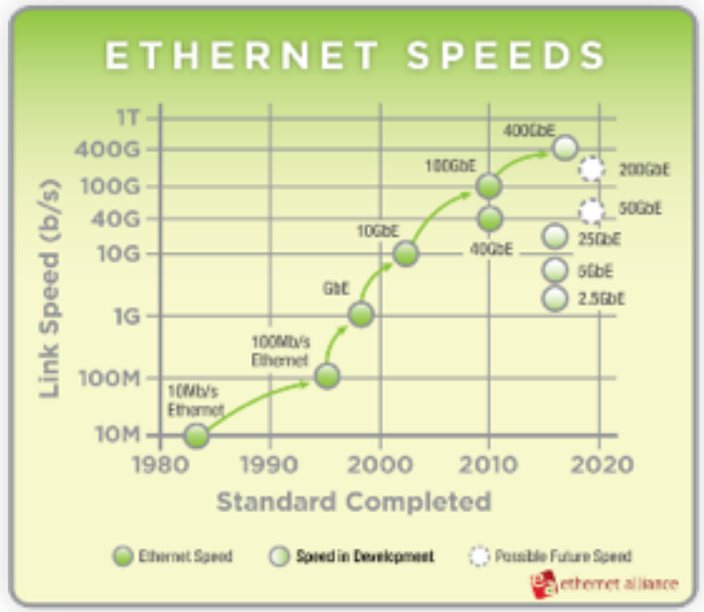
100G Networking Technology Overview

Christopher Lameter <cl@linux.com>
Fernando Garcia <fgarcia@dasgunt.com>

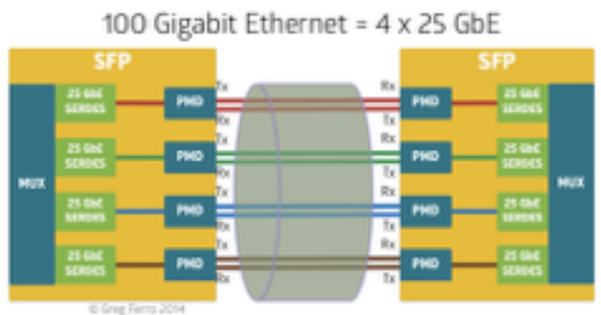
Toronto, August 23, 2016

Why 100G now?

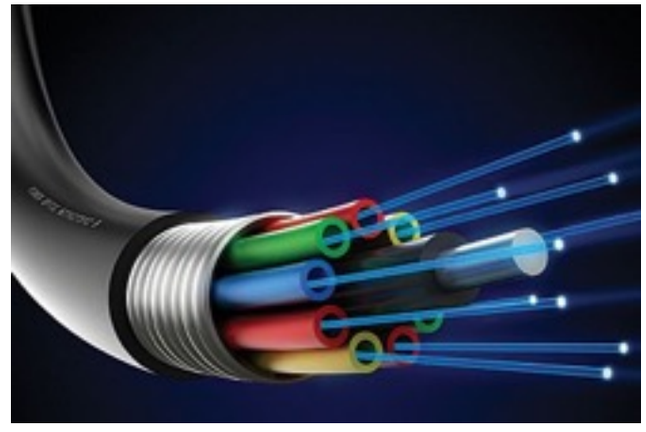
- Capacity and speed requirements on data links keep increasing.
- Fiber link reuse in the Connectivity providers (Allows Telcos to make better use of WAN links)
- Servers have begun to be capable of sustaining 100G to memory (Intel Skylake, IBM Power8+)
- Machine Learning Algorithms require more bandwidth
- Exascale Vision for 2020 of the US DoE to move the industry ahead.



