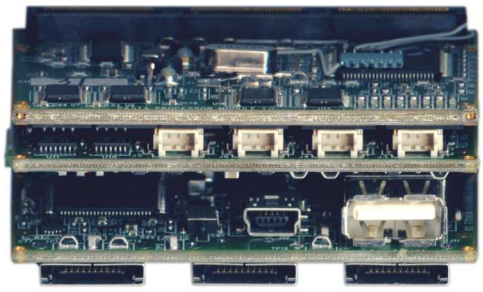


Figure 5: Stacked modules

bus so that it acts as a daughter-board of the PXA module rather than an independent module. Figure 6 shows a stack consisting of IOB and PXA modules along with a CF board populated with a CF ethernet adapter.

Figure 6: Stack including CF board

Table 1 shows design-time estimates of the power consumed by each module for various operational modes. In the "off" mode everything on the module is powered off except for the power-control MCU. The PXA module has the widest operational power range due to the the range of processor clock rates and various possibilities in processor and memory utilization. Figure 7 provides more details of the how the PXA module power can scale from 0.05 mW to 1.5 W. The innermost portion of this diagram includes figures for the overhead of power conversion and the 32kHz MCU on the IOB module.

Module	Mode	Power
PXA	off	0.05 mW
PXA	suspended	2.5 - 7.5 mW
PXA	active	150 - 1530 mW
IOB	active	0.1 mW
ADC	off	0.05 mW
ADC	active	40 mW

Table 1: Power modes for various modules

5 Impact on Linux

In a conventional hub-and-spoke model, Linux runs on a central processor and manages some number of peripheral devices. In contrast, our distributed architecture places Linux on an autonomous module which is a peer to other modules. Any other module might request a power transition of the Linux module, from off to powered, from suspended to active, etc.

The efficiency of these power transitions is a critical component of the average system power. The distributed platform is designed to achieve low average system power through aggressive duty-cycling of high-powered components. The time and energy spent during power-mode transitions is overhead that must be amortized, imposing limits on practical duty cycles that can be used.

We are currently using Linux version 2.4.21 with the standard ARM and PXA patches as well as customizations for our PXA module. The user-level software distribution is derived primarily from the handhelds.org[11] Familiar distribution. Our reference sensor application (see Section 6) does not turn the PXA module off, but does suspend/resume the processor aggressively to achieve active operation for a few milliseconds once per second. The time spent during suspend/resume is divided between time spent in driver callbacks and time spent in the kernel proper. Table 2 shows the times we have measured for these transitions on our PXA module.

Transition	Time
Suspend drivers	507 ms
Suspend kernel	13 ms
Resume kernel	78 μs
Resume drivers	25 ms

