Overview of RDMA on Windows

Wenhao Wu
Program Manager
Windows HPC team

Agenda

- Microsoft’s Commitments to HPC
- RDMA for HPC Server
- RDMA for Storage in Windows 8
- Microsoft Talks in OFA Theatre

Microsoft’s Commitments to HPC

2004年 Beginning of HPC Journey

2006 V1

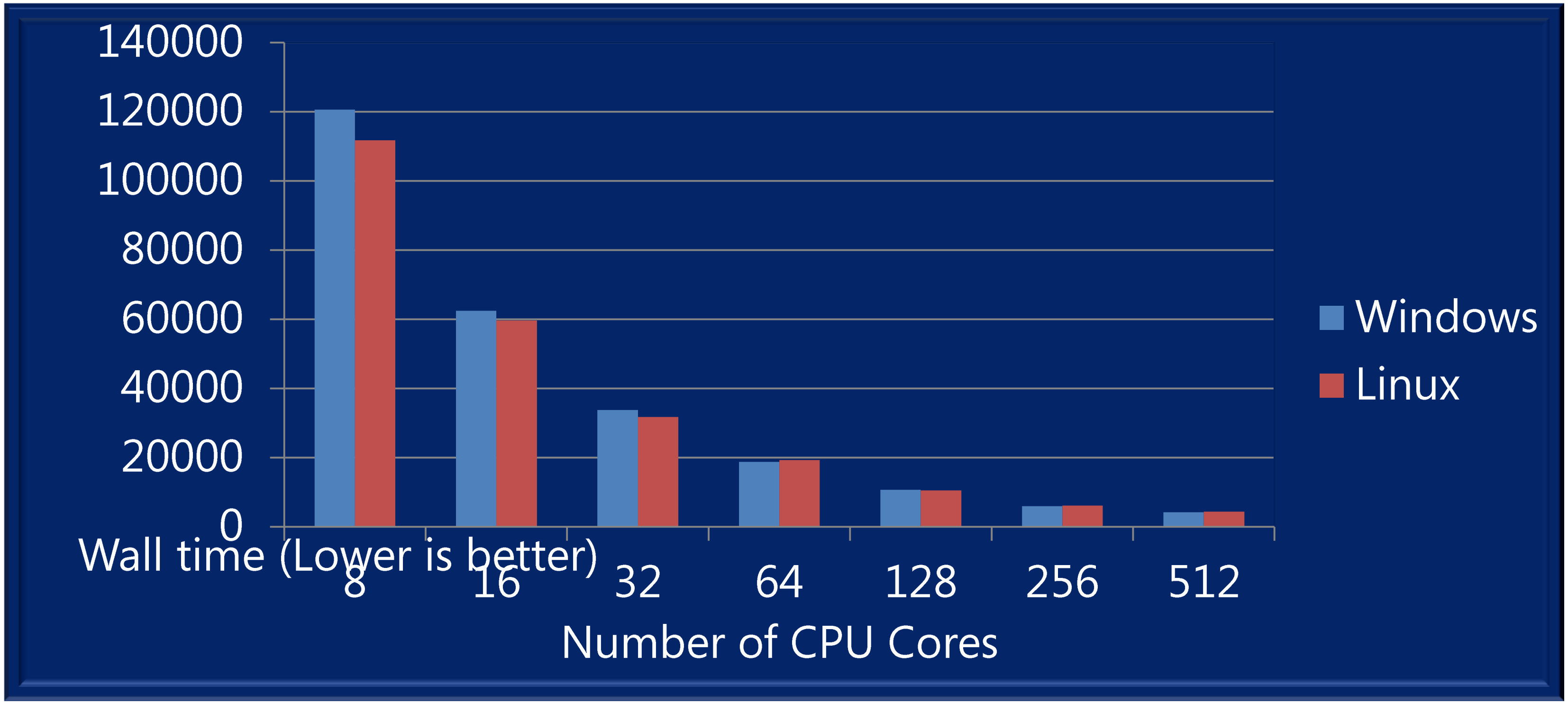
2008 HPC Server 2008

2010 HPC Server 2008 R2

2010-10 SP1 和 SP2

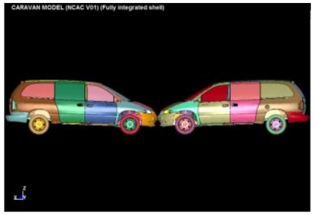
Performance and Scale

Windows Matches Linux Performance on LSTC LS-DYNA®



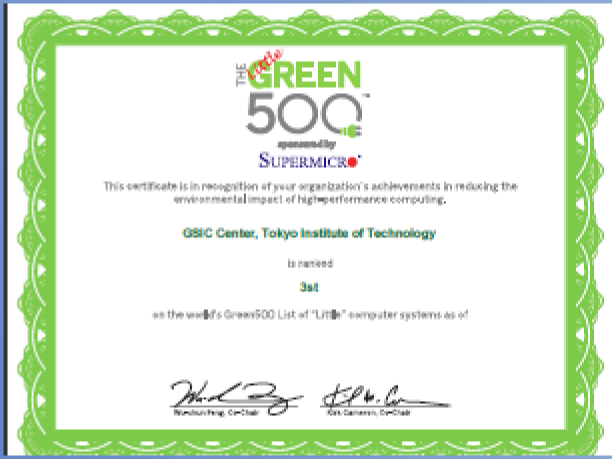
Reference:

- Dataset is car2car, public benchmark. LSTC LS-DYNA data to be posted at http://www.topcrunch.org. LS-DYNA version mpp971.R3.2.1.
- Similar hardware configuration was used for both Windows HPC Server 2008 and Linux runs: Windows HPC: 2.66GHz Intel® Xeon® Processor X5550, 24GB DDR3 memory per node, QDR InfiniBand Linux: 2.8GHz Intel® Xeon® Processor X5560, 18GB DDR3 memory per node, QDR InfiniBand



Windows HPC capture #3 and #5 spots on Green500

- Little Green500 strives to raise awareness in the energy efficiency of supercomputing
