

Open Fabrics For Windows Panel



OPENFABRICS
ALLIANCE

Eric Lantz (elantz@microsoft.com)

Senior Program Manager, HPC team, Microsoft Corp.

www.openfabrics.org

Agenda

- Windows HPC Server 2008
 - Goals
 - Using HPC in new ways
 - Proofpoints
 - Performance
 - Case Studies
 - HPC cluster as part of a larger picture
- OFA WinOF is central to MS's HPC efforts
- MS HPC: Topics of Interest

Windows HPC Server 2008

➤ Goals

- Traditional HPC market
 - Everything to get started is in the “box”
- Using HPC in new ways:
 - CCP Games – 40,000 players in single virtual environment
 - Service Oriented Architecture
 - cluster as a web service
 - Workgroup clusters (100's of nodes)
 - Simple install/maintenance
 - Integration with enterprise networks

➤ Released: September 2008

MONTE CARLO SIMULATION FOR A OPTIONS PORTFOLIO

	To horizon	At horizon	Total
Calend Days	14	16	30
Trading Days	10	11	21
σ	0.03968	σ_{T-t} 0.04384	0.083518
$s =$	37.00%	$s \text{ SQR}(T) =$ 0.078003	$K =$ 94
$s \text{ SQR}(t) =$	0.07371	$PV(T) =$ 0.997312	$s =$ 37.26%
$\mu + t =$	0.00236	$PVDiv(T) =$ 1.000000	$r =$ 6.14%

Share Quantities	Stock	Option
	-400	-400

Initial Stock Price \$ 93,900

99% VAR

Iteration No	N(0,s)	Stock P	ln(S/K*PV)	N(d1)	N(d2)	Option Price	Stock Value	Option Value	Portfolio Value
1	0.06422	100.128	0.065844	0.81142	0.78963	\$ 0.840	\$ (40,051)	\$ (336)	\$ (40,387)
2	0.06405	100.111	0.065675	0.81083	0.78900	\$ 0.843	\$ (40,044)	\$ (337)	\$ (40,381)
3	-0.0555	88.828	-0.053896	0.25722	0.23271	\$ 5.951	\$ (35,531)	\$ (2,380)	\$ (37,912)
4	0.17506	111.864	0.176685	0.98939	0.98700	\$ 0.032	\$ (44,746)	\$ (13)	\$ (44,759)
5	-0.1095	84.157	-0.107918	0.08939	0.07744	\$ 9.854	\$ (33,863)	\$ (3,941)	\$ (37,604)
6	-0.0353	90.645	-0.033653	0.34737	0.31902	\$ 4.682	\$ (36,258)	\$ (1,873)	\$ (38,131)
7	0.05468	99.177	0.056306	0.77663	0.75265	\$ 1.035	\$ (39,671)	\$ (414)	\$ (40,085)
8	-0.0966	85.252	-0.094987	0.11925	0.10442	\$ 6.872	\$ (34,101)	\$ (3,549)	\$ (37,650)
9	-0.0149	92.511	-0.013272	0.44783	0.41717	\$ 3.557	\$ (37,005)	\$ (1,423)	\$ (38,427)
10	-0.0211	91.942	-0.019447	0.41671	0.38655	\$ 3.880	\$ (36,777)	\$ (1,552)	\$ (38,329)
11	-0.0486	89.622	-0.045006	0.29530	0.26895	\$ 5.377	\$ (35,849)	\$ (2,151)	\$ (37,999)
12	-0.054	88.962	-0.052395	0.26346	0.23863	\$ 5.852	\$ (35,585)	\$ (2,341)	\$ (37,926)
13	-0.0725	87.330	-0.070906	0.19215	0.17156	\$ 7.114	\$ (34,932)	\$ (2,846)	\$ (37,778)
14	0.08959	102.701	0.091219	0.88656	0.87085	\$ 0.457	\$ (41,080)	\$ (183)	\$ (41,263)
15	-0.065	87.988	-0.063407	0.21950	0.19714	\$ 6.592	\$ (35,195)	\$ (2,637)	\$ (37,832)
16	0.03709	97.448	0.038714	0.70378	0.67628	\$ 1.483	\$ (38,979)	\$ (593)	\$ (39,572)
17	-0.0005	93.851	0.001107	0.52121	0.49010	\$ 2.867	\$ (37,540)	\$ (1,147)	\$ (38,687)
18	0.02913	96.675	0.030755	0.66760	0.63881	\$ 1.725	\$ (38,670)	\$ (690)	\$ (39,360)
19	0.00575	94.442	0.007379	0.55314	0.52217	\$ 2.593	\$ (37,777)	\$ (1,037)	\$ (38,814)

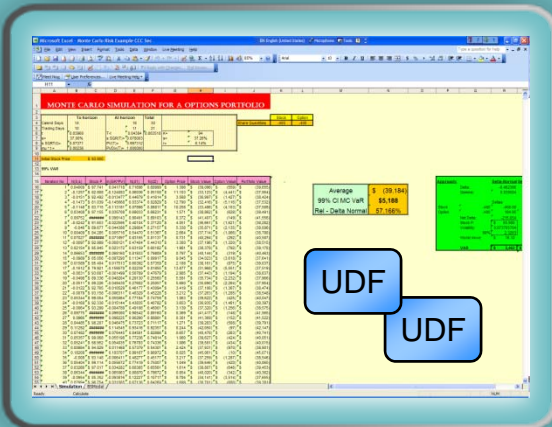
Average 99% CI MC VaR Rel - Delta Normal