Table 2: Linux transitions with driver callbacks

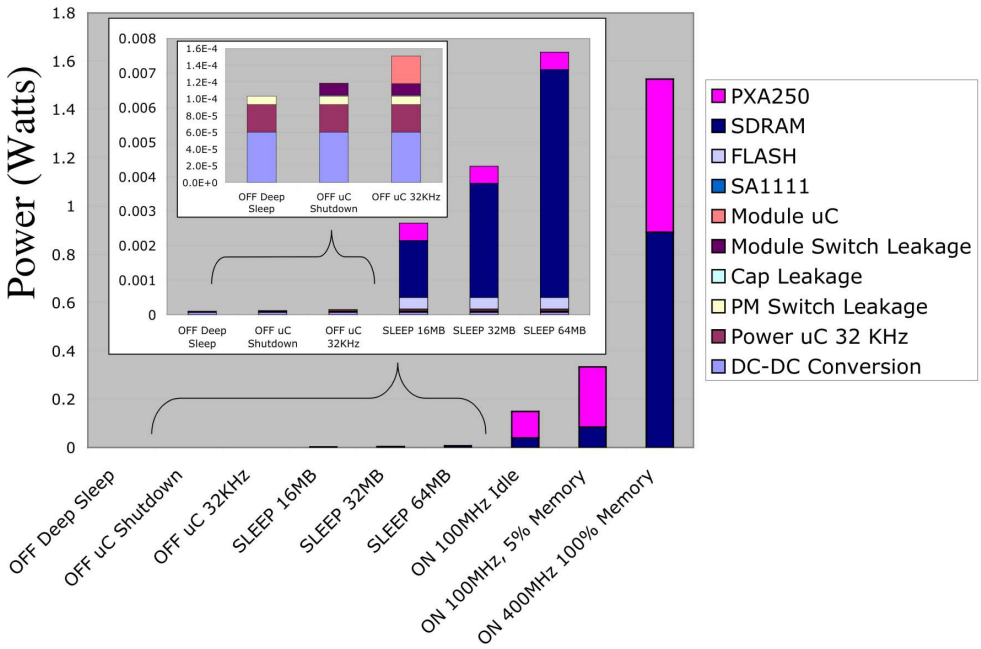


Figure 7: PXA module power states

We expect the suspend/resume transitions of the Linux kernel itself to behave in a symmetric fashion. However, we measured a very responsive resume time of 78 μs and a much slower suspend time of 13 ms. We do not yet have a complete explanation for why the suspend process is so much slower, although we have accounted for a 1 ms delay that is caused by the MCU on our PXA board, and therefore not a feature of the standard Linux kernel.

A much more significant problem is the time spent in the power management callbacks of various subsystems and drivers. A suspend time of 520ms spells disaster for an application such as the one described in Section 6.

We quickly tracked down the source of this long suspend time to the USB OHCI driver (hcd). An ill-fated decision in our design was the choice of the SA-1111 coprocessor. One difficulty we encountered was that we had to provide our own suspend/resume callbacks as they do not exist for the SA-1111-based hcd in the standard kernel. To simplify the task, we have ported the PCI-based OHCI code which contains a call to `mdelay (500)` along with the comment, "Suspend chip and let things settle down a bit." This single 500ms delay accounts for over 98% of the time required to suspend drivers. We suspect that this constant can be safely reduced so that much of the time lost during suspend can be recovered. Even so, USB-related timeouts, etc. are on the order of milliseconds—orders of magnitude more than the time required by the kernel.

Clearly, this poses a serious problem for applications with a high duty cycling requirement. We are currently working around the long driver suspend times by simply removing drivers for non-essential devices and subsystems, (such as SA-1111), prior to running an application with a restricted power budget. This allows the convenience of things such as using a USB 802.11 adapter during development and debugging without the long suspend times of the USB drivers during execution.

6 Results

We implemented a vehicle tracking algorithm using 4-channel acoustic beamforming. In this application, data is continually sampled at a rate of 1kHz, but signal processing only needs to be performed at a maximum rate of 1Hz.

In previous work[13], this algorithm was implemented on a successor to the WINS node, (a hub-and-spoke platform with an Intel SA-1110 processor). Although efficient signal processing software was developed, system power savings were modest since the processor had to remain active (yet mostly idle) at all times simply to drive data collection.

We have ported the algorithm to the distributed platform described in this paper. On this platform, the signal processing for one second's worth of data can be completed in 3 ms by the PXA255 processor running at 100MHz. Since this is a newer processor than the SA-1110 of the WINS platform, direct comparison of power numbers between the two platforms would be unfair. Instead, we estimate the power needed for two implementations of the algorithm on the distributed node.

The first version is intended to behave as if in a hub-and-spoke system. The processor remains active at all times to store samples into main memory. The second version takes advantage of the distributed nature of the platform. Linux on the PXA module is suspended as much as possible while the ADC module continues to sample and buffer data. This approach adds the overhead needed to suspend/resume Linux and to transfer data from the ADC module to the PXA module over the SPI channel. The amount of data to be transferred is 8192 bytes, (4 channels $*$ 1024 samples/s $*$ 2 bytes/sample $*$ 1 s). The SPI transfer rate is 1.8 Mbps yielding a total transfer time of 36.4 ms.

We measured the power consumed by the PXA module at two different stages in the algorithm. During active computation the PXA module consumes 528 mW. When mostly idle, (e.g. when transferring 1kHz data from the ADC module), it consumes 370 mW. We have not yet measured the average power consumed during the suspend or resume transitions, but we use an estimate of 370 mW. This estimate should be conservative as the actual power usage should ramp down to less than 10 mW during the transition.

Combining these measurements with estimates from Table 1 and the time measurements from Table 2, we compute the total energy spent to compute one result per second. From this we can determine the average power necessary for the complete algorithm. These results are given for both versions of the algorithm in Tables 3 and 4.