- Smaller TSUBAME 2.0 running Windows improves the efficiency from 958.35 Mflops/Watt on the Green500 to 1,031.92 on Little Green500
- CASPUR running Windows become the 1st European system on Little Green500



THE Little GREEN 500

sponsored by SUPERMICRO

This certificate is in recognition of your organization's achievements in reducing the environmental impact of high-performance computing.

GSIC Center, Tokyo Institute of Technology

is ranked

3st

on the world's Green500 List of "Little" computer systems as of

Wu-chun Feng, Co-Chair

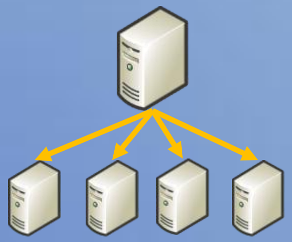
Kirk Cameron, Co-Chair

Little Green500 List - November 2010*

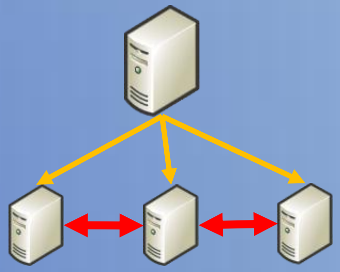
Rank	System Description	Vendor	OS	MFLOPS/ Watt	Total (kw)
1	NNSA/SC Blue Gene/Q Prototype	IBM	LINUX	1684.20	39
2	GRAPE-DR accelerator Cluster	NAOJ	LINUX	1448.03	25
3	TSUBAME 2.0 - HP ProLiant GP/GPU	NEC/HP	Windows	1031.92	26
4	EcoG	NCSA	LINUX	1031.92	37
5	CASPUR-Jazz Cluster GP/GPU	Clustervision/HP	Windows	933.06	26

* Source: http://www.green500.org/lists/2010/11/little/list.php

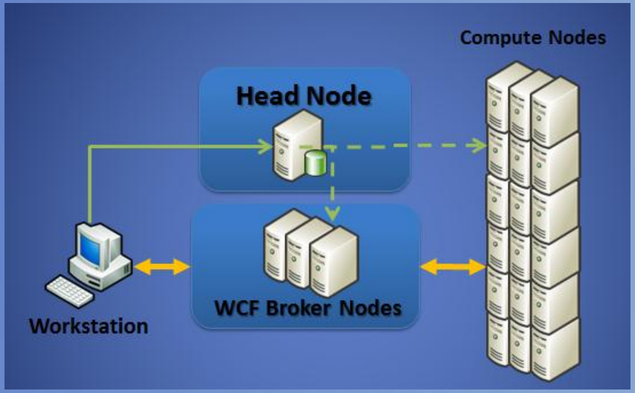
Parallel Applications patterns on HPC Server



