Windows OpenFabrics (WinOF) Update

OPENFABRICS ALLIANCE

Eric Lantz, Microsoft

(elantz@microsoft.com)

April 2008

Agenda

- OpenFabrics and Microsoft
- Current Events
- HPC Server 2008 Release
- NetworkDirect - RDMA for Windows

OpenFabrics and Microsoft

- Value to the Microsoft HPC team
  - Focus our dev/test resources on a single code base
  - Single Win stack simpler for OEMs & Customers
    - Removes adoption roadblock for high-perf networking
- Value to OpenFabrics
  - Advocate/Liaison into Microsoft
    - The HPC team are the high-perf “activists” within MS
  - Insight to Windows environment details to leverage:
    - Win Architecture, Win Mgmt Infrastructure, MS Support team, MS marketing machine for high-perf networking
  - Fab!

Microsoft HPC Current Events

- Actively working with OFA’s WinOF stack
  - Hundreds of machines run WinOF each week
  - Discovering issues in kernel drivers, IPoIB driver, OpenSM, compat with other SM’s
    - Working Contribution Agreement with OFA LWG to enable direct contribution of fixes
- Early stages of membership discussion
- Focus on HPC Server 2008 release

HPC Server 2008

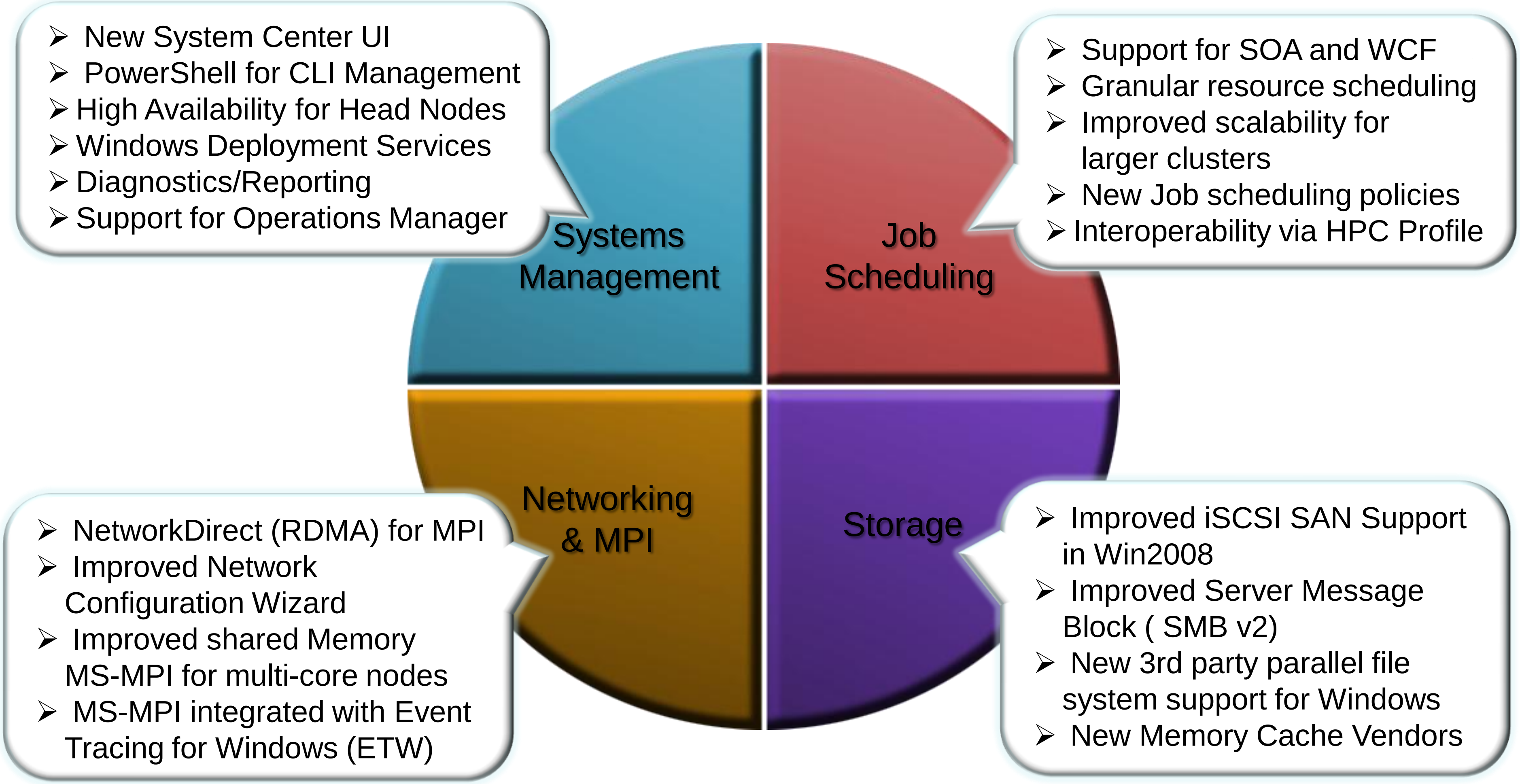
- Release in Summer 2008
- Committed to RDMA Networking for Windows Clusters → NetworkDirect (verbs for Windows)
  - See HPC Server 2008 SDK (Apr posting) on MS Connect website for docs & test apps (w/ source): https://connect.microsoft.com/default.aspx
  - Logo requirements posted for comment in WinQual's Logopoint: http://www.microsoft.com/whdc/winlogo/LogoTools.mspx
  - Reviewed individually with IB, iWARP, and other vendors
- New Mgmt Infrastructure
  - Configuration Database = SAME (System Definition Model)
  - User Interface = NEW (System Center framework)
  - Node Imaging = NEW (Windows Deployment Services)