Module/Mode	Power	Time	Energy
IOB active	0.1 mW	1 s	0.1 mJ
ADC active	40 mW	1 s	40.0 mJ
PXA active	528 mW	3 ms	1.6 mJ
PXA idle	370 mW	997 ms	368.9 mJ
Estimated energy per second			410.6 mJ
Estimated system power: 411 mW			

Table 3: Hub-and-spoke power requirements for beamforming

Module/Mode	Power	Time	Energy
IOB active	0.1 mW	1.0 s	0.1 mJ
ADC active	40 mW	1.0 s	40.0 mJ
PXA suspended	7.5 mW	948 ms	7.1 mJ
PXA resuming	370 mW	78 μs	28.9 μJ
PXA transferring	370 mW	36.4 ms	13.5 mJ
PXA processing	528 mW	3 ms	1.6 mJ
PXA suspending	370 mW	13 ms	4.8 mJ
Estimated energy per second			95.9 mJ
Estimated system power: 96 mW			

Table 4: Distributed power requirements for beamforming

7 Conclusion

The 96 mW beamforming result marks a success for the distributed sensor platform—a 4X power reduction over the 411 mW required for the hub-and-spoke platform. This shows that it is possible to take advantage of 32-bit, Linux processing without average power exceeding the 10-100 mW range of a less-capable sensor based on a 8-bit microcontroller, (i.e. a Mote).

8 Future Work

The field of power-aware sensing is rich, and we have only just begun to explore the possibilities, even within our own platform. Many applications require a much smaller power budget than the 96 mW result we have demonstrated. Our long-term goal is to design sensors capable of operating entirely from scavenged energy, (e.g. solar), which requires operation in the range of 1 mW[5].

We are currently building a low-power "tripwire" module which will implement single-channel acoustic vehicle detection. This will allow the PXA module to be completely off when a vehicle is not present. We anticipate that this will allow beamforming within a power budget as low as 10 mW.

As mentioned in Section 5, there remains a fair amount of engineering and research with regards to the role of Linux within a distributed sensor. This includes reducing the time required in the power management callbacks of all relevant drivers as much as possible. An additional task is to move from Linux version 2.4 to 2.6. The dynamic nature of the new unified device model in 2.6 should make a natural fit with a platform consisting of autonomous modules that can be powered on and off at any point.

Additionally, new power-scheduling research could better take advantage of the wide range of power modes in this platform. For example, Linux could monitor the frequency and duty cycles of the power-up/shutdown and suspend/resume cycles. It would then be possible to provide intelligent feedback into these cycles based on the relative overhead of each transition. This work would complement current efforts that dynamically adjust voltage and clock rate based on system load.

9 Availability

This work has been developed as part of the *Power-Aware Sensing Tracking and Analysis (PASTA)* project, but we hope that it will be useful to researchers and hobbyists with a wide range of applications. To that end we are working toward making the hardware modules available at cost. Further details will be made available at the PASTA website, `http://pasta.east.isi.edu`. All of the software developed under PASTA is also available there under the terms of the GNU General Public License (GPL).

10 Acknowledgements

This research is sponsored by the Defense Advanced Research Projects Agency and Air Force Research Laboratory, Air Force Materiel Command, USAF, under agreement number F33615-02-2-4005. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Defense Advanced Research Projects Agency, the Air Force Research Laboratory, or the U.S. Government.

References

[1] Jonathan R. Agre, Loren P. Clare, Gregory J. Pottie, and Nikolai P. Romanov. Development platform for self-organizing wireless sensor networks. In *Proceedings of Aerosense*, pages 257–268. International Society of Optical Engineering, 1999.

[2] Alberto Cerpa, Jeremy Elson, Deborah Estrin, Lewis Girod, Michael Hamilton, and Jerry Zhao. Habitat monitoring: Application driver for wireless communications technology. In *ACM SIGCOMM Workshop on Data Communications in Latin America and the Caribbean*, April 2001.

[3] PC/104 Consortium. PC/104 Embedded-PC Modules, 2004. `http://www.pc104.org/`.

[4] Intrinsic Corporation. CerfCube embedded strongarm system. `http://www.intrinsyc.com/products/cerfcube/`.

[5] L. Doherty, B.A. Warneke, B. Boser, and K.S.J. Pister. Energy and performance considerations for smart dust. In *International Journal of Parallel and Distributed Sensor Networks*, 2001.

[6] Jessica Feng, Farinaz Koushanfar, and Miodrag Potkonjak. System-architectures for sensor networks issues, alternatives, and directions. In *International Conference on Computer Design: VLSI in Computers and Processors*, page 226. IEEE, 2002.

[7] Jason Hill, Robert Szewczyk, Alec Woo, Seth Hollar, David E. Culler, and Kristofer S.J. Pister. System architecture directions for networked sensors. In *Architectural Support for Programming Languages and Operating Systems*, pages 93–104, 2000.