- Embarrassingly Parallel
  - Parametric Sweep Jobs
    - CLI**: job submit /parametric:1-1000:5 MyApp.exe /MyDataset=*
    - Search: hpc job scheduler @ http://www.microsoft.com/downloads



- Message Passing
  - MPI Jobs
    - CLI**: job submit /numcores:64 mpiexec MyApp.exe
    - Search: hpc mpi



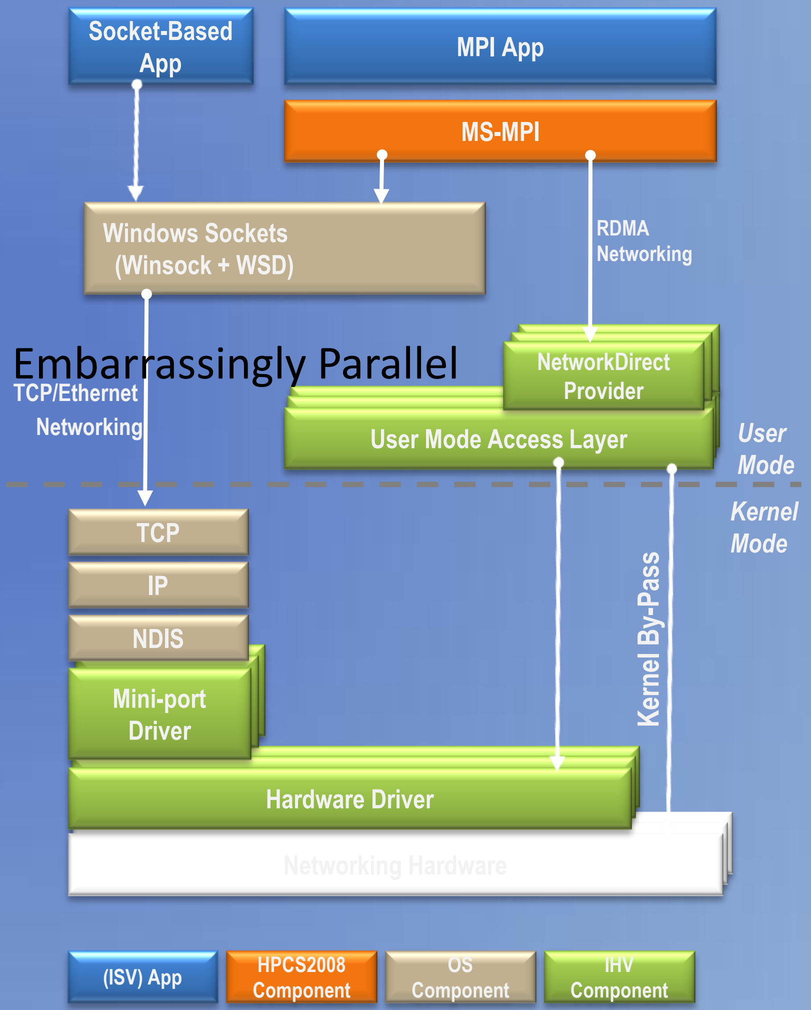
- Interactive Applications
  - Service-Oriented Architecture (SOA) Jobs
    - .NET call to a Windows Communications Foundation (WCF) endpoint
    - Search: hpc soa

- Big Data
  - LINQ-To-HPC (L2H) Jobs
    - HPC application calls L2H .Net APIs
    - Search: hpc dryad

** CLI = HPC Server 2008 Command Line Interface

NetworkDirect

- **Verbs-based design** for native, high-perf networking interfaces
- **Equal to Hardware-Optimized stacks** for MPI
- NetworkDirect drivers for key high-performance fabrics:
  - Infiniband
  - 10 Gigabit Ethernet (iWARP-enabled)



Socket-Based App

MPI App

MS-MPI

Windows Sockets (Winsock + WSD)