\$ (39,234) \$5,427 56.739%

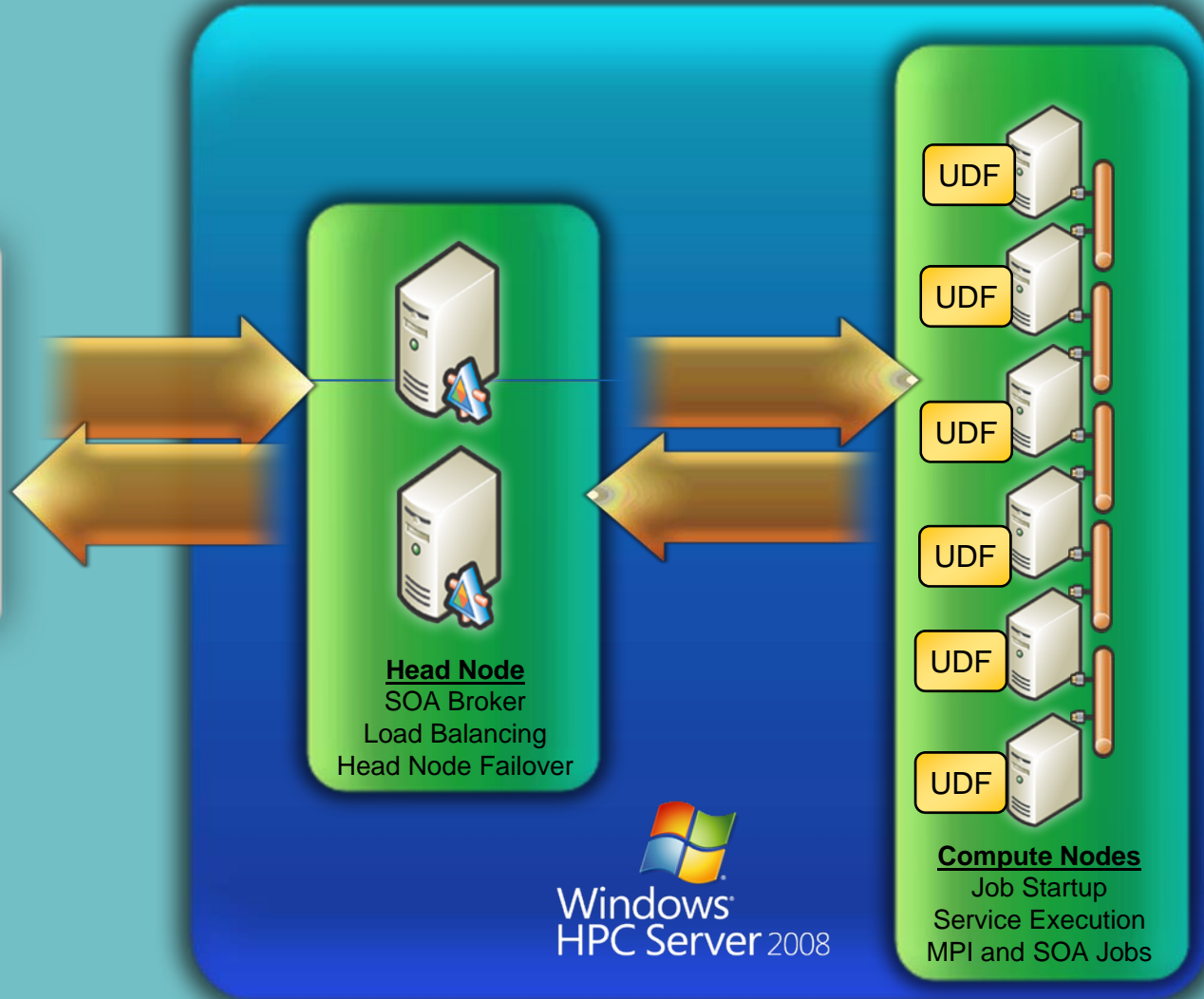
Approach:	Delta-Normal	Delta-Gamma
Delta:	-0.462366	
Gamma:	0.039584	
Deltas		
Stock	-400	-400.00
Option	-400	184.95
		-15.83
Net Delta: -215.054		
Stock P:	\$ 93,900	
Volatility:	0.073705764	
	99%	2.3263
Worst move:	\$ 16.10	
VAR \$ 3,462 \$ 5,515		

59%

65506	65491	-0.0744	87.170	-0.072742	0.18578	0.16584	\$ 7.244	\$ (34,968)	\$ (2,898)	\$ (37,766)
65507	65492	-0.2094	76.159	-0.207782	0.00434	0.00344	\$ 17.596	\$ (30,464)	\$ (7,039)	\$ (37,502)
65508	65493	-0.0552	88.859	-0.053551	0.25865	0.23406	\$ 5.928	\$ (35,544)	\$ (2,371)	\$ (37,915)
65509	65494	-0.0028	93.640	-0.001150	0.50967	0.47857	\$ 2.969	\$ (37,456)	\$ (1,188)	\$ (38,643)
65510	65495	-0.0525	89.096	-0.050888	0.26891	0.24466	\$ 5.754	\$ (35,638)	\$ (2,302)	\$ (37,940)
65511	65496	-0.0443	89.831	-0.042675	0.30569	0.27890	\$ 5.231	\$ (35,932)	\$ (2,092)	\$ (38,025)
65512	65497	-0.0706	87.496	-0.069004	0.19888	0.17784	\$ 6.980	\$ (34,999)	\$ (2,792)	\$ (37,791)
65513	65498	-0.069	87.640	-0.067367	0.20479	0.18336	\$ 6.866	\$ (35,056)	\$ (2,746)	\$ (37,802)
65514	65499	0.10697	104.502	0.108600	0.92382	0.91201	\$ 0.288	\$ (41,801)	\$ (115)	\$ (41,916)
65515	65500	-0.0287	91.245	-0.027056	0.37909	0.34980	\$ 4.300	\$ (36,498)	\$ (1,720)	\$ (38,218)
65516	65501	0.00761	94.618	0.009240	0.56256	0.53166	\$ 2.515	\$ (37,847)	\$ (1,006)	\$ (38,853)
65517	65502	0.00545	94.413	0.007077	0.55161	0.52063	\$ 2.606	\$ (37,765)	\$ (1,042)	\$ (38,808)
65518	65503	0.18563	113.053	0.187253	0.99265	0.99090	\$ 0.022	\$ (45,221)	\$ (9)	\$ (45,230)
65519	65504	-0.0952	85.370	-0.093609	0.12281	0.10786	\$ 8.769	\$ (34,148)	\$ (3,507)	\$ (37,655)
65520	65505	-0.0064	93.300	-0.004786	0.49108	0.46003	\$ 3.139	\$ (37,320)	\$ (1,256)	\$ (38,575)
65521	65506	-0.0732	87.271	-0.071587	0.18977	0.16934	\$ 7.162	\$ (34,908)	\$ (2,865)	\$ (37,773)
65522	65507	-0.0138	92.611	-0.012199	0.45328	0.42254	\$ 3.503	\$ (37,044)	\$ (1,401)	\$ (38,445)
65523	65508	0.12102	105.980	0.122651	0.94645	0.93741	\$ 0.193	\$ (42,392)	\$ (77)	\$ (42,469)
65524	65509	0.04056	97.787	0.042189	0.71900	0.69212	\$ 1.385	\$ (39,115)	\$ (554)	\$ (39,669)
65525	65510	-0.0252	91.565	-0.023558	0.39627	0.36655	\$ 4.104	\$ (36,826)	\$ (1,642)	\$ (38,268)
65526	65511	-0.1047	84.562	-0.103113	0.09976	0.08677	\$ 9.486	\$ (33,825)	\$ (3,795)	\$ (37,620)
65527	65512	0.00344	94.224	0.005068	0.54140	0.51036	\$ 2.692	\$ (37,689)	\$ (1,077)	\$ (38,766)
65528	65513	-0.0981	85.126	-0.096476	0.11549	0.10101	\$ 8.984	\$ (34,050)	\$ (3,594)	\$ (37,644)
65529	65514	-0.0016	93.750	0.000026	0.51569	0.48458	\$ 2.915	\$ (37,500)	\$ (1,166)	\$ (38,666)
65530	65515	0.10638	104.440	0.108009	0.92273	0.91080	\$ 0.293	\$ (41,776)	\$ (117)	\$ (41,893)
65531	65516	-0.0344	90.721	-0.032816	0.35134	0.32286	\$ 4.633	\$ (36,288)	\$ (1,853)	\$ (38,141)
65532	65517	0.03513	97.257	0.036756	0.69505	0.66721	\$ 1.540	\$ (38,903)	\$ (616)	\$ (39,519)
65533	65518	-0.1193	83.344	-0.117630	0.07091	0.06093	\$ 10.602	\$ (33,337)	\$ (4,241)	\$ (37,578)
65534	65519	0.00603	94.468	0.007659	0.55456	0.52360	\$ 2.581	\$ (37,787)	\$ (1,033)	\$ (38,820)
65535	65520	-0.0399	90.224	-0.038312	0.32558	0.29800	\$ 4.962	\$ (36,089)	\$ (1,985)	\$ (38,074)
65536	65521	-0.0892	85.887	-0.087576	0.13924	0.12267	\$ 8.320	\$ (34,355)	\$ (3,328)	\$ (37,682)