[8] J.M. Kahn, R.H. Katz, and K.S.J. Pister. Next century challenges: mobile networking for smart dust. In *Proceedings of the fifth annual ACM/IEEE international conference on Mobile computing and networking*, pages 271–278. ACM Press, 1999.

[9] Alan Mainwaring, David Culler, Joseph Polastre, Robert Szewczyk, and John Anderson. Wireless sensor networks for habitat monitoring. In *Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications*, pages 88–97. ACM Press, 2002.

[10] R. Min, M. Bhardwaj, S. Cho, A. Sinha, E. Shih, A. Wang, and A. Chandrakasan. An architecture for a power-aware distributed microsensor node, 2000.

[11] The Handhelds.org Project. Handhelds.org (web document). `http://www.handhelds.org`.

[12] V. Raghunathan, C. Schurgers, S. Park, and M. Srivastava. Energy aware wireless microsensor networks, 2002.

[13] R. Riley, S. Thakkar, J. Czarnaski, and B. Schott. Power-aware acoustic beamforming. In *Proceedings of the Fifth International Military Sensing Symposium*, 2002.

[14] Crossbow Technology. Stargate xscale processor platform. `http://www.xbow.com`.

[15] A. Wang and A. Chandrakasan. Energy-efficient dsps for wireless sensor networks, 2002.

Virtually Linux

Virtualization Techniques in Linux

Chris Wright
OSDL
chrisw@osdl.org

Abstract

Virtualization provides an abstraction layer mapping a virtual resource to a real resource. Such an abstraction allows one machine to be carved into many virtual machines as well as allowing a cluster of machines to be viewed as one. Linux provides a wealth of virtualization offerings. The technologies range in the problems they solve, the models they are useful in, and their respective maturity. This paper surveys some of the current virtualization techniques available to Linux users, and it reviews ways to leverage these technologies. Virtualization can be used to provide things such as quality of service resource allocation, resource isolation for security or sandboxing, transparent resource redirection for availability and throughput, and simulation environments for testing and debugging.

1 Introduction

Virtualization has many manifestations in computer science. At the simplest level it can be viewed as a layer of abstraction which helps delegate functionality—typically handling resource utilization. This abstraction layer often helps map a *virtual* resource to a *physical* or *real* resource. The virtual resource is then presented directly to the resource consumer obscuring the existence of the real resource. This can be implemented through hardware[1] or software [16, 21, 19], may include any subset of a machine's resources, and has a wide variety of applications. Such usages include machine emulation, hardware consolidation, resource isolation, quality of service resource allocation, and transparent resource redirection. Applications of these usage models include virtual hosting, security, high availability, high throughput, testing, and ease of administration.

It is interesting to note that differing virtualization models may have inversely correlated proportions of virtual to physical resources. For example, the method of carving up a single machine into multiple machines—useful in hardware consolidation or virtual hosting—looks quite different from a single system image (SSI) [15]—useful in clustering. This paper primarily focuses on providing multiple virtual instances of a single physical resource, however, it does cover some examples of a single virtual resource mapping to multiple physical resources.

Modern processors are sufficiently powerful to provide ample resources to more than one operating environment at a time. Of course, timesharing systems have always allowed for concurrent application execution. However, there are many ways in which these concurrent applications may effect one another. Because the

[1] For example, an MMU helps with translation of virtual to physical memory addresses.

operating system provides access to shared resources such as the CPU, memory, I/O devices, file system, network, etc., one application's use of the system's resources may effect another's. This can have negative effects on both quality of service and security. Carving a single machine into a series of independent virtual machines can eliminate the quality of service and security issues.

At the same time, modern computing systems, inclusive of both hardware and software, are subject to failures and scalability problems. The application of virtualization can hide these shortcomings by distributing computing loads across a cluster of physical systems which may present a single *virtual* interface to an application.

The remainder of this paper is organized as follows. Section 2 presents a variety of virtualization techniques. Section 3 gives a detailed comparison of some of these techniques. Section 4 presents conclusions drawn from the comparisons.

2 Virtualization Techniques

The term "virtual" is one of those horribly overloaded terms in computing. For the purpose of this paper, we will define virtualization as a technique for mapping virtual resources to real resources. These virtual resources are then used by the resource consumer, fully decoupled from any real resources that may or may not exist. As discussed in Section 1 the virtual resource may be some or all of a system's resources.