HPC Server 2008 (continued)

- Node Imaging = NEW (Windows Deployment Services)
  - Requires INF-based installer
  - HPCS2008 layers **Node Templates** onto WDS
    - Inject drivers into OS images
    - Arbitrary command line execution(s)
    - Apply named templates to any number of compute nodes in a couple “clicks”

- Apply OFA WinOF drivers to any number of compute nodes in a couple “clicks”!!

What’s new in HPC Server 2008?



Systems Management

- New System Center UI
- PowerShell for CLI Management
- High Availability for Head Nodes
- Windows Deployment Services
- Diagnostics/Reporting
- Support for Operations Manager

Job Scheduling

- Support for SOA and WCF
- Granular resource scheduling
- Improved scalability for larger clusters
- New Job scheduling policies
- Interoperability via HPC Profile

Networking & MPI

- NetworkDirect (RDMA) for MPI
- Improved Network Configuration Wizard
- Improved shared Memory MS-MPI for multi-core nodes
- MS-MPI integrated with Event Tracing for Windows (ETW)

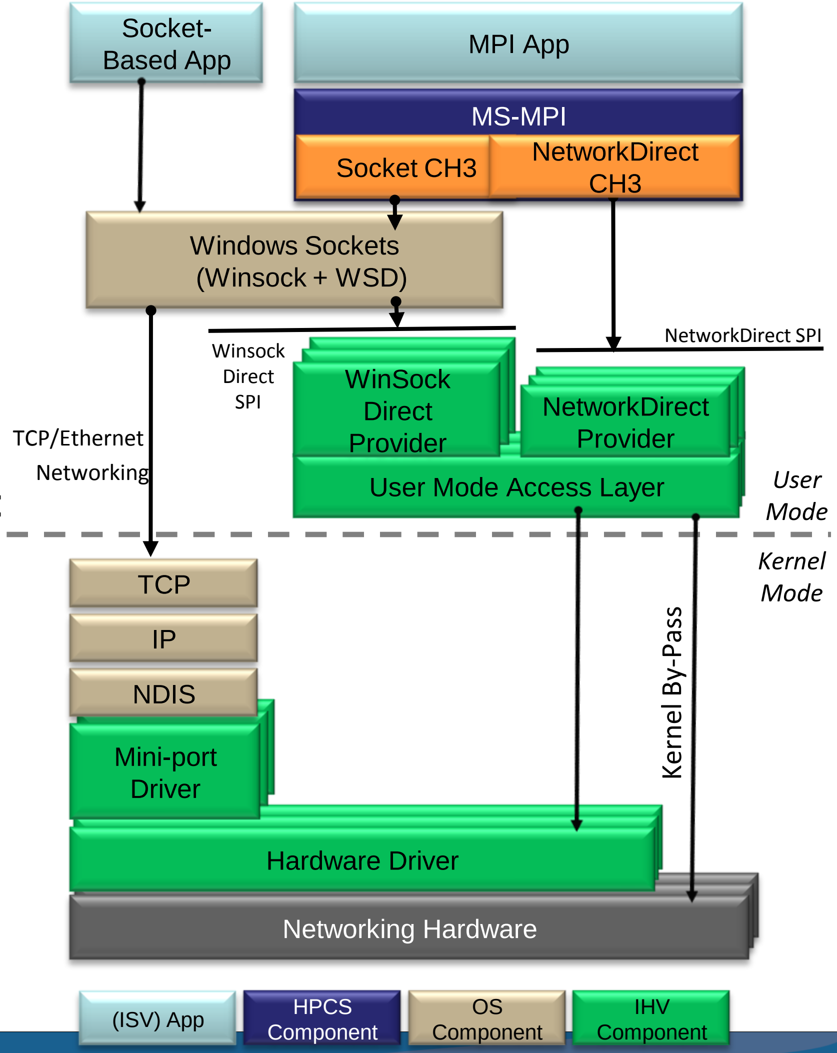
Storage

- Improved iSCSI SAN Support in Win2008
- Improved Server Message Block (SMB v2)
- New 3rd party parallel file system support for Windows
- New Memory Cache Vendors

NetworkDirect

A *new* RDMA networking interface built for speed and stability

- Priorities
  - Equal to Hardware-Optimized stacks for *MPI micro-benchmarks*
    - Focus on **MPI-Only Solution for HPCS 2008**
  - Verbs-like design for close fit with native, high-perf networking interfaces
  - Coordinated w/ Win Networking team's long-term plans
- Implementation
  - MS-MPIv2 capable of 4 networking paths:
    - Shared Memory
      between processors on a motherboard
    - TCP/IP Stack ("normal" Ethernet)
    - Winsock Direct (and SDP)
      for sockets-based RDMA
    - NetworkDirect interface
  - HPC team partnering with networking IHVs to develop/distribute drivers for this new interface