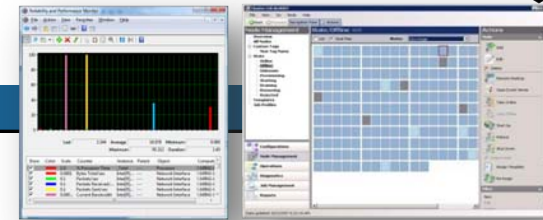


Microsoft[®]
Office Excel.2007

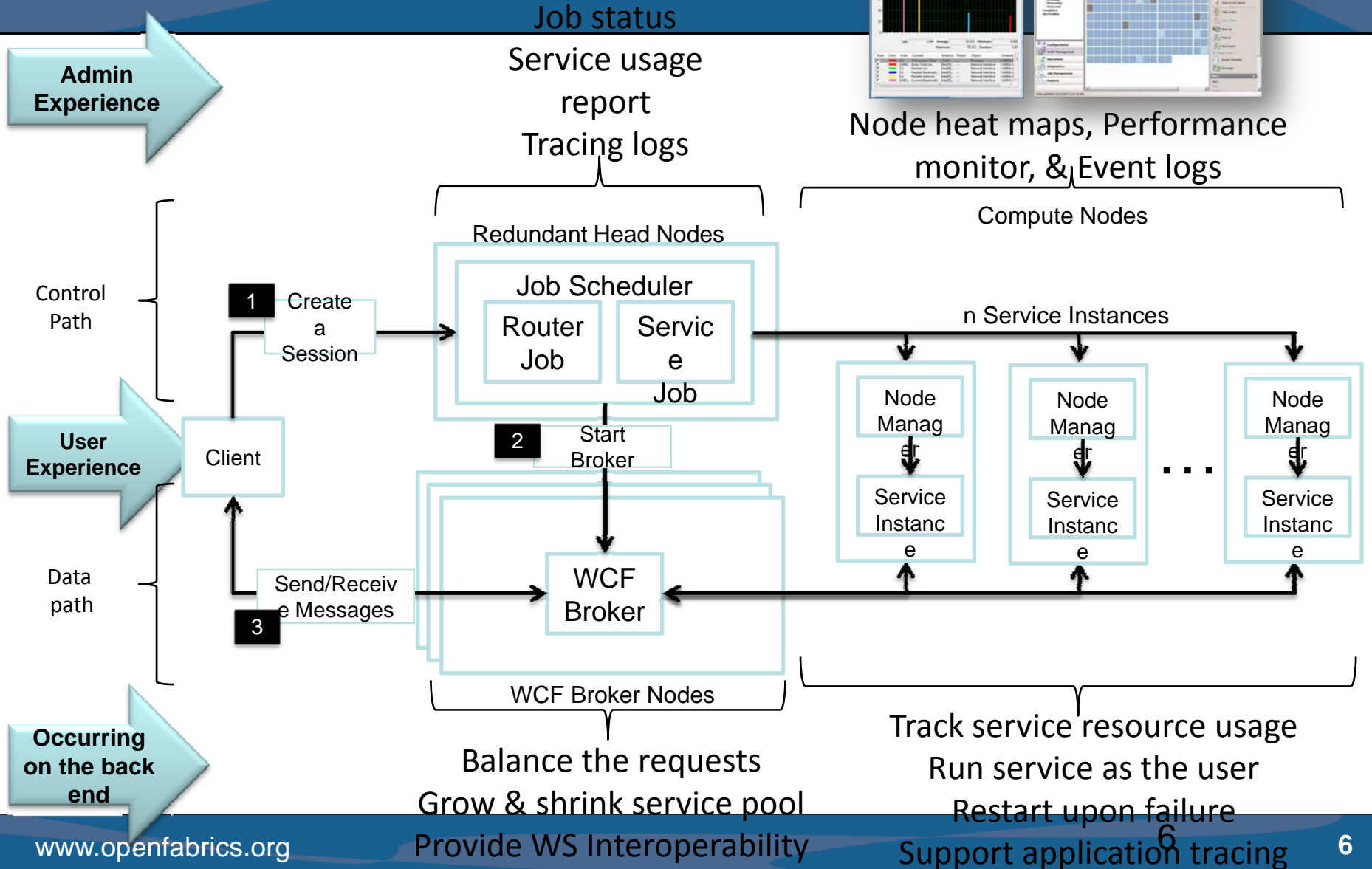


Windows[®]
HPC Server 2008

Service-Oriented Jobs



Node heat maps, Performance monitor, & Event logs



```
string headNode = Environment.GetEnvironmentVariable("CCP_SCHEDULER");
if (headNode.Equals("") == true)
    headNode = "localhost";

SessionStartInfo info = new SessionStartInfo(headNode, "AsianOptionsService");

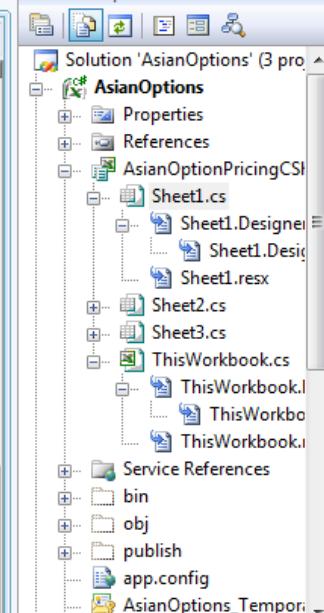
info.ResourceUnitType = JobUnitType.Core;
info.MinimumUnits = 44;
info.MaximumUnits = 44;
info.Secure = false;
info.BrokerSettings.SessionIdleTimeout = 12 * 60 * 60; // 12 hours
info.TransportScheme = TransportScheme.NetTcp;

Session session = Session.CreateSession(info);
Client = new Service1Client(new NetTcpBinding(SecurityMode.None, false), session.EndpointReference);
client.InnerChannel.OperationTimeout = new TimeSpan(1, 0, 0);

foreach (string col in cols)
{
    if (runOnCluster.Equals("Yes"))
    {
        client.BeginPriceAsianOptions(initial, exercise, up, down, interest, periods, runs,
#region callback
        (IAsyncResult result) =>
        {
            double price = client.EndPriceAsianOptions(result);
            this.Range[(string)result.AsyncState, missing].Value2 = price;

            if (count == cols.Length)
                finishedEvt.Set();
        },
#endregion
        );
    }
    else
    {
        double price = PriceAsianOptions(initial, exercise, up, down, interest, periods, runs);
        this.Range[string.Format("{0}{1}", col, i), missing].Value2 = price;
    }
}

if (runOnCluster.Equals("Yes"))
    finishedEvt.WaitOne();
```



Windows HPC Server 2008

➤ Performance

➤ #10: Shanghai Supercomputer Center, Shanghai, China

180.6 TeraFLOPS on 31,200 cores at 77.5% efficiency - with commodity hardware.

➤ #23: National Center for Supercomputing Applications, Illinois, USA

68.5 TeraFLOPS on 9,472 cores at 77.7% efficiency

NetworkDirect ran hour-after-hour at full scale while we tuned.

#40: UMEA University, Sweden

46 TeraFLOPS on 5,376 cores at 85.5% efficiency

Best efficiency score at the time for an x86 architecture cluster on the Top 500 list- regardless of Operating System.