There are many virtualization techniques available to Linux users, and these techniques can be leveraged through a variety of applications. The techniques reviewed in this paper fall roughly into two categories: *complete* virtualization, Section 2.1, which provides all or nearly all of a system's resources; and *partial* virtualization, Section 2.2, which provides only a specialized subset of resources. Understanding the different techniques helps identify which technique is the best given a specific set of requirements.

2.1 Complete Virtualization

Complete virtualization techniques involve creating a fully functional and isolated virtual system which can support an OS. This instance of the OS may have no indication that it is not being run natively on real hardware, and it is often referred to as the *guest*. Host-based virtual systems run atop an existing *host* OS. Others run atop a thin supervisor which just helps multiplex resources to the virtual systems. Typically the host machine is capable of supporting many concurrent virtual systems, each with its own guest OS instance. These virtual systems can be created by simple software emulation or by more complicated methods. These types of complete virtualization techniques differ in terms of efficiency and performance, portability for either the host or the guest OS, and functional goals.

The rest of this section is organized at follows. Section 2.1.1 is a look at pure software processor emulation techniques. Section 2.1.2 looks at the virtual OS approach taken by User-mode Linux. Finally, Section 2.1.3 reviews techniques using virtual machines and virtual machine monitors.

2.1.1 Processor Emulation

Processor emulation is one technique used to provide complete virtualization. In this case, the CPU is emulated entirely in software. Additionally, it is typical to find peripheral devices such as keyboard, mouse, VGA, network, timer chips, etc. supported by the emulator.

The emulation is done in user-space software, which makes it a rich environment for debugging system level software running in the emulator. Also, this technique has great advantages for portability at the cost of runtime performance. The emulator may easily run on various hardware architectures, as all emulation is done in software. Further, because these are hardware emulators, there is often little to no restriction on what OS software can be executed. However, the dynamic translation required to translate hardware instructions from the emulated processor to the native processor is pure overhead, and thus can be hundreds of times slower than native instructions [22].

An exhaustive survey of processor emulators is beyond the scope of this paper. Here we take a brief look at a few of the prevalent emulators often used to host virtual Linux instances:

- QEMU CPU emulator
- Bochs
- PearPC
- Valgrind.

QEMU [21] is a CPU emulator that does dynamic instruction translation. It maintains a translation cache for efficiency. It can be used as a user-mode emulator which will run Linux binaries compiled for the CPU that QEMU is emulating regardless of the host platform. Also, QEMU can do full system emulation, which allows one to boot an OS on the QEMU emulated CPU. While the QEMU user-mode is available for many architectures, the complete system emulation mode is only available for x86 and is in testing for PowerPC. The x86 emulator provides all the PC peripheral devices needed to boot an OS, and can easily run an unmodified Linux kernel. It also features debugger support which can be quite useful for debugging a Linux kernel.

Bochs [2] is an IA-32[2] CPU emulator. It does dynamic compilation and is often cited as being rather slow [3]. Similar to QEMU, Bochs provides full platform emulation sufficient for running an OS, and it can boot an unmodified Linux kernel. While Bochs is highly portable, it targets only the IA-32 processor.

PearPC [17] is a PowerPC CPU emulator. The generic PearPC CPU emulator can be ported and is slow. PearPC also provides a PowerPC CPU emulator that is specific to x86 hosts. This version uses dynamic instruction translation and caching techniques (similar to QEMU) which improve the speed substantially.

Valgrind [14] is worthy of mentioning as it is both a very useful tool and contains an x86-to-x86 just-in-time (JIT) compiler, thus emulating the x86 CPU. However, this tool has been historically used like Purify [10] as a memory checker, and not typically used for bringing up a virtual instance of Linux on the emulated CPU[3]. It handles user-space emulation, but not full system emulation. Valgrind is developed as an instrumentation framework around the JIT, so it can been expanded to be a general purpose “meta-tool for program supervision.” [14]

2.1.2 Virtual OS

The virtual OS is rather specific to User-mode Linux (UML) [6]. In this case, the physical machine is controlled by a normal Linux kernel. The host kernel provides hardware resources to each UML instance. The UML kernel provides virtual hardware resources to all the processes within a UML instance. The processes on a UML instance can run native code

[2] IA-32 and x86 are used interchangeably in this paper.