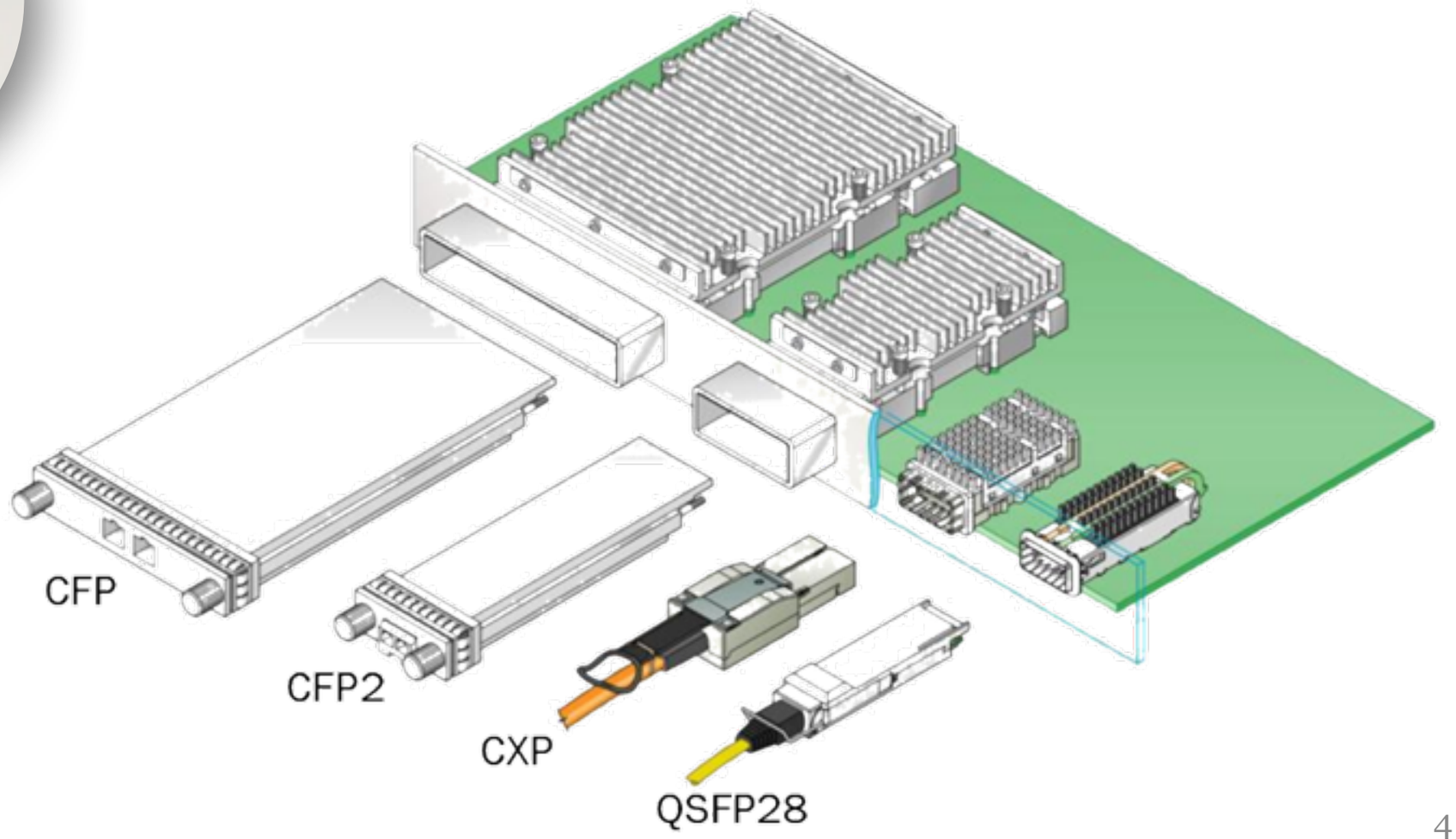
100G Networking Technologies



- 10 x 10G Link old standard CFP C??. Expensive. Lots of cabling. Has been in use for awhile for specialized uses.
- New 4 x 28G link standards "QSFP28". Brings down price to ranges of SFP and QSFP. Compact and designed to replace 10G and 40G networking.
- Infiniband (EDR)
  - Standard pushed by Mellanox.
  - Transitioning to lower Infiniband speeds through switches.
  - Most mature technology to date. Switches and NICs are available.
- Ethernet
  - Early deployment in 2015.
  - But most widely used chipset for switches recalled to be respun.
  - NICs are under development. Mature one is the Mellanox EDR adapter that can run in 100G Ethernet mode.
  - Maybe ready mid 2016.
- Omnipath (Intel)
  - Redesigned serialization. No legacy issues with Infiniband. More nodes. Designed for Exascale vision. Immature. Vendor claims production readiness but what is available has the character of an alpha release with limited functionality. Estimate that this is going to be more mature at the end of 2016.



CFP vs QSFP28: 100G Connectors



Splitting 100G Ethernet to 25G and 50G

- 100G is actually 4x25g (*QSFP28*), so 100G Ports can be split with "octopus cables" to lower speed.

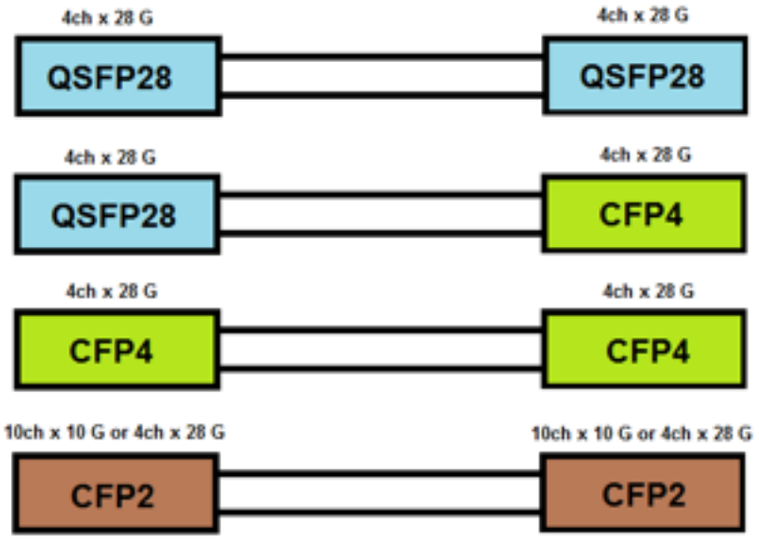
- 50G (2x25) and 25G (1x25G) speeds are available which doubles or quadruples the port density of switches.

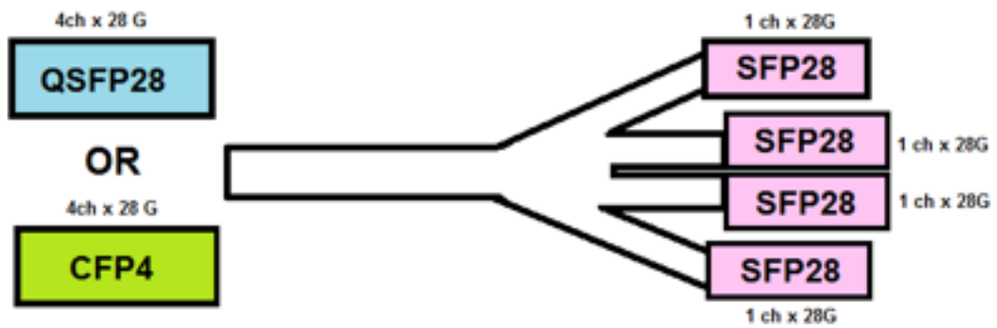
- Some switches can handle 32 links of 100G, 64 of 50G and 128 of 25G.