➤ #100: Aachen University, Germany

18.8 TeraFLOPS on 2,096 cores at 76.5% efficiency

Matched the best Linux efficiency on this cluster but with simpler cluster mgmt

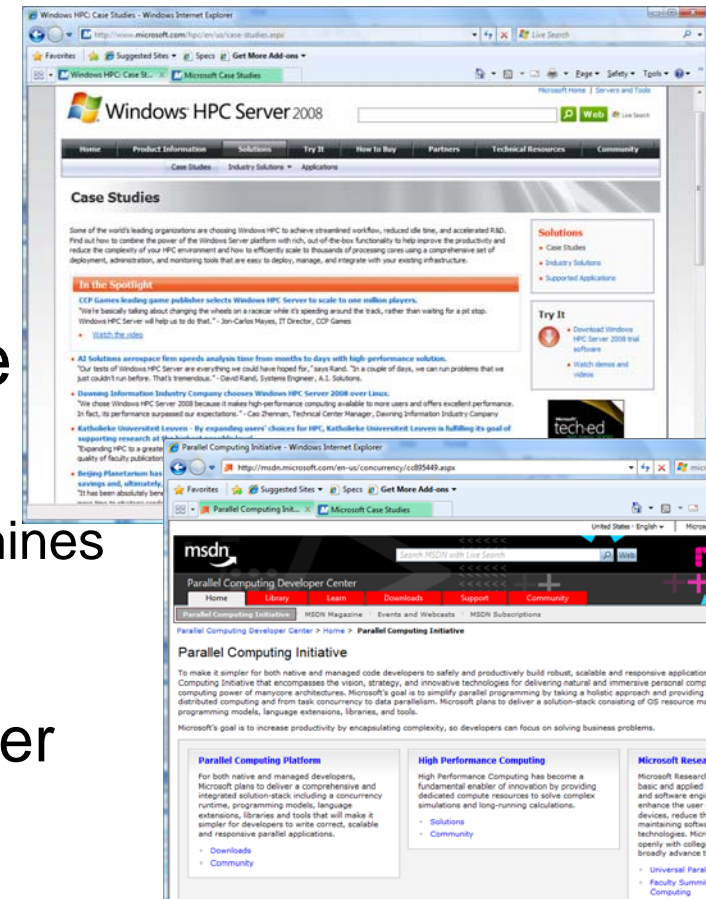
Windows HPC Server 2008

➤ Case Studies

- <http://www.microsoft.com/hpc/en/us/case-studies.aspx>

➤ HPC as Part of a Larger Picture

- Parallel compute initiative
 - Scale on core to many to many machines
 - <http://msdn.microsoft.com/en-us/concurrency/default.aspx>
- Enterprise mgmt via System Center
 - <http://www.microsoft.com/systemcenter>



What's new in the HPC Pack 2008?



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

Systems
Management

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

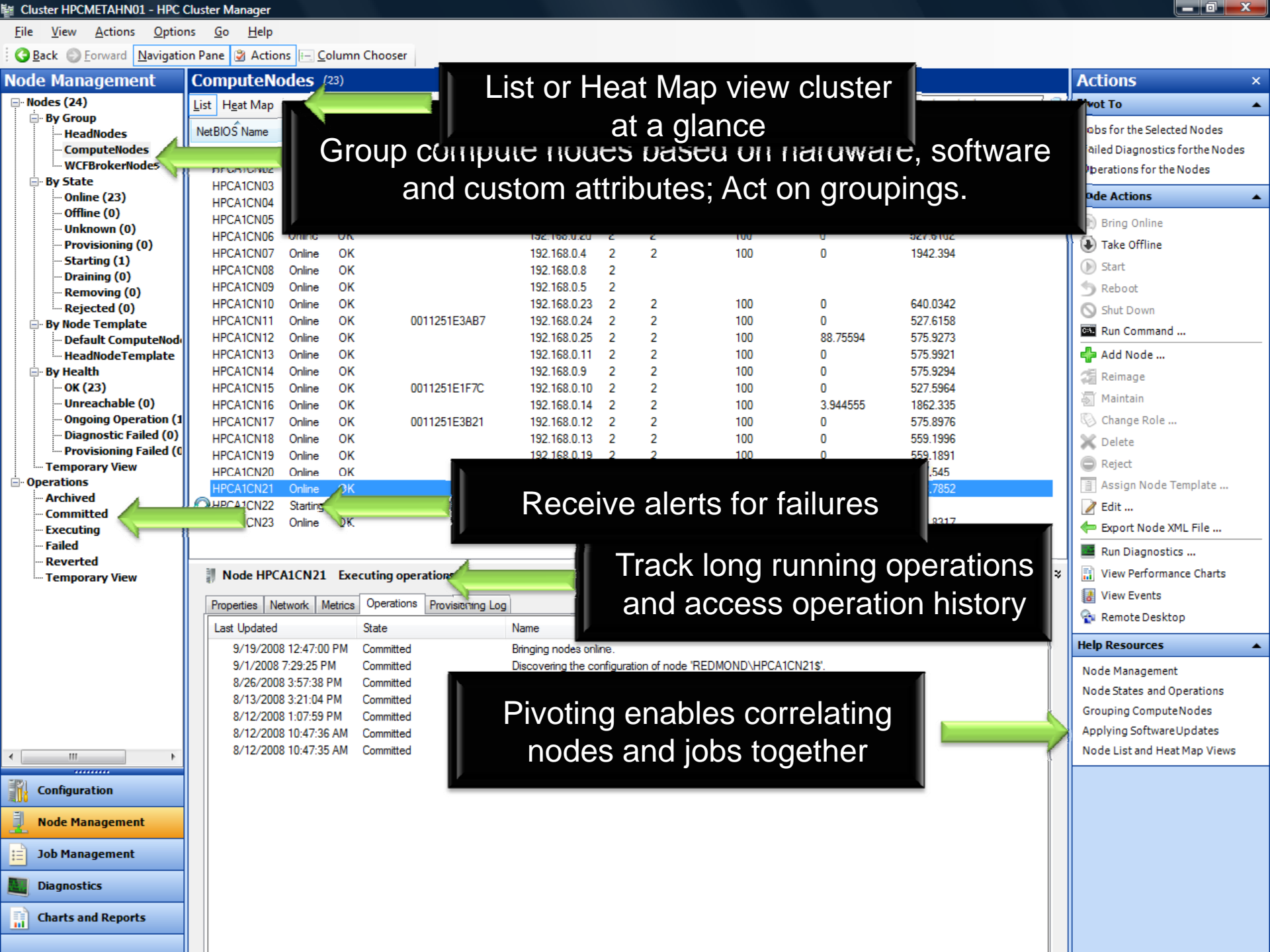
Job
Scheduling

Networking
& MPI

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

Storage

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



List or Heat Map view cluster at a glance

Group compute nodes based on hardware, software and custom attributes; Act on groupings.