[3] Efforts have been made to run UML under Valgrind.

on the processor, avoiding pure emulation, and UML kernel traps all privileged needs. The UML kernel is, in fact, just an architectural port—*ARCH=um*—of the normal Linux kernel. The architecture specific code in UML is actually user-space code which uses the host Linux kernel system call interface. In other words, it is a port of the Linux kernel to the Linux kernel. This form of virtualization can be used for security[4], debugging, or virtual hosting.

2.1.3 Virtual Machine

The virtual machine (VM) has been studied for well over thirty years [8, 9]. It is a powerful abstraction that gives the illusion of running on dedicated real hardware without such physical requirements. In its early incarnations it provided a safe and convenient way to share expensive hardware resources. The well-known IBM VM/370 [5] simulated the System/390 hardware, presenting multiple independent VM's to the user. The VM/370 was aided by the System/370 hardware design, a luxury which is often not available to the modern world of low-priced, powerful commodity processors based on the x86 architecture [23]. However, it is precisely this type of environment which can benefit from consolidating multiple hardware servers to a single amply powered machine.

The typical architecture includes a physical platform which runs a virtual machine monitor (VMM). This monitor carves up the physical resources and makes them available to each virtual machine. In some cases, the VMM is *host-based* requiring a host OS, host specific drivers and user-space code to launch a VM [13, 7]. As with processor emulation in Section 2.1.1, it is beyond the scope of this paper to give an exhaustive survey of virtual machine technologies. Here we take a brief look at a few of the prevalent projects which can be used to run Linux in a virtual machine:

- Plex86
- VMware[5]
- Xen

Plex86 [18] is one project that provides an x86 virtual machine. This project provides a hosted virtual machine monitor, requiring a host OS to run the plex86 VMM. Plex86 is quite specific to Linux. The host OS may be Linux (although other host kernels are supported) and requires a kernel module to help implement the VMM. It also makes some key assumptions regarding usage of the virtual x86 hardware and patches the guest Linux kernel to conform to these assumptions. Plex86 does very little to virtualize hardware I/O. Instead, Plex86 uses a Hardware Abstraction Layer (HAL) to handle virtual I/O to the hardware devices. This eliminates the need to provide any kind of virtual devices in the VM, and being host-based eliminates the need for the VMM to understand all the possible hardware on the host. I/O which is started in the guest OS is passed through the HAL using fairly simple guest kernel drivers which issue an `int $0xff`—which must not be used for other purposes on the host OS. The host VMM traps that software interrupt and handles the request accordingly. As noted by the project's author, Plex86 is still in a prototype state, and not really ready for meaningful benchmarking yet.

VMware [7] is worthy of mention, despite the fact that it is a commercial product. VMware™ Workstation [12] provides an x86 virtual machine and is in some ways similar to Plex86. It is a hosted virtual machine monitor, however,

[4] To be secure, UML must run in `skas` mode which requires a small patch to the host kernel

[5] VMware is a commercial product.

the goals of VMware Workstation include the ability to run a complete x86 OS without making any modifications. Therefore, it makes no assumptions about the guest OS. By emulating very standard hardware such as the PS/2 keyboard and mouse, the AMD PCnet™ network interface card or the Soundblaster 16 sound card the VM provides virtual hardware devices that can be run by standard guest OS drivers. Another x86 virtual machine from VMware is the ESX Server [26]—a pure virtual machine monitor that is not host-based. This method eliminates some of the overhead involved with running atop a host OS at the cost of requiring more hardware support in the VMM itself. As with Workstation, ESX requires no modifications to the guest OS. The lower overhead of ESX makes it a contender for a data center virtual hosting environment, where it could easily run multiple VM's on a single physical system.

Xen [16] is an x86 virtual machine monitor that provides a virtual hardware interface to the virtual machine. Typically, the virtual machine provides a hardware interface which is identical to the underlying hardware. However, the Xen VM hardware abstraction is similar but not identical to the underlying x86 hardware. This allows the VMM to overcome some of the shortcomings of the x86 architecture which make it difficult to virtualize [23]. A similar method was used for the Denali [1] isolation kernel. However, unlike Denali, the Xen VM supports a notion of a virtual address space. So the guest OS and applications may share resources just like a normal OS environment. In addition, guest kernels running in a Xen VM preserve the ABI to their applications. So, while there is a need to port the guest OS kernel to the Xen VM virtual hardware abstraction, the porting effort ends there. Further, given the similarity to the x86 architecture, the effort to port to Xen results in a very small amount of new OS code—well below 2% of the OS code base [16]. This method has proven to be quite effective when considering the minimal porting effort coupled with the impressive performance benchmarks [16].