- It was a late idea. So 25G Ethernet standards are scheduled to be completed in 2016 only. Vendors are designing to a proposed standard.
- 50G Ethernet standard is in the works (2018-2020). May be the default in the future since storage speeds and memory speeds increase.
- 100G Ethernet done

- 25G Ethernet has a new connector standard called *SFP28*

100G Cabling and Connectors





100G Switches

	Ports	Status	Name
Mellanox Infiniband	EDR x 36	Released. Stable.	7700 Series
Broadcom	100G x 32 50G x 64 25G x 128	Rereleased after silicon problem.	Tomahawk Chip
Mellanox Ethernet	100G x 32 50G x 64	2Q ?	Spectrum
Intel	Omnipath x 48	Available	100 Series

Overwhelmed by data

Ethernet	10M	100M (Fast)	1G (Gigabit)	10G	100G
Time per bit	100 ns	10 ns	1 ns	0.1 ns	0.01 ns
Time for a MTU size frame 1500 bytes	1500 us	150 us	15 us	1.5 us	150 ns
Time for a 64 byte packet	64 us	6.4 us	640 ns	64 ns	6.4 ns
Packets per second	~10 K	~100 K	~1 M	~10 M	~100 M
Packets per 10 us		2 (small)	20 (small)	6 (MTU)	60 (MTU)

No time to process what you get?

- NICs have the problem of how to get the data to the application
- Flow Steering in the kernel allows the distribution of packets to multiple processors so that the processing scales. But there are not enough processing cores for 100G.
- NICs have extensive logic to offload operations and distribute the load.
- One NIC supports multiple servers of diverse architectures simultaneously.
- Support for virtualization. SR-IOV etc.
- Switch like logic on the chip.

1 us = 1 microsecond
= 1/1000000 seconds

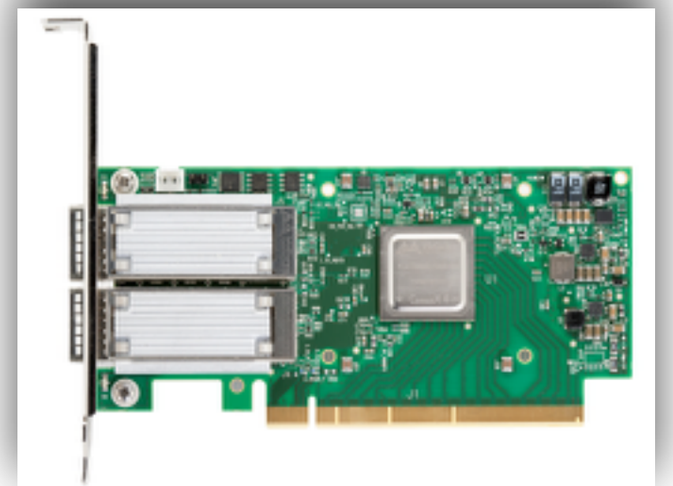
1 ns = 1 nanosecond
= 1/1000 us

Network send or receive syscall:
10-20 us

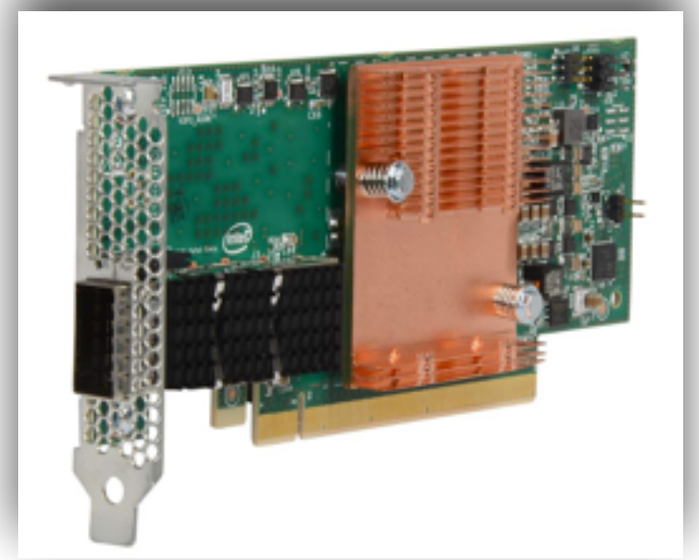
Main memory access:
~100 ns

Available 100G NICs

- **Mellanox ConnectX4 Adapter**
  - 100G Ethernet
  - EDR Infiniband
  - Sophisticated offloads.
  - Multi-Host
  - Evolution of ConnectX3



- **Intel Omnipath Adapter**
  - Focus on MPI
  - Omnipath only
  - Redesign of IB protocol to be a “Fabric”
  - “Fabric Adapter”. New Fabric APIs.
  - More nodes larger transfer sizes



Application Interfaces and 100G

1. Socket API (Posix)

   Run existing apps. Large code base. Large set of developers that know how to use the programming interface