Receive alerts for failures

Track long running operations and access operation history

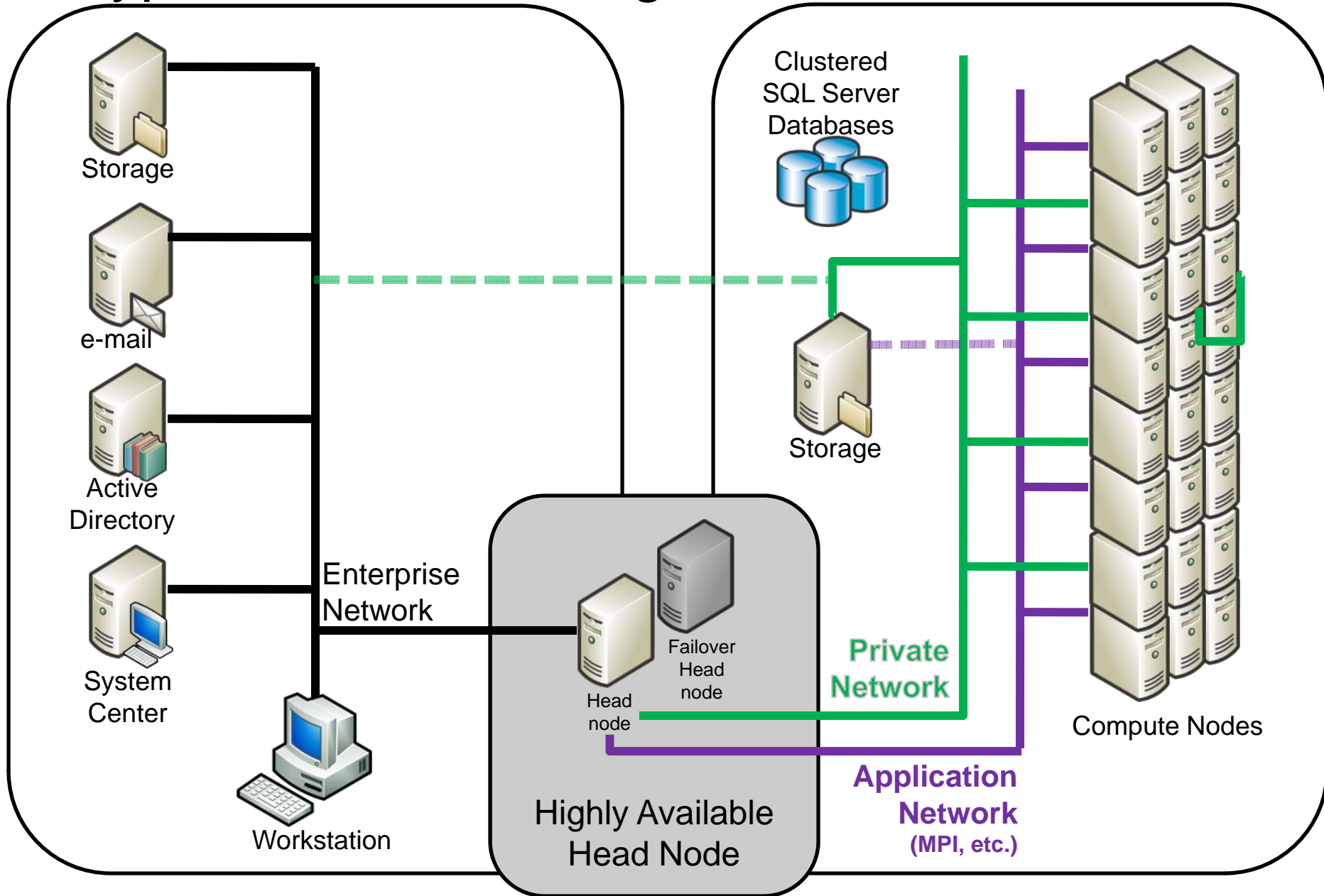
Pivoting enables correlating nodes and jobs together

Node HPCA1CN21 Executing operations

Last Updated	State	Name
9/19/2008 12:47:00 PM	Committed	Bringing nodes online.
9/1/2008 7:29:25 PM	Committed	Discovering the configuration of node 'REDMOND\HPCA1CN21\$'.
8/26/2008 3:57:38 PM	Committed	
8/13/2008 3:21:04 PM	Committed	
8/12/2008 1:07:59 PM	Committed	
8/12/2008 10:47:36 AM	Committed	
8/12/2008 10:47:35 AM	Committed	

- Node Management
- Node States and Operations
- Grouping ComputeNodes
- Applying Software Updates
- Node List and Heat Map Views

Typical Cluster Configuration



WinOF Stack Is Central

- WinOF leverages our dev efforts & focuses our testing
- OEMs demand proof points before committing fully
 - WinOF “concentrates” our experience.
- Breadth – ND, WSD, IPoIB, SRP, uDAPL, Tools
- Simplicity – One stack that works on all IB hardware
??and iWARP too??

MS HPC: Topics of Interest

- Improved OpenSM
- Better diagnostics
 - Closer parity w/ Linux tools
 - Simpler, more integrated results
- Network Boot (& PXE boot)
- Faster IPoIB (connection-based)
- NDIS6 (currently at NDIS5.x)
- Clearer understanding of iWARP/IB delivery
- Windows Logo offered for organizations (OFA)
- Windows is part of OFA InterOp testing

Interpreting vstat

```
hca_idx=0
uplink={BUS=PCI_E, SPEED=2.5 Gbps, WIDTH=x8, CAPS=2.5*x8}
  vendor_id=0x05ad
  vendor_part_id=0x6278
  hw_ver=0xa0
  fw_ver=0x400070258
  PSID=MT_00A0000001
  node_guid=0005:ad00:000b:5e18
  num_phys_ports=2
    port=2
      port_state=PORT_ACTIVE (4)
      Link_speed=2.5 Gbps (1)
      link_width=4x (2)
      rate=10 Gbps
      port_phys_state=LINK_UP (5)
      active_speed=2.5 Gbps (1)
      sm_lid=0x0001
      port_lid=0x0001
```

ConnectX best with and QDR
requires PCIe 2.0
Shows as 5Gbps speed

2.5 = SDR
5.0 = DDR
10.0 = QDR

Less than 4x implies a bad cable

Can detect DDR running as SDR

Diagnostics

Tests

- Scheduler
- Services
- Connectivity
- System Configuration
- SOA
- Performance
- Test Results**
- Running
- Success
- Warning
- Failure
- FailedToRun
- Complete
- Temporary View

Test Results (38)

Filter: Test suite Failed node Last updated

Test Name	Result	Test Suite	Target	Last Updated
✓ MPI Ping-Pong: Lightwe...	Success	Performance	22 nodes	9/19/2008 2:11:40 PM
⚠ MPI Ping-Pong: Quick ...	Warning	Performance	22 nodes	9/19/2008 2:11:11 PM
SOA Model Latency	Success	SOA	23 nodes	9/5/2008 6:49:43 PM
MPI Ping-Pong: Quick ...	Success	Performance	22 nodes	9/5/2008 6:45:39 PM
MPI Ping-Pong: Quick ...	Failure	Performance	22 nodes	9/5/2008 6:18:35 PM
All Services Running	Success	Services	23 nodes	9/5/2008 6:18:27 PM
MPI Ping-Pong: Lightwe...	Failure	Performance	HPCA1CN03	9/5/2008 6:01:23 PM
MPI Ping-Pong: Lightwe...	Failure	Performance	HPCA1CN03	9/5/2008 5:56:10 PM
MPI Ping-Pong: Lightwe...	Failure	Performance	HPCA1CN03	9/5/2008 5:34:52 PM

MPI Ping-Pong: Quick Check

Result

Summary

Warning

A "Warning" assessment indicates that at least one node is performing poorly relative to the other nodes in the cluster. A poorly performing node meets BOTH of the following criteria:

Average latency/throughput over all network links for the node is at least one standard deviation away from the mean value for the cluster AND

Latency is at least 20% higher or throughput is at least 20% lower than the cluster mean. This avoids unwarranted Warnings on highly-uniform cluster networks.

Result Summary

This table summarizes the test results for the nodes.

Results	No. of Nodes
Warning	4
Success	18

Latency Summary

Warning

Average = 87.894 usecs

Std Dev = 44.304 usecs

Best Link = 53.551 usecs (HPCA1CN03 <-> HPCA1CN05)

Worst Link = 258.291 usecs (HPCA1CN06 <-> HPCA1CN21)

Variability = High

Packet size for determining latency: 4 Bytes

Link Latency Histogram

This table shows the distribution of measured ping-pong latencies for all node-to-node communication links in the cluster.