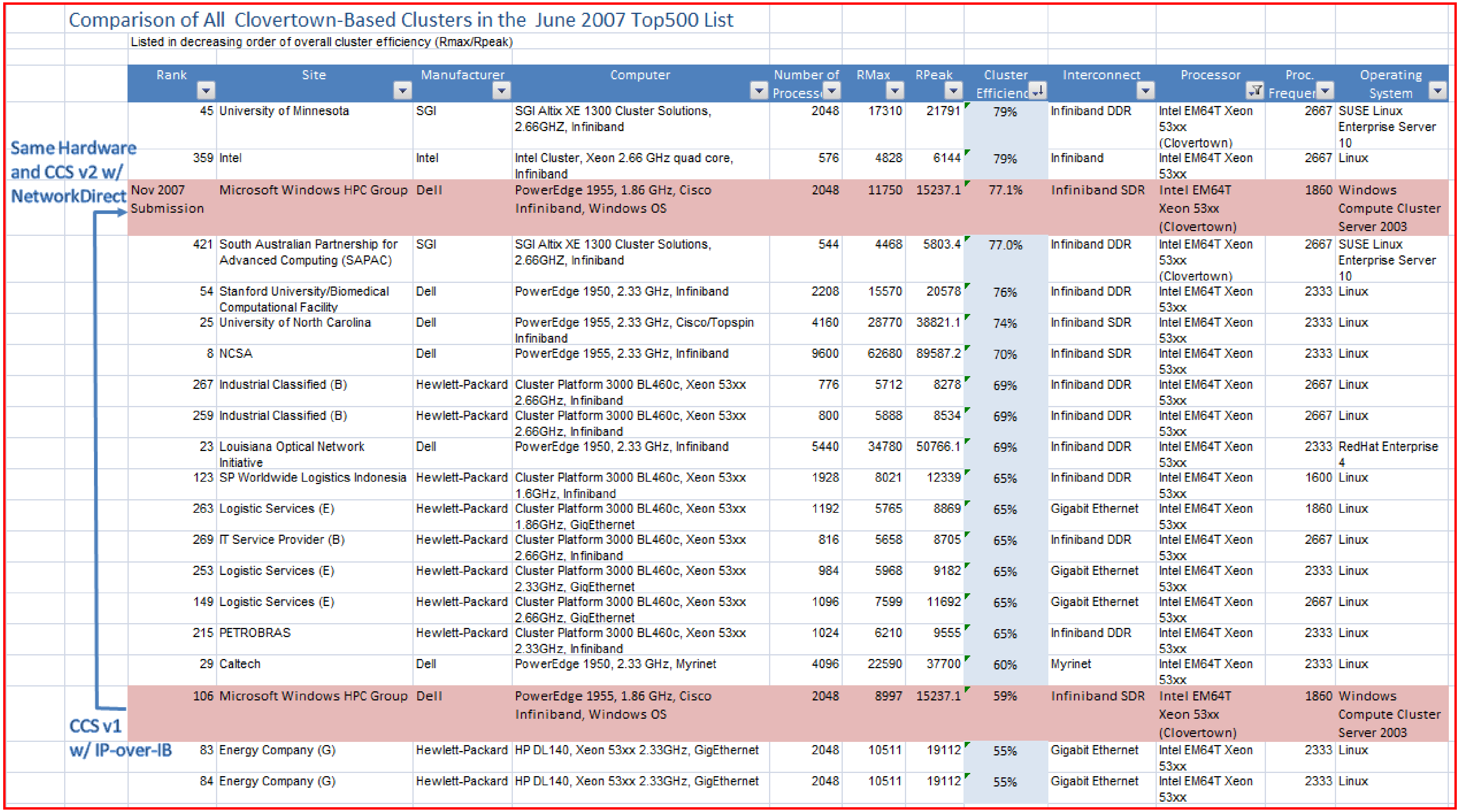
NetworkDirect is…**Verbs for MS-MPI**

- Defined via a published NetworkDirect Service Provider Interface (SPI)
- Aligned with industry-standard verbs
  - Some changes for simplicity
  - Some changes for work with both Infiniband and iWARP (Ethernet RDMA)
- Windows-centric design
  - Leverage Windows asynchronous I/O capabilities
- Transparent to MPI Applications

What Makes NetworkDirect So Fast?

- The application, in this case MS-MPI [which understands it’s data and messaging patterns], controls communication policies
  - Register/deregister of memory (scatter-gather lists)
  - Association of endpoints with completion queues
  - Association of memory windows with registered memory
  - Endpoint request limits
  - Endpoint scatter/gather limits
  - Completion polling (if/when)
- Wafer Thiiiiinnn layer over native hardware API’s

HPCS2008 Software Boosts Performance By 30% On Microsoft Cluster To Achieve 77.1% Overall Cluster Efficiency



Comparison of All Clovertown-Based Clusters in the June 2007 Top500 List

Listed in decreasing order of overall cluster efficiency (Rmax/Rpeak)