2. Block level File I/O

   Another POSIX API. Remote filesystems like NFS may use NFSoRDMA etc

3. RDMA API

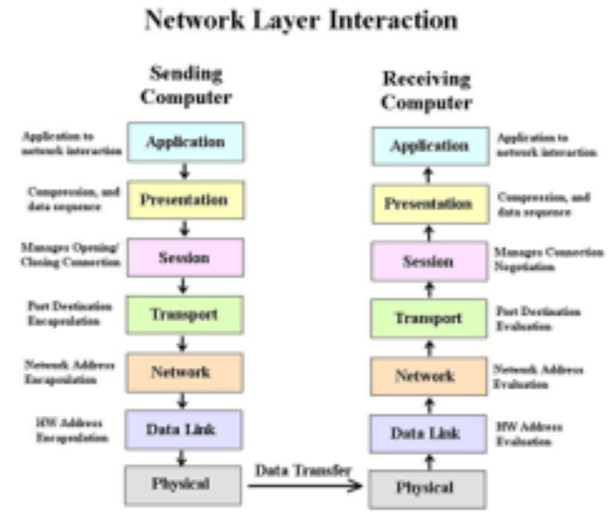
   1. One sided transfers
   2. Receive/SendQ in user space
   3. Talk directly to the hardware.

4. OFI

   Fabric API designed for application interaction not with the network but the “Fabric”

5. DPDK

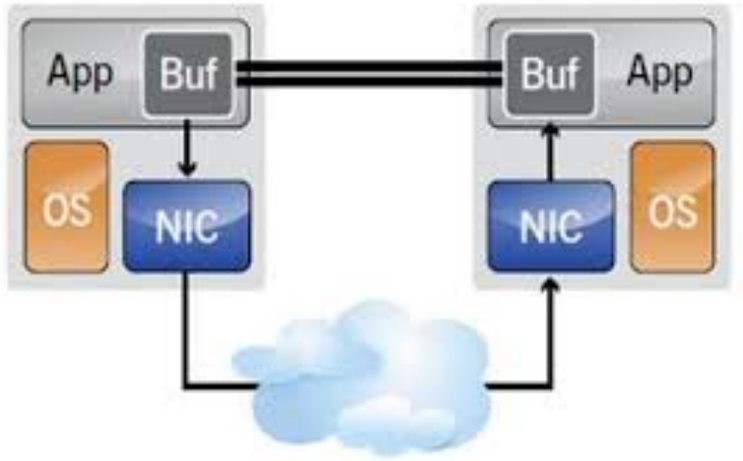
   Low level access to NIC from user space.



Using the Socket APIs with 100G

- Problem of queuing if you have a fast talker.
- Flow steering to scale to multipe processors
- Per processor queues to scale sending.
- Exploit offloads to send / receive large amounts of data
- May use protocol with congestion control (TCP) but then not able to use full bandwidth. Congestion control not tested with 100G.
- Restricted to Ethernet 100G. Use on non Ethernet Fabrics (IPoIB, IPoFabric) has various non Ethernet semantics. F.e. Layer 2 behaves differently and may offer up surprises.

RDMA / Infiniband API



- Allow use of native Infiniband functionality designed for higher speed.
- Supports both Infiniband and Onmipath.
- Single sided transfers via memory registration or classic messaging.
- Offload behavior by having RX and TX rings in user space.
- Group send / receive possible.
- Control of RDMA/Infiniband from user space with API that is process safe but allows direct interaction with an instance of the NIC.
- Can be used on Ethernet via ROCE and/or ROCEv2
- Generally traffic is not routable (ROCE V2 and Ethernet messaging of course is). Problem of getting into and out of fabric. Requires specialized gateways.

OPENFABRICS ALLIANCE

INFINIBAND TRADE ASSOCIATION

OFI (aka libfabric)

- Recent project by OFA to design a new API.
- Based on RDMA concepts.
- Driven by Intel to have an easier API than the ugly RDMA APIs. OFI is focusing on the application needs from a Fabric.
- Tested and developed for the needs of MPI at this point.
- Ability to avoid the RDMA kernel API via “native” drivers. Drivers can define API to their own user space libraries.
- Lack of some general functionality like Multicast.
- Immature at this point. Promise for the future to solve some of the issue coming with 100G networking.