Lower Bound (usecs)	Upper Bound (usecs)	Number of links measured within this interval
04.400	114.073	12

Actions

Pivot To

Failed Nodes of the Test
Progress of the Test

Diagnostics

- Cancel Test
- Clear Alert
- Rerun Test
- Export Results ...

Help Resources

Diagnostics
Understanding Diagnostic Tests
Running Diagnostic Tests
Understanding Test Results
Filtering Test Results

Configuration

Node Management

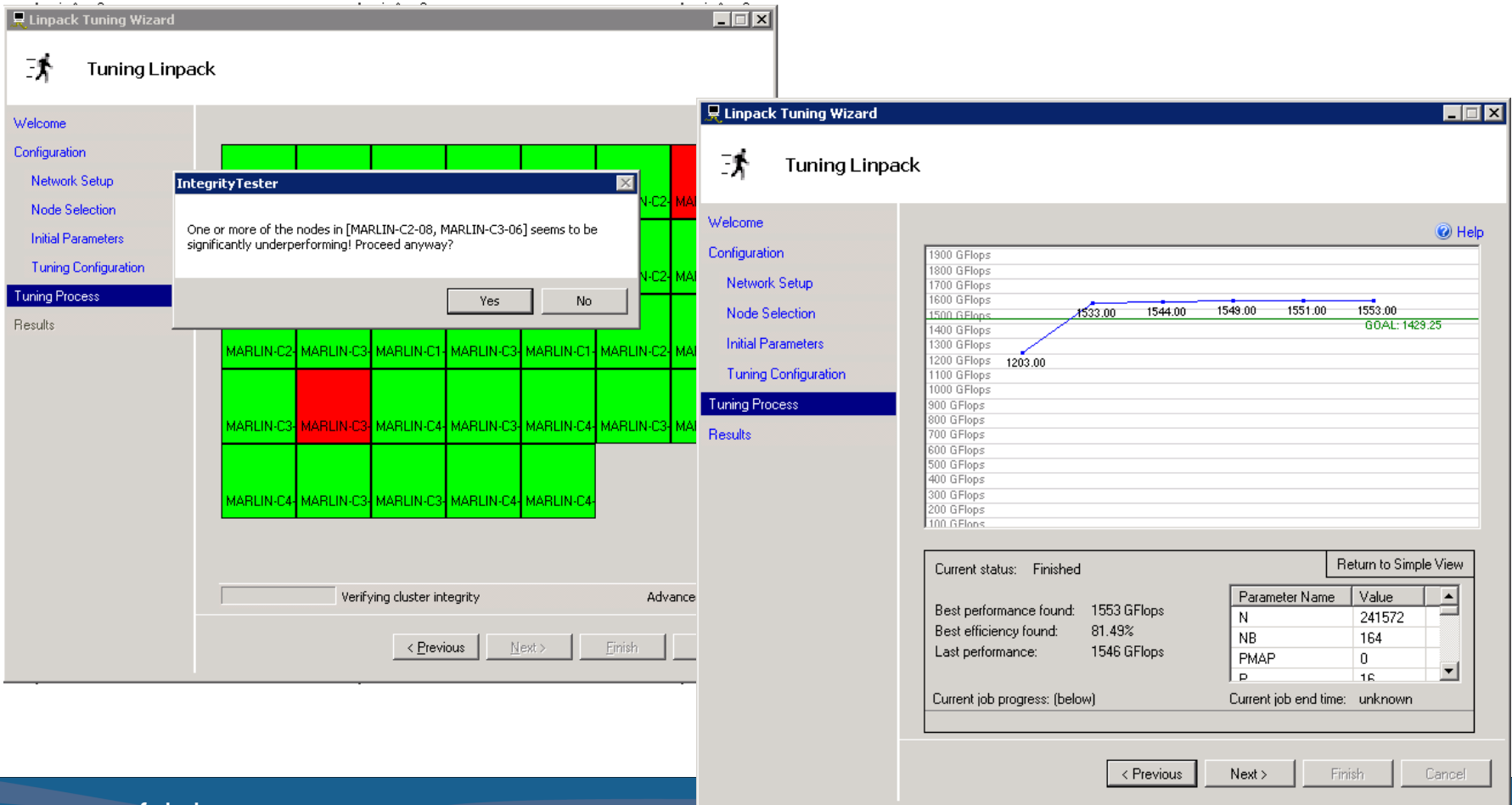
Job Management

Diagnostics

Charts and Reports

Cluster Sanity Testing

➤ Upcoming toolpack tools can help here



The image shows two overlapping screenshots of the Linpack Tuning Wizard interface. The left screenshot shows the 'Tuning Linpack' window with a grid of nodes. A dialog box titled 'IntegrityTester' is open, displaying the message: 'One or more of the nodes in [MARLIN-C2-08, MARLIN-C3-06] seems to be significantly underperforming! Proceed anyway?' with 'Yes' and 'No' buttons. The right screenshot shows the 'Tuning Process' step, featuring a line graph of performance in GFlops over time. The graph shows a peak of 1553.00 GFlops and a goal of 1429.25 GFlops. Below the graph, the current status is 'Finished' and a table lists the parameters used for the test.

Parameter Name	Value
N	241572
NB	164
PMAP	0
P	16

Node Management

Nodes (24)

- By Group
 - HeadNodes
 - ComputeNodes
 - WCFBrokerNodes
- By State
 - Online (23)
 - Offline (0)
 - Unknown (0)
 - Provisioning (0)
 - Starting (1)
 - Draining (0)
 - Removing (0)
 - Rejected (0)
- By Node Template
 - Default ComputeNode
 - HeadNodeTemplate
- By Health
 - OK (23)
 - Unreachable (0)
 - Ongoing Operation (1)
 - Diagnostic Failed (0)
 - Provisioning Failed (0)
- Temporary View
 - Archived
 - Committed
 - Executing
 - Failed
 - Reverted
 - Temporary View

ComputeNodes (23)

List Heat Map Search nodes by name

Metric: Context switches / second [Add to heat map](#) [Customize metric display](#) Zoom

- Context switches / second
- Cores in use
- Disk Queue Length
- Disk Throughput (Bytes/second)
- Free Disk Space (%)
- Memory Paging (Hard Faults/second)
- Running Jobs
- Running Tasks
- System calls / second
- WCF Broker Calls / second
- WCF Failed Broker Calls / second

CPU Usage (%) 0 100

Available Physical Memory (MBytes) 0 8000

Network Usage (Bytes/second) 0 1000000

100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
1508.00	1554.00	1543.00	1545.00	1577.00	1571.00	1545.00	1558.00	1556.00	
527.59	527.65	527.61	527.64	527.63	527.62	527.64	420.01	457.27	
HPCA1CN01	HPCA1CN02	HPCA1CN03	HPCA1CN04	HPCA1CN05	HPCA1CN06	HPCA1CN07	HPCA1CN08	HPCA1CN09	
100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
1546.00	1552.00	1543.00	1543.00	1568.00	1564.00	1381.00	1559.00	1556.00	
560.16	527.64	527.64	527.57	527.60	527.64	527.65	570.02	971.51	
HPCA1CN10	HPCA1CN11	HPCA1CN12	HPCA1CN13	HPCA1CN14	HPCA1CN15	HPCA1CN16	HPCA1CN17	HPCA1CN18	
100.00	100.00	100.00	0.00	100.00					
1560.00	1575.00	1343.00	0.00	1356.00					
527.62	570.05	439.69	0.00	527.62					
HPCA1CN19	HPCA1CN20	HPCA1CN21	HPCA1CN22	HPCA1CN23					