2.2 Partial Virtualization

Partial virtualization techniques create virtualized resources that are a specialized subset of a complete system's resources rather than a complete virtual machine. These methods are typically used to present a virtual interface to clients or applications when limited isolation or virtualization is sufficient. Partial virtualization can have very different applications depending on the resource which is being virtualized. These techniques vary widely in the problems they solve, and in some cases can be used with alongside of complete virtualization. The remainder of this Section reviews these techniques.

2.2.1 Linux-Vserver

The Linux-Vserver [20] project takes some of the basic ideas of isolation from a virtual machine and implements them in a single host OS. The Linux kernel is patched to allow for multiple concurrent execution contexts, often called Virtual Private Servers[6] (VPS). This method eliminates any overhead associated with running multiple operating systems, multiple VM's and the supervisor VMM. Each context can have its own file system, its own network addresses, its own set of Linux Capabilities [25], and its own set of resource limits. With this level of software isolation, it is possible to run two concurrent contexts that are unable to interact with each other directly. It may still be possible to generate some indirect QoS degradation from *crosstalk* [24], however these effects should be largely mitigated by

[6] This is also the name given to Ensim's commercial product [4].

proper setting of each context's resource limits. While this solution does require a reasonably large kernel patch (a 337K patch against Linux 2.6.6), it is a very thin virtualization layer that efficiently isolates execution contexts.

2.2.2 Linux Virtual Server

The Linux Virtual Server Project [11] takes a very different view of server virtualization from Linux-Vserver, Section 2.2.1. Rather than creating a virtual operating environment for each server, it behaves as a network load balancer. The Linux Virtual Server, also referred to as IP Virtual Server (IPVS), presents a single network address for the network service and distributes client requests transparently to a hardware cluster of network servers. With IPVS, the client can be redirected to the next available resource using a variety of algorithms such as round robin and least connected. This is an example of virtualization used to provide enhanced availability throughput, or scalability. Further, this project in contrast with Linux-Vserver helps illustrate the difficulty in defining a "Virtual Server."

2.2.3 File system and Disks

The UNIX file system provides the basic namespace that applications use to interact with significant portions of the system. The root of a file system can be relocated in Linux using `chroot()`. This may be a stretch of the definition of virtualization, but this technique does allow a single server to give different views into the system global namespace. Tools like `chroot()` or the BSD `jail()` system[7] allow multiple applications to have completely private file system namespaces, which becomes an effective tool towards system virtualization. In fact, Linux-Vserver, Section 2.2.1, makes use of `chroot()` as key to its file system isolation. Linux has native support for per process private namespaces. This gives each process its own virtual or logical view of the system's global namespace, in a more powerful, flexible and secure manner than `chroot()`. Linux-Vserver is considering moving to namespaces as a replacement for `chroot()` isolation [20]. It would not be surprising to find other virtualization systems using the same technique for file system isolation.

Another layer of virtualization can be found in the disk or block device layer of the Linux kernel. The device-mapper allows administrators to create a virtual block device which is backed by one or more physical block devices. This type of virtualization is typically used for ease of administration.

3 Comparisons

Having reviewed a variety of virtualization techniques in Section 2, it is now useful to pick a representative subset and see how they compare with one another. For the sake of comparison, this paper will focus on QEMU, Usermode Linux, Xen, and Linux-Vserver. All four of these technologies can provide a virtual execution environment comprehensive enough to run either a complete OS, or at a minimum user-space applications.

3.1 QEMU

Pros:

- Portable to numerous architectures.
- Can be used to cross platforms.
- Can run guest OS unmodified.
- Can run on unmodified host OS.

[7] An implementation of BSD jail has been ported to Linux.

- Flexible, can run a full system or just isolated user-space programs.
- Very easy to debug system software.
- Security through isolation.

Cons:

- Processor emulation is much slower than virtualization.

3.2 User-mode Linux

Pros:

- Portable to numerous architectures.
- Can run on unmodified host OS.
- Efficient enough to run multiple instances on single host in virtual hosting environment.
- Very easy to debug system software.
- Security through isolation.

Cons:

- Still slower than a virtual machine.
- The guest OS kernel is not the same as a native one.

3.3 Xen

Pros:

- True virtual machine monitor for best performance.
- The guest OS user-space applications are binary compatible.
- No host OS, very clean virtual machine separation.
- Security through isolation.
- Ideal for virtual hosting environment, can scale up to 100 virtual machines.