Software Support for 100G technology

EDR via Mellanox ConnectX4 Adapter
- Linux 4.3. Redhat 7.2

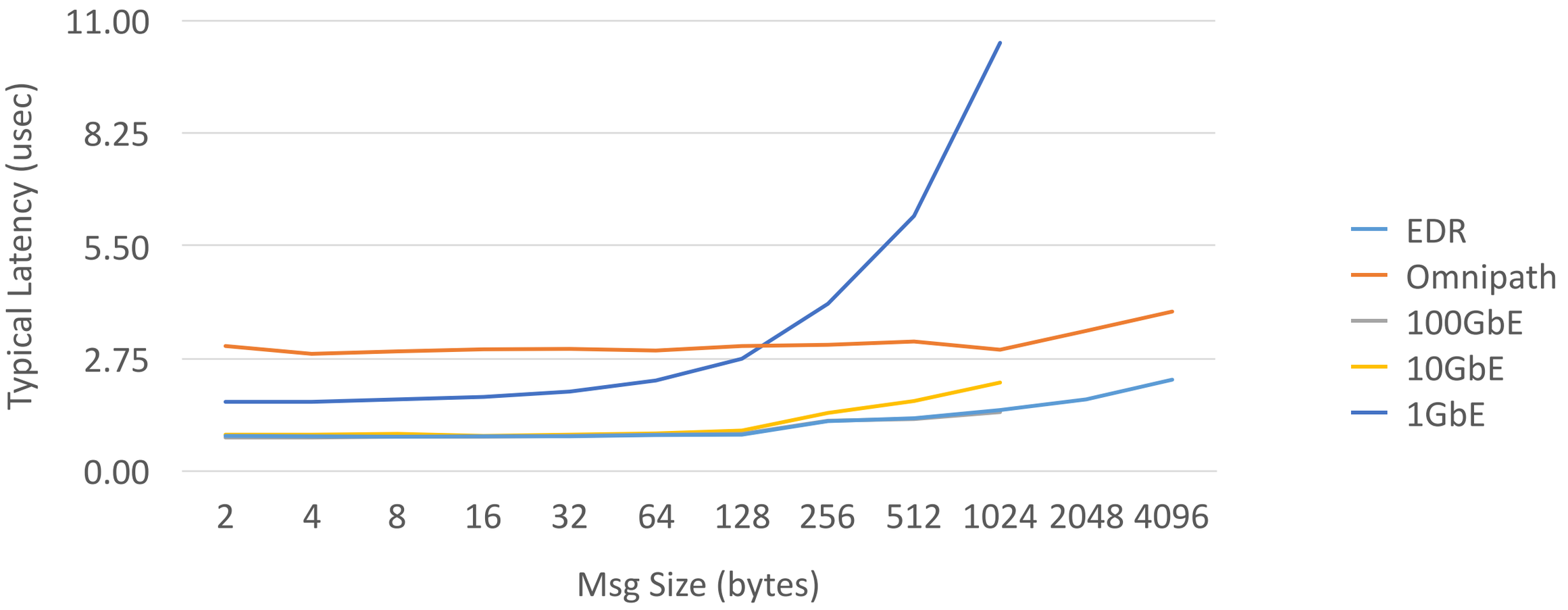
Ethernet via Mellanox ConnectX4 Adapter
- Linux 4.5. Redhat 7.3.
(7.2 has only socket layer support).

Omnipath via Intel OPA adapter
- Out of tree drivers, in Linux 4.4 staging.
Currently supported via Intel OFED distribution

Test Hardware

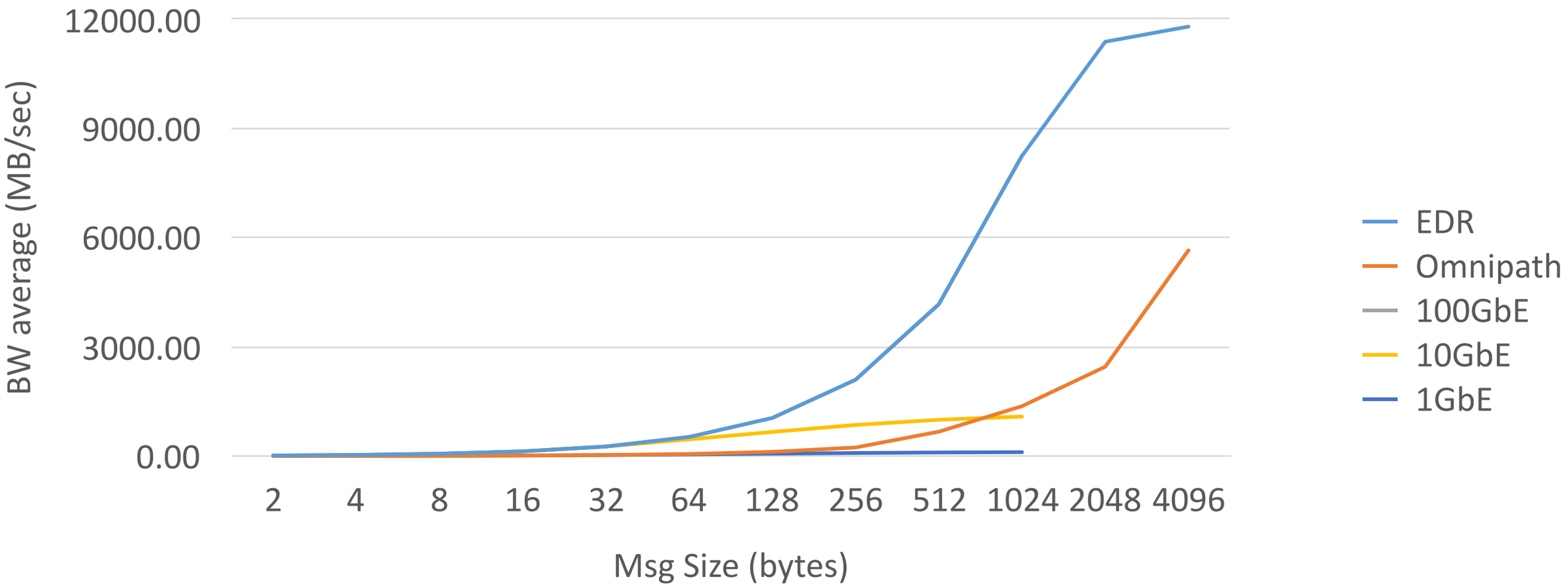
- Intel(R) Xeon(R) CPU E5-2667 v3 @ 3.20GHz
- Adapters
  - Intel Omni-Path Host Fabric Interface Adapter
    - Driver Version: 0.11-162
    - Opa Version: 10.1.1.0.9
  - Mellanox ConnectX-4 VPI Adapter
    - Driver Version: Stock RHEL 7.2
    - Firmware Version: 12.16.1020
- Switches
  - Intel 100 OPA Edge 48p
    - Firmware Version: 10.1.0.0.133
  - Mellanox SB7700
    - Firmware Version: 3.4.3050