- Configuration
- Node Management**
- Job Management
- Diagnostics
- Charts and Reports

Charts and Reports

- Monitoring Charts
- Reports
 - Node Availability
 - Job Resource Usage
 - Job Throughput
 - Job Turnaround

- Configuration
- Node Management
- Job Management
- Diagnostics
- Charts and Reports

Monitoring Charts

[Help](#)

Summary

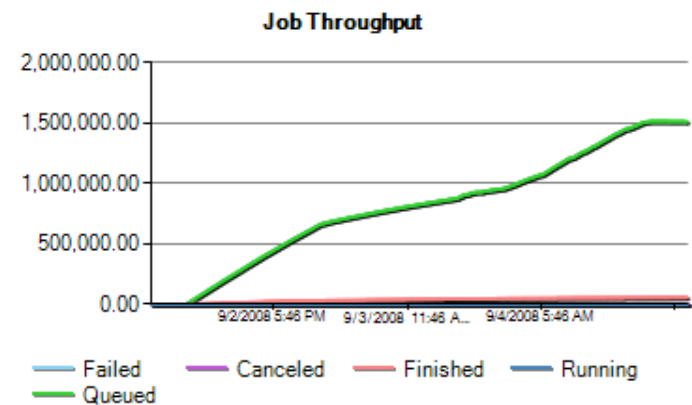
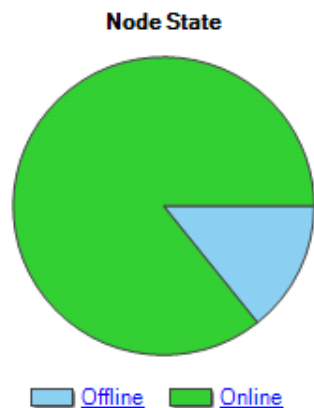
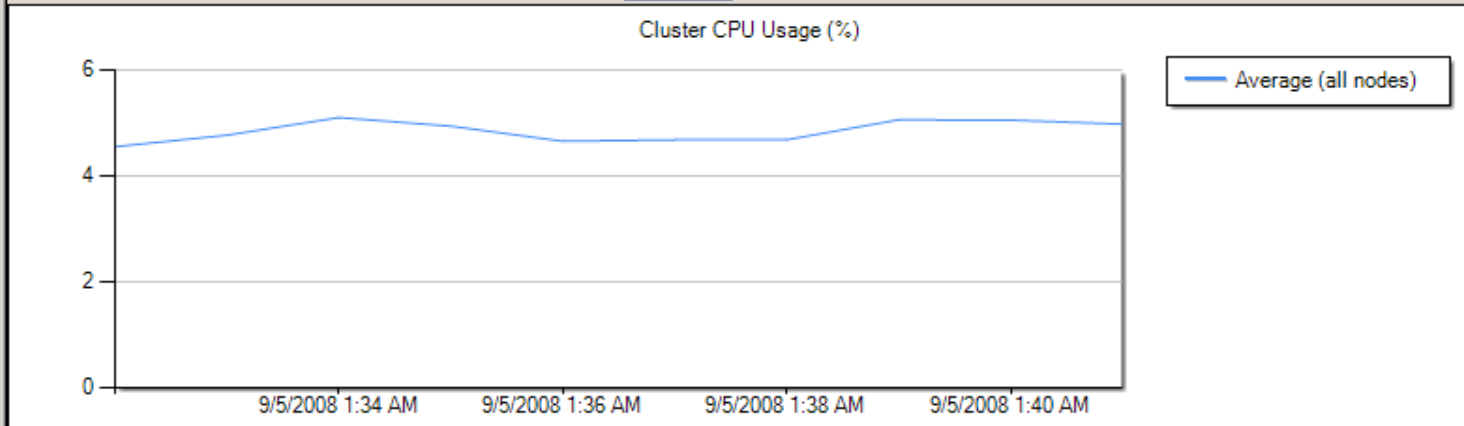


Chart details

Metric: Scheduler Jobs Add

X axis (minutes): 10 | Y axis: 0 - 0 | Autoscale Refresh



Charts and Reports

- Monitoring Charts
- Reports
 - Node Availability
 - Job Resource Usage
 - Job Throughput**
 - Job Turnaround

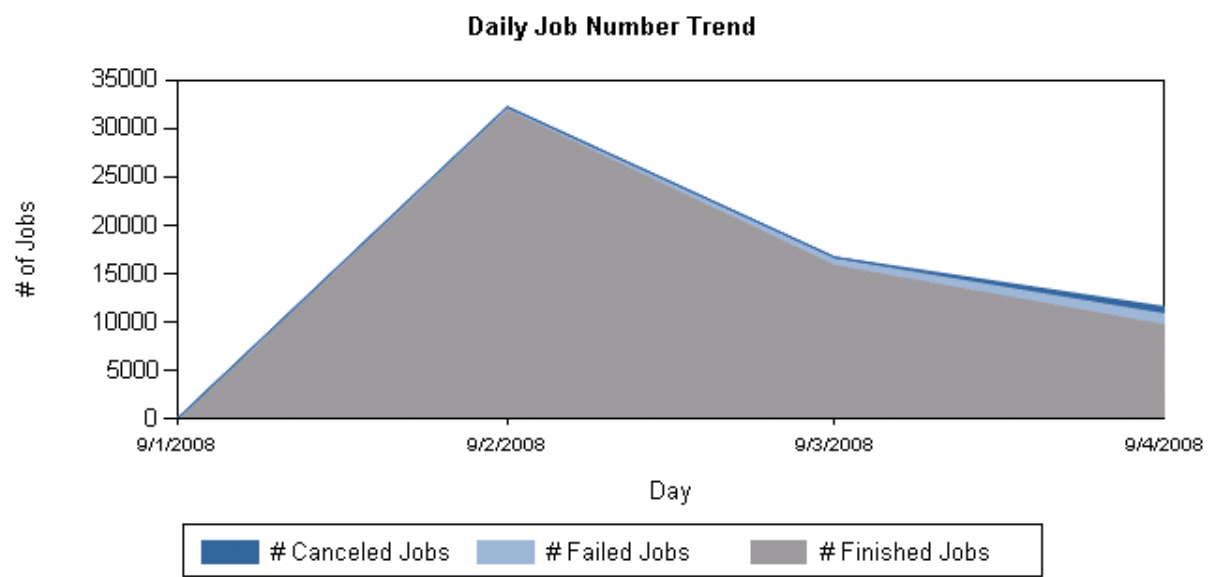
- Configuration
- Node Management
- Job Management
- Diagnostics
- Charts and Reports

Job Throughput

Date Range: Custom From: 9/ 1/2008 To: 9/ 4/2008 Group By: User Filter: All View Report Help

3 of 4 Page Width Find Next

Trend - Daily Trend within the time period for User/All



Expand to see data

Time Period	# of Finished Jobs	% of Finished Jobs	# of Failed Jobs	% of Failed Jobs	# of Canceled Jobs	% of Canceled Jobs	Total # of Jobs
9/1/2008	6	18.75 %	2	6.25 %	24	75.00 %	32
9/2/2008	32,010	99.57 %	137	0.43 %	1	0.00 %	32,148
9/3/2008	15,941	95.87 %	676	4.07 %	11	0.07 %	16,628
9/4/2008	9,840	85.44 %	1,101	9.56 %	576	5.00 %	11,517

Charts and Reports

- Monitoring Charts
- Reports
 - Node Availability
 - Job Resource Usage
 - Job Throughput
 - Job Turnaround**

- Configuration
- Node Management
- Job Management
- Diagnostics
- Charts and Reports