RDMA Networking

Embarrassingly Parallel

TCP/Ethernet Networking

NetworkDirect Provider

User Mode Access Layer

User Mode

Kernel Mode

TCP

IP

NDIS

Mini-port Driver

Kernel By-Pass

Hardware Driver

Networking Hardware

(ISV) App

HPCS2008 Component

OS Component

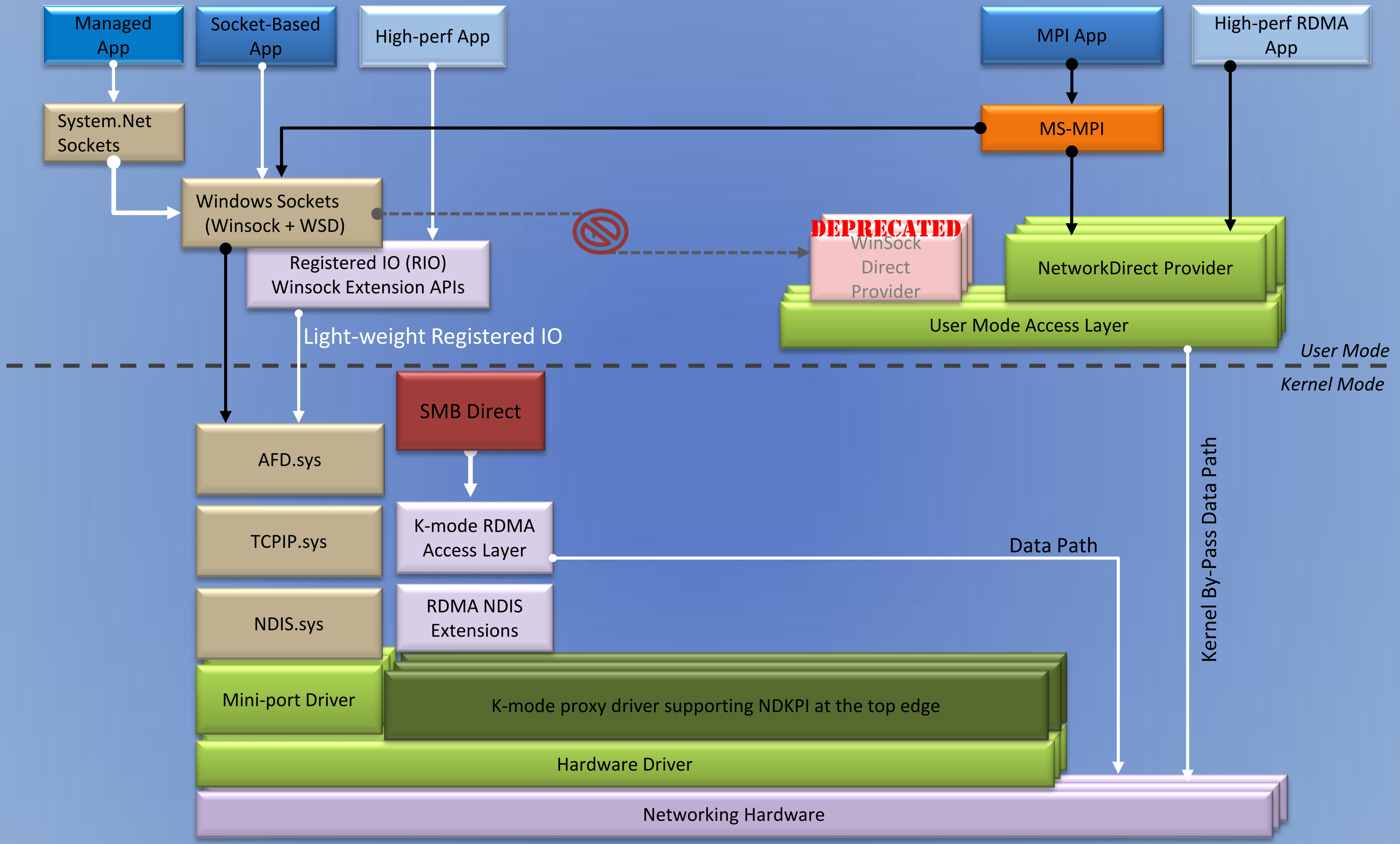
IHV Component

Scaling the network traffic

- What is left to solve?
  - – CPU Utilization and throughput concerns
    - Large I/Os – CPU utilization at high data rates (throughput is great)
    - Small I/Os – CPU utilization and network throughput at high data rates
- Solution: Remote Direct Memory Access (RDMA)