Latency Tests via RDMA APIs(ib_send_lat)



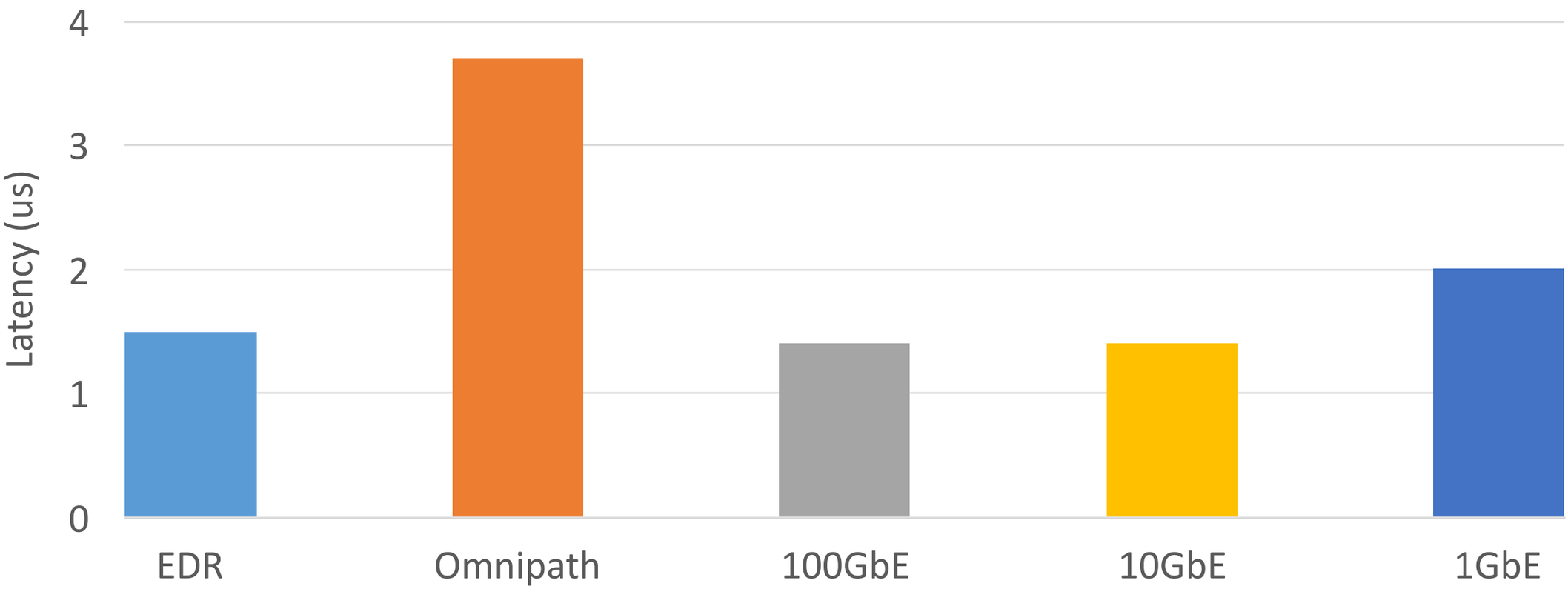
- Consistently low latency below 50% of 1G Ethernet.
- Higher packet sizes benefit significantly.
- Omnipath has higher latency due to software processing of send/receive requests.

Bandwidth Tests using RDMA APIs (ib_send_bw)