Cons:

- The guest OS kernel must be ported to Xen virtual hardware architecture.

3.4 Linux-Vserver

Pros:

- Highly efficient way to isolate resources.
- Can conserve on disk and memory by sharing basic resources like shared libraries.
- Security through context separation.

Cons:

- Only one kernel instance, so quality of service may be hard to guarantee.

4 Conclusions

Virtualization is an old yet resurging technology. Virtual machine research is alive and well, and Linux provides a great testbed for new virtualization technologies. With a wealth of choices, Linux users are sure to find a virtualization technique that suits their requirements. From running as a guest OS on a virtual machine, to providing thin isolation environments for applications, to single system image clusters, Linux is thriving in this virtual reality.

References

[1] M. Shaw A. Whitaker and S. D. Gribble. Scale and Performance in the Denali Isolation Kernel. In *5th Symposium on Operating Systems Design and Implementation (OSDI)*, December 2002.

[2] Bochs IA-32 Emulator Project. `http://bochs.sourceforge.net/`.

[3] Bochs FAQ. `http://bochs.sourceforge.net/doc/docbook/user/faq.html#AEN273`.

[4] Ensim Corporation. Ensim Virtual Private Server, June 2004. `http://www.ensim.com/products/privateservers/index.html`.

[5] R. J. Creasy. The Origin of the VM/370 Time-Sharing System. *IBM Journal of Research and Development*, 25(5):483–490, September 1981.

[6] Jeff Dike et al. User-Mode Linux, July 2000. `http://user-mode-linux.sourceforge.net/`.

[7] Mendel Rosenblum et al. VMWare. `http://www.vmware.com/`, February 1998.

[8] R. Goldberg. *Architectural Principles for Virtual Computer Systems*. PhD thesis, Harvard University, 1972.

[9] R. P. Goldberg. Architecture of Virtual Machines. In *Proceedings of the Workshop on Virtual Computer Systems*, pages 74–112, 1973.

[10] Reed Hastings and Bob Joyce. Purify: Fast Detection of Memory Leaks and Access Errors. In *Proceedings of the Winter USENIX Conference*, 1992. Also available at `http://www.rational.com/support/techpapers/fast_detection/`.

[11] The Linux Virtual Server Project. `http://www.linuxvirtualserver.org/`.

[12] Ganesh Venkitachalam Jeremy Sugerman and Beng-Hong Lim. Virtualizing I/O Devices on VMware Workstation's Hosted Virtual Machine Monitor. In *USENIX Annual Technical Conference*, June 2001.

[13] Mac-on-Linux. `http://www.maconlinux.org/`.

[14] Julian Seward Nicholas Nethercote. Valgrind: A Program Supervision Framework. In *Electronic Notes in Theoretical Computer Science 89 No. 2*, 2003.

[15] Single System Image Clusters (SSI) for Linux. `http://openssi.org`.

[16] K. Fraser S. Hand T. Harris A. Ho R. Neugebauer I. Pratt P. Barham, B. Dragovic and A. Warfield. Xen and the Art of Virtualization. In *Symposium on Operating Systems Principles (SOSP)*, 2003.

[17] PearPC - PowerPC Architecture Emulator. `http://pearpc.sourceforge.net/`.

[18] Plex86 x86 Virtual Machine Project. `http://plex86.sourceforge.net/`.

[19] Herbert Poetzl. Linux-VServer Project, October 2001. `http://www.linux-vserver.org`.

[20] Herbert Poetzl. Linux-VServer Technology. In *LinuxTAG 2004*, June 2004.

[21] QEMU CPU Emulator. `http://fabrice.bellard.free.fr/qemu/`.

[22] QEMU Benchmarks. `http://fabrice.bellard.free.fr/qemu/benchmarks.html`.

[23] J. S. Robin and C. E. Irvine. Analysis of the intel pentium's ability to support a secure virtual machine monitor. In *9th USENIX Security Symposium*, pages 129–144, August 2000.

[24] D. L. Tennenhouse. Layered multiplexing considered harmful. In

Rudin and Williamson, editors, *Protocols for High Speed Networks*. May 1989.

[25] Winfried Trumper. Summary about POSIX.1e. `http://wt.xpilot.org/publications/posix.1e`, July 1999.

[26] C. A. Waldspurger. Memory Resource Management in VMware ESX Server. In *5th Symposium on Operating Systems Design and Implementation (OSDI)*, December 2002.