RDMA is "Remote Direct Memory Access" – a secure way to enable a DMA engine to transfer buffers

Windows 8 RDMA Networking Architecture



Managed App

Socket-Based App

High-perf App

MPI App

High-perf RDMA App

System.Net Sockets

MS-MPI

Windows Sockets (Winsock + WSD)

Registered IO (RIO) Winsock Extension APIs

DEPRECATED

WinSock Direct Provider

NetworkDirect Provider

User Mode Access Layer

Light-weight Registered IO

User Mode

Kernel Mode

SMB Direct

AFD.sys

TCPIP.sys

NDIS.sys

K-mode RDMA Access Layer

RDMA NDIS Extensions

Data Path

Kernel By-Pass Data Path

Mini-port Driver

K-mode proxy driver supporting NDKPI at the top edge

Hardware Driver

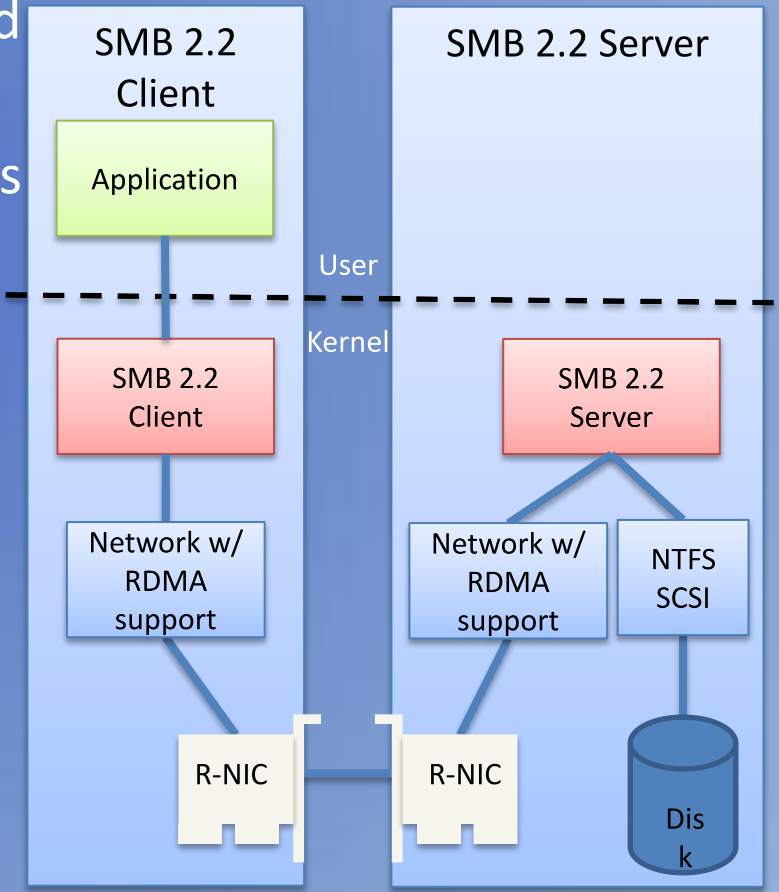
Networking Hardware

Legend

(ISV) App	HPCS2008 Component	OS Component	IHV Component today	New IHV Component s	New OS Component s	K-mode RDMA Component s	New High-perf app

SMB2 Direct (SMB2 over RDMA)