Rank	Site	Manufacturer	Computer	Number of Process	RMax	RPeak	Cluster Efficienc	Interconnect	Processor	Proc. Frequer	Operating System
45	University of Minnesota	SGI	SGI Altix XE 1300 Cluster Solutions, 2.66GHZ, Infiniband	2048	17310	21791	79%	Infiniband DDR	Intel EM64T Xeon 53xx (Clovertown)	2667	SUSE Linux Enterprise Server 10
359	Intel	Intel	Intel Cluster, Xeon 2.66 GHz quad core, Infiniband	576	4828	6144	79%	Infiniband	Intel EM64T Xeon 53xx	2667	Linux
Nov 2007 Submission	Microsoft Windows HPC Group	Dell	PowerEdge 1955, 1.86 GHz, Cisco Infiniband, Windows OS	2048	11750	15237.1	77.1%	Infiniband SDR	Intel EM64T Xeon 53xx (Clovertown)	1860	Windows Compute Cluster Server 2003
421	South Australian Partnership for Advanced Computing (SAPAC)	SGI	SGI Altix XE 1300 Cluster Solutions, 2.66GHZ, Infiniband	544	4468	5803.4	77.0%	Infiniband DDR	Intel EM64T Xeon 53xx (Clovertown)	2667	SUSE Linux Enterprise Server 10
54	Stanford University/Biomedical Computational Facility	Dell	PowerEdge 1950, 2.33 GHz, Infiniband	2208	15570	20578	76%	Infiniband DDR	Intel EM64T Xeon 53xx	2333	Linux
25	University of North Carolina	Dell	PowerEdge 1955, 2.33 GHz, Cisco/Topspin Infiniband	4160	28770	38821.1	74%	Infiniband SDR	Intel EM64T Xeon 53xx	2333	Linux
8	NCSA	Dell	PowerEdge 1955, 2.33 GHz, Infiniband	9600	62680	89587.2	70%	Infiniband SDR	Intel EM64T Xeon 53xx	2333	Linux
267	Industrial Classified (B)	Hewlett-Packard	Cluster Platform 3000 BL460c, Xeon 53xx 2.66GHz, Infiniband	776	5712	8278	69%	Infiniband DDR	Intel EM64T Xeon 53xx	2667	Linux
259	Industrial Classified (B)	Hewlett-Packard	Cluster Platform 3000 BL460c, Xeon 53xx 2.66GHz, Infiniband	800	5888	8534	69%	Infiniband DDR	Intel EM64T Xeon 53xx	2667	Linux
23	Louisiana Optical Network Initiative	Dell	PowerEdge 1950, 2.33 GHz, Infiniband	5440	34780	50766.1	69%	Infiniband DDR	Intel EM64T Xeon 53xx	2333	RedHat Enterprise 4
123	SP Worldwide Logistics Indonesia	Hewlett-Packard	Cluster Platform 3000 BL460c, Xeon 53xx 1.6GHz, Infiniband	1928	8021	12339	65%	Infiniband DDR	Intel EM64T Xeon 53xx	1600	Linux
263	Logistic Services (E)	Hewlett-Packard	Cluster Platform 3000 BL460c, Xeon 53xx 1.86GHz, GigEthernet	1192	5765	8869	65%	Gigabit Ethernet	Intel EM64T Xeon 53xx	1860	Linux
269	IT Service Provider (B)	Hewlett-Packard	Cluster Platform 3000 BL460c, Xeon 53xx 2.66GHz, Infiniband	816	5658	8705	65%	Infiniband DDR	Intel EM64T Xeon 53xx	2667	Linux
253	Logistic Services (E)	Hewlett-Packard	Cluster Platform 3000 BL460c, Xeon 53xx 2.33GHz, GigEthernet	984	5968	9182	65%	Gigabit Ethernet	Intel EM64T Xeon 53xx	2333	Linux
149	Logistic Services (E)	Hewlett-Packard	Cluster Platform 3000 BL460c, Xeon 53xx 2.66GHz, GigEthernet	1096	7599	11692	65%	Gigabit Ethernet	Intel EM64T Xeon 53xx	2667	Linux
215	PETROBRAS	Hewlett-Packard	Cluster Platform 3000 BL460c, Xeon 53xx 2.33GHz, Infiniband	1024	6210	9555	65%	Infiniband DDR	Intel EM64T Xeon 53xx	2333	Linux
29	Caltech	Dell	PowerEdge 1950, 2.33 GHz, Myrinet	4096	22590	37700	60%	Myrinet	Intel EM64T Xeon 53xx	2333	Linux
106	Microsoft Windows HPC Group	Dell	PowerEdge 1955, 1.86 GHz, Cisco Infiniband, Windows OS	2048	8997	15237.1	59%	Infiniband SDR	Intel EM64T Xeon 53xx (Clovertown)	1860	Windows Compute Cluster Server 2003
83	Energy Company (G)	Hewlett-Packard	HP DL140, Xeon 53xx 2.33GHz, GigEthernet	2048	10511	19112	55%	Gigabit Ethernet	Intel EM64T Xeon 53xx	2333	Linux
84	Energy Company (G)	Hewlett-Packard	HP DL140, Xeon 53xx 2.33GHz, GigEthernet	2048	10511	19112	55%	Gigabit Ethernet	Intel EM64T Xeon 53xx	2333	Linux

Same Hardware and CCS v2 w/ NetworkDirect

CCS v1 w/ IP-over-IB

Performance improvement was demonstrated with exactly the same hardware and is attributed to :