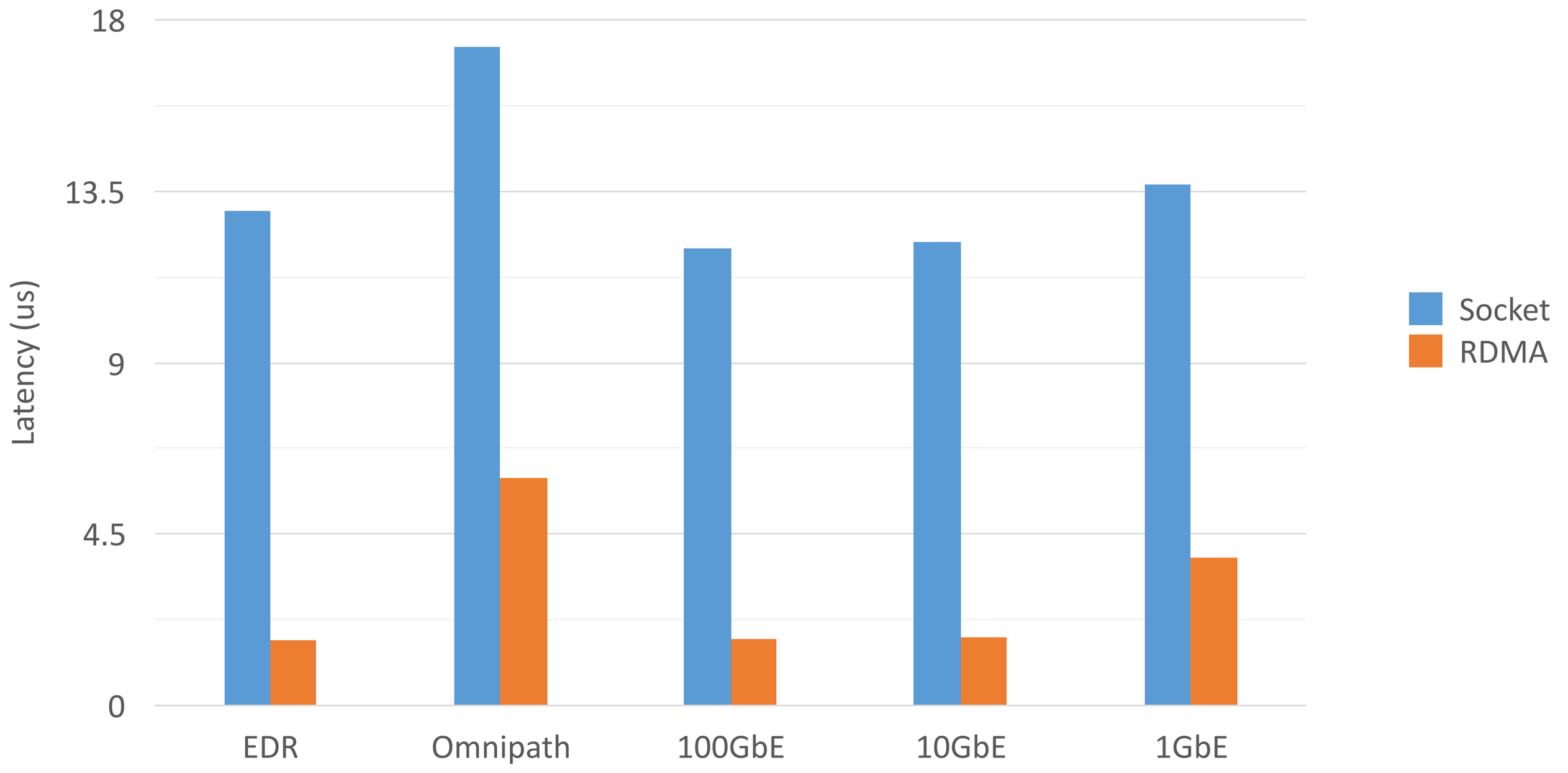
- EDR can reach line saturation (~12GB/sec) at ~ 2k packet size
- Small packet processing is superior on EDR.
- 10GE (1GB/sec) and 1G (120GB/sec) saturate the line with small packets early