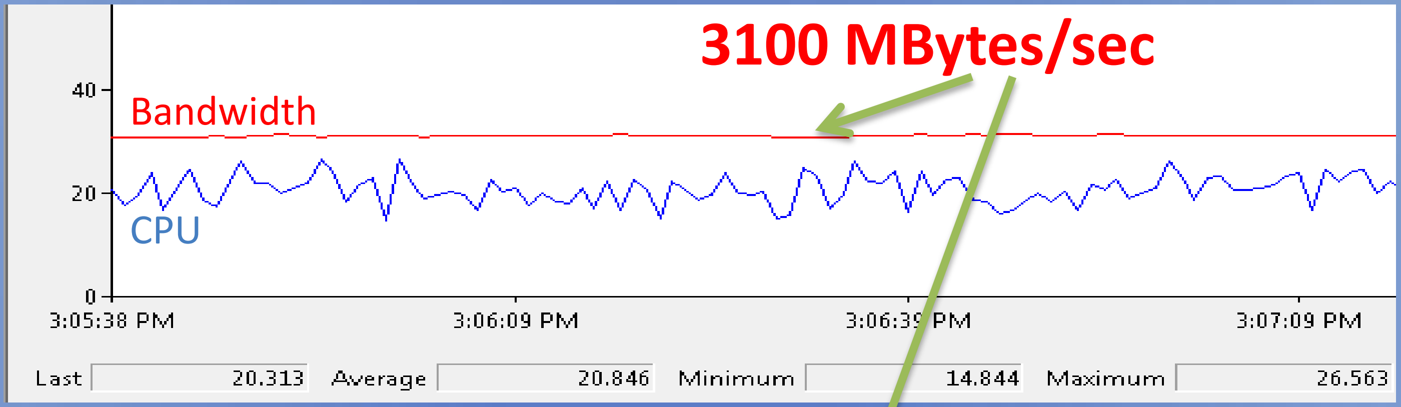
- Used by File Server and Clustered Shared Volumes
- Scalable, fast and efficient storage access
- Minimal CPU utilization for I/O
- High throughput with low latency
- Required hardware
  - InfiniBand
  - 10G Ethernet w/ RDMA
    - iWARP – RDMA on top of TCP
    - ROCE (RDMA Over Ethernet)



SMB Direct 2.2 over RDMA

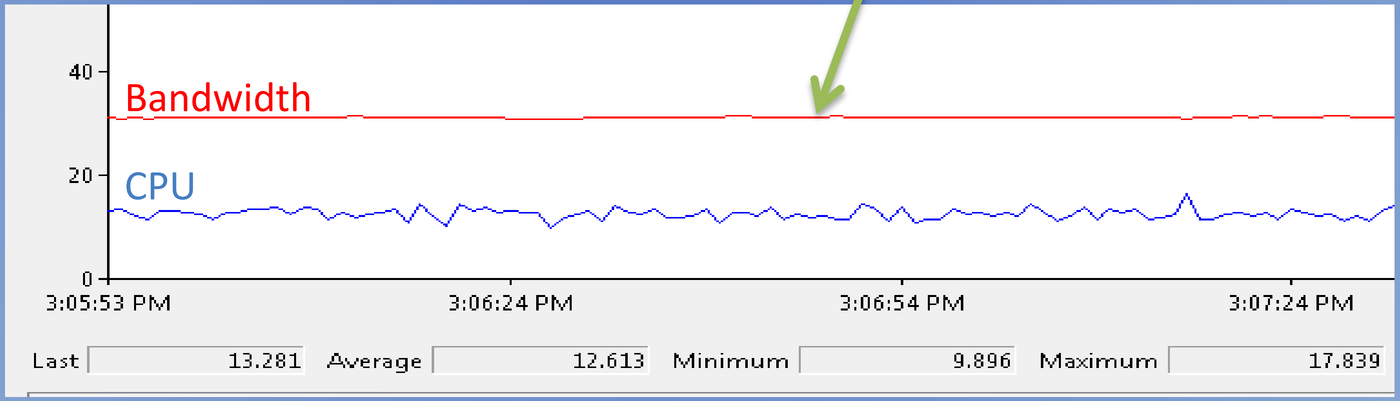
Preliminary 512 KB I/O Results:

Client: ~20% of 4 cores, or 80% of one core



Last 20.313 Average 20.846 Minimum 14.844 Maximum 26.563

Server: ~12% of 12 cores, or 140% of one core



Last 13.281 Average 12.613 Minimum 9.896 Maximum 17.839

Microsoft Talks in OFA Theatre

Time	Topic	Presenter	Title
Tue 1:15-1:45 pm	Overview of RDMA on Windows	Wenhao Wu	Program Manager
Tue 2:15 - 2:45 pm	Windows HPC Update	Greg Burgess	Development Manager
Wed 2:15-2:45 pm	SMB 2.2 Over RDMA	Dan Lovinger	Program Manger