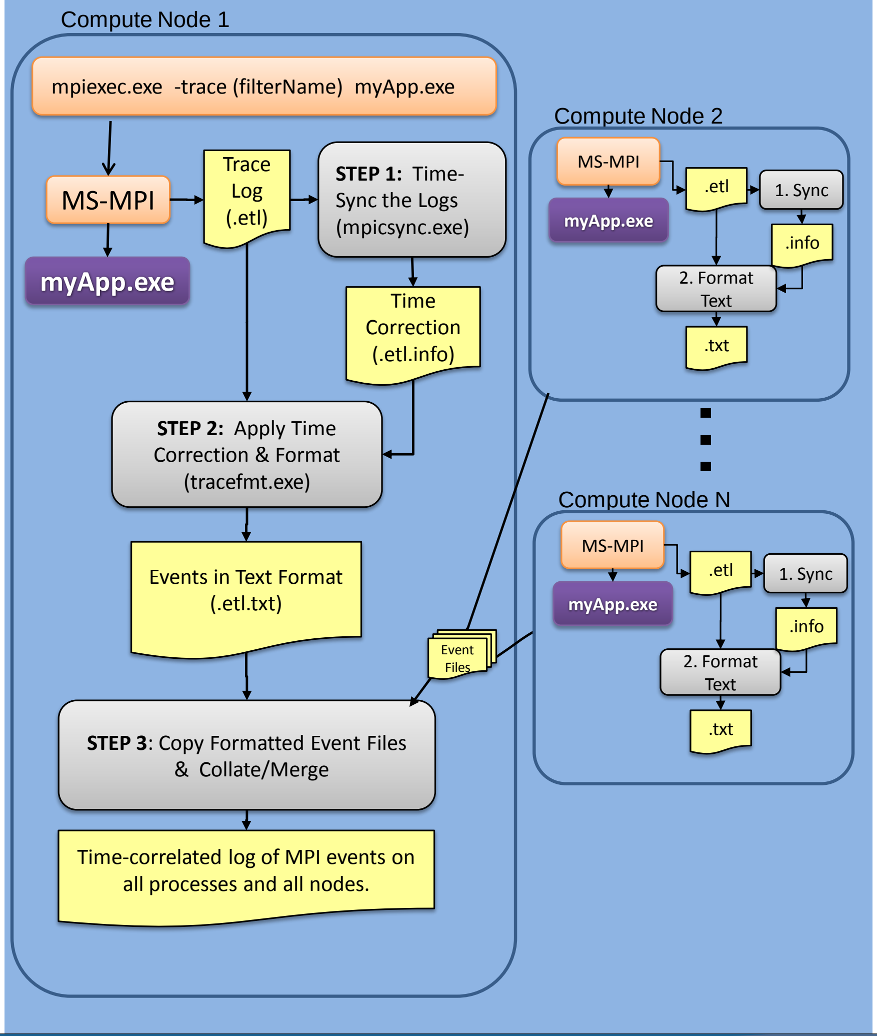
- Improved networking performance of MS-MPI's NetworkDirect interface
- Entirely new MS-MPI implementation for shared memory communications
- Windows Server 2008 improvements in querying completion port status
- Use of Visual Studio's Profile Guided Optimization (POGO) on the Linpack, MS-MPI, and the ND provider binaries
- Tools and scripts to optimize process placement and tune the Linpack parameters for this 256-node, 2048-processor cluster
  - Characterization and Optimization of the 29-parameter Linpack Tuning Space
  - Automated Linpack executions with Excel and HPCS 2008 Scheduler API
  - **>160 continuous execution hours (cluster pegged at 100%) with 0.0 failures using HPC Server 2008 alpha version**

OPENFABRICS ALLIANCE

THANK YOU!

MS-MPI Tracing (w/ ETW)

- Create Traces in Production Environments
- Create time-correlated logs of MPI events from all processes on all nodes running an MPI application
- HPC-specific addition to ETW: High-precision CPU clock correction for MPI (mpicsync)
- Tap into events "live"


Compute Node 1
mpiexec.exe -trace (filterName) myApp.exe
MS-MPI
myApp.exe
Trace Log (.etl)
STEP 1: Time-Sync the Logs (mpicsync.exe)
Time Correction (.etl.info)
STEP 2: Apply Time Correction & Format (tracefmt.exe)
Events in Text Format (.etl.txt)
Event Files
STEP 3: Copy Formatted Event Files & Collate/Merge
Time-correlated log of MPI events on all processes and all nodes.
Compute Node 2
MS-MPI
myApp.exe
.etl
1. Sync
.info
2. Format Text
.txt
Compute Node N
MS-MPI
myApp.exe
.etl
1. Sync
.info
2. Format Text
.txt