Multicast latency tests

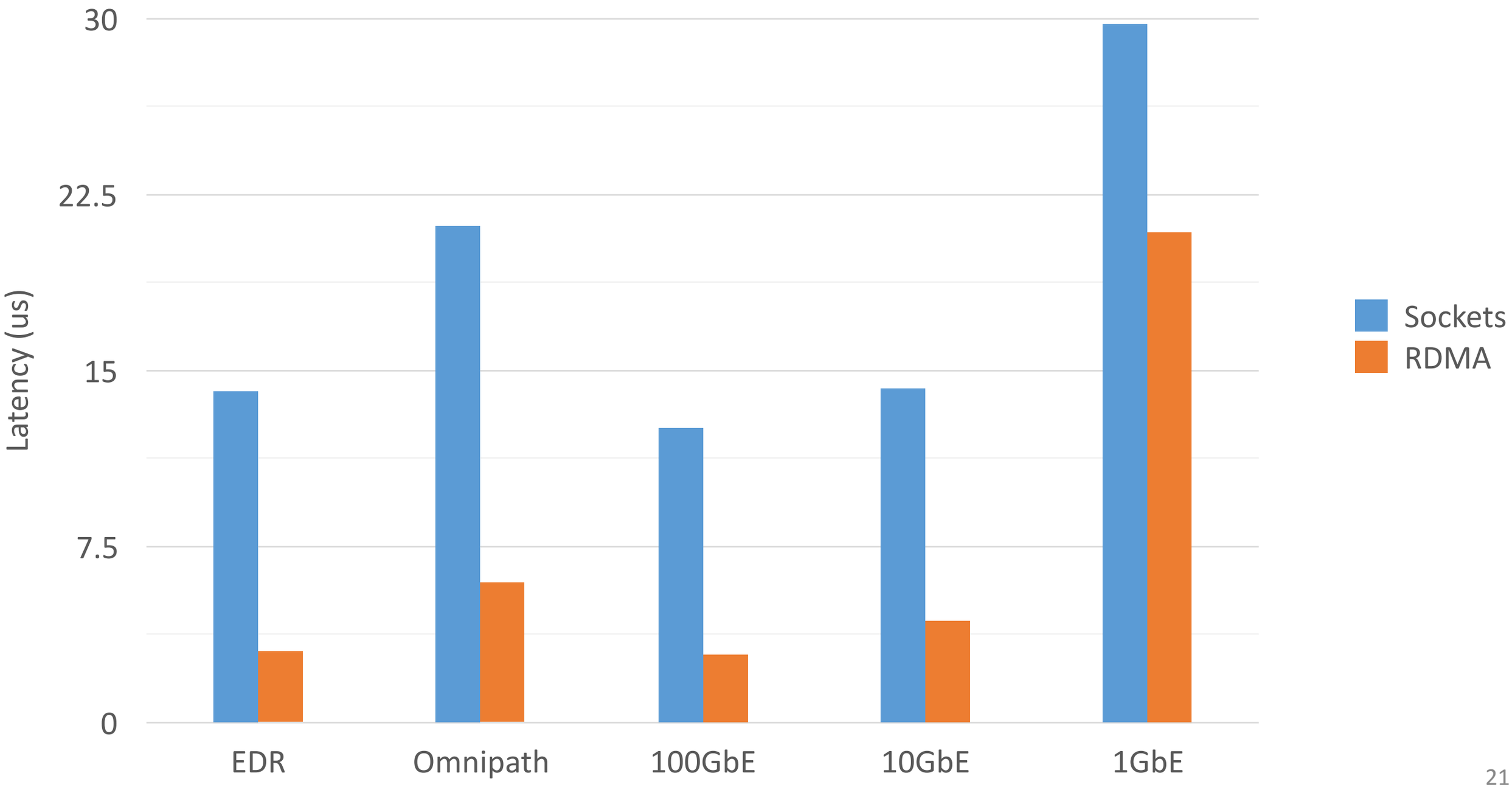


- Lowest latency with 100G and 10G Ethernet
- Slightly higher latency of EDR due to Forward Error Correction
- Software processing increases packet latency on Omnipath

RDMA vs. Posix Sockets (30 byte payload)



RDMA vs. Posix Sockets (1000 byte Payload)



Further Reading Material

http://presentations.interop.com/events/las-vegas/2015/open-to-all---keynote-presentations/download/2709

https://en.wikipedia.org/wiki/100_Gigabit_Ethernet

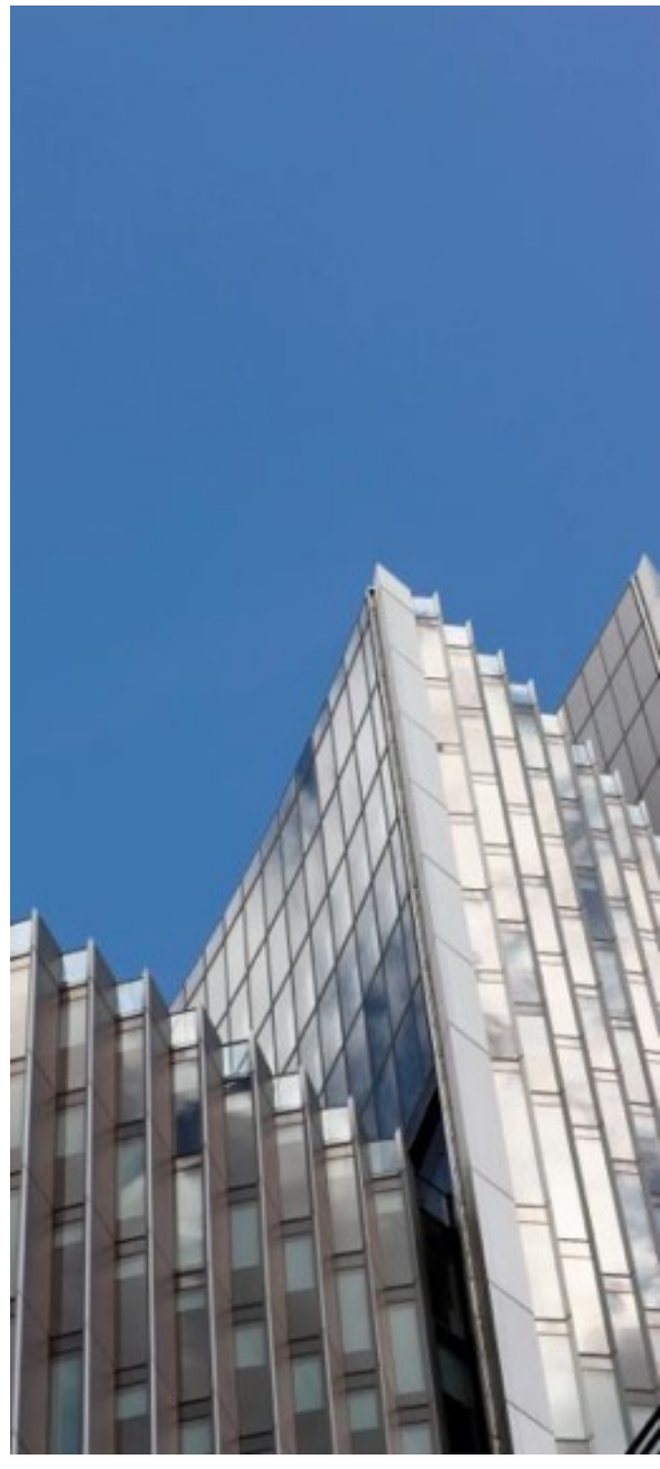
http://www.ieee802.org/3/index.html

Memory Performance issues with 100G

- 100G NIC can give you 12.5G Byte per second of throughput
- DDR3 memory in basic configuration at 6.5 Gbyte per sec. High end at 17G byte per second.
- DDR4 12.8G - 19.2G byte per sec.
- Some adapter have dual 100G connectors.
- Memory via the NIC traffic may be able to saturate the system.

Looking Ahead

- 100G is maturing.
- 200G available in 2017/2018.
- Terabit links by 2022.
- Software needs to mature. Especially the OS network stack to handle these speeds.
- Issues
  - Memory throughput
  - Proper APIs
  - Deeper integration of cpu/memory/io



Q&A

- Issues
- Getting involved
- How to scale the OS and software
- What impact will this speed have on software
- Contact information

  cl@linux.com