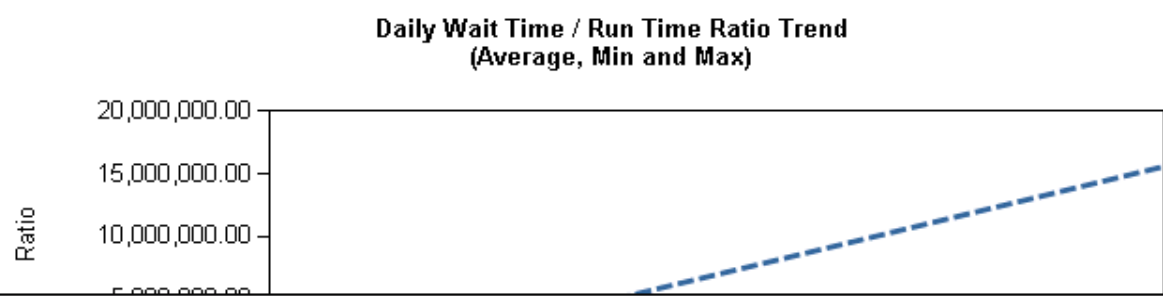
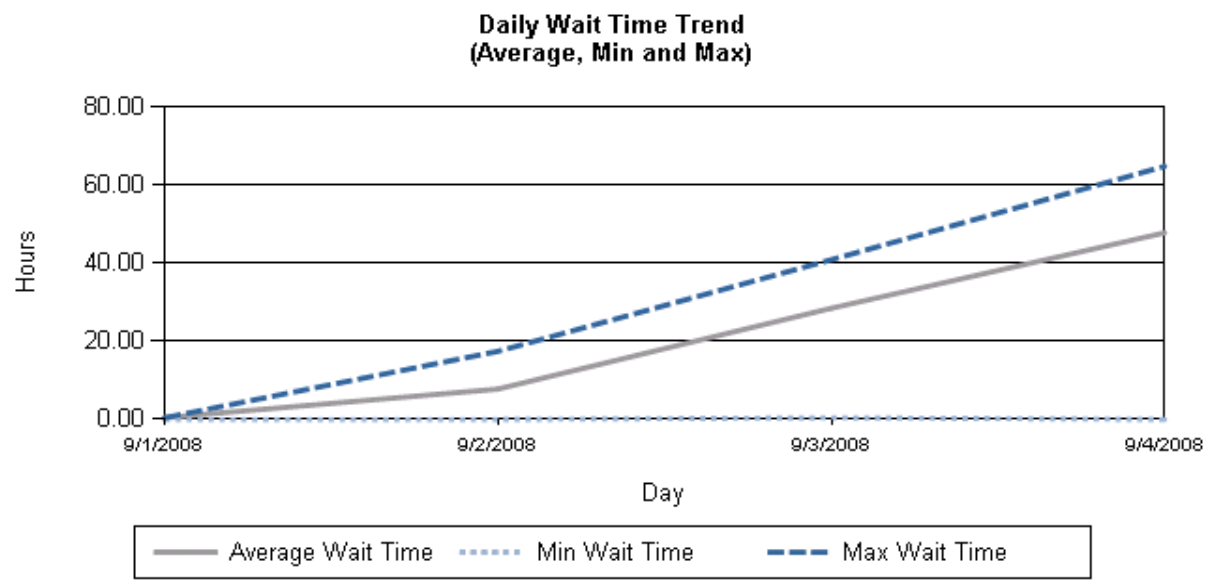
Job Turnaround

Date Range: Custom From: 9/ 1/2008 Group By: User View Report Help

To: 9/ 4/2008 Filter: All

3 of 4 Page Width Find Next

Trend - Daily Trend within the time period for User/All



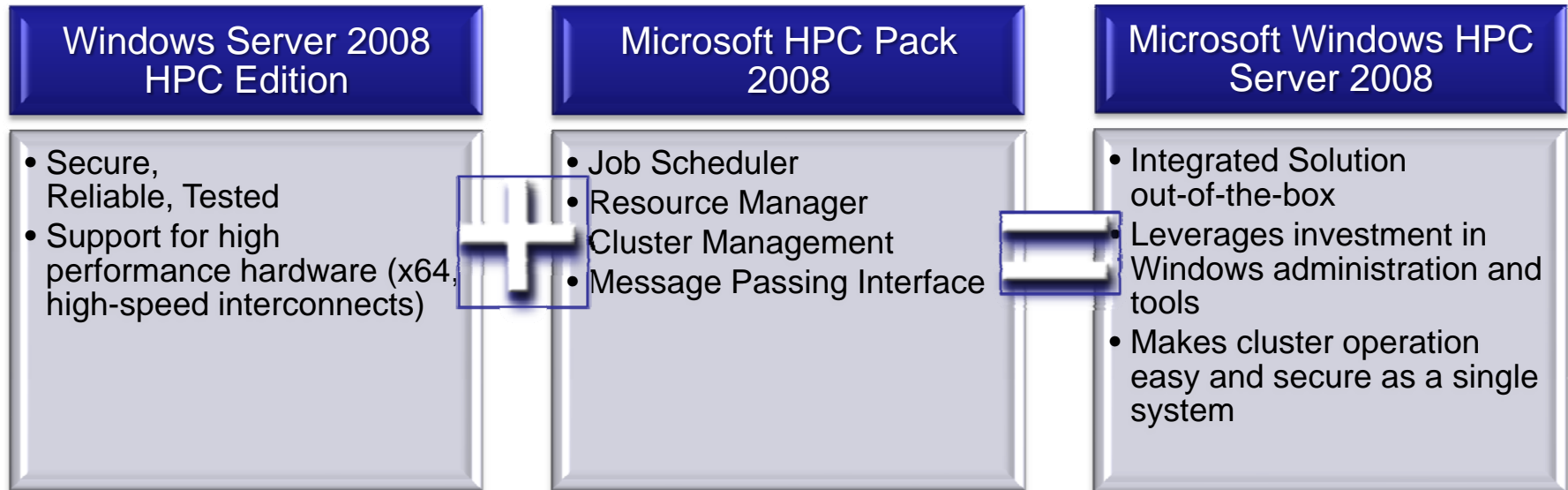
MS HPC: Topics of Interest

- Improved OpenSM
- Better diagnostics
 - Closer parity w/ Linux tools
 - Simpler, more integrated results
- Network Boot (& PXE boot)
- Faster IPoIB (connection-based)
- NDIS6 (currently at NDIS5.x)
- Clearer understanding of iWARP/IB delivery
- Windows Logo offered for organizations (OFA)
- Windows is part of OFA InterOp testing

OTHER INTERESTING BITS

Windows HPC Server 2008

- Complete, integrated platform for computational clustering
- Built on top the proven Windows Server 2008 platform
- Integrated development environment



Evaluation available from <http://www.microsoft.com/hpc>

NetworkDirect- 3 Points to Remember



➤ **NetworkDirect is fast- really fast**

HPC Server 2008 stack produced world-class cluster efficiencies in June 2008 Top500 runs at: NCSA (#23), UMEA (#39), Aachen (#100) and Nov2008 Top500: Shanghai Supercomputing Center (#10)

➤ **And Stable**

The MS HPC team have significant mileage on ND-enabled clusters

- 2,000 cores routinely
- max. tested to date: 30,000 cores

for hours/days without fail (MPI failures are easy to spot! ;)

➤ **And Logo Tested**

MS HPC, Core Networking, and Windows Logo teams have created a logo program for NetworkDirect drivers. The first submissions are coming in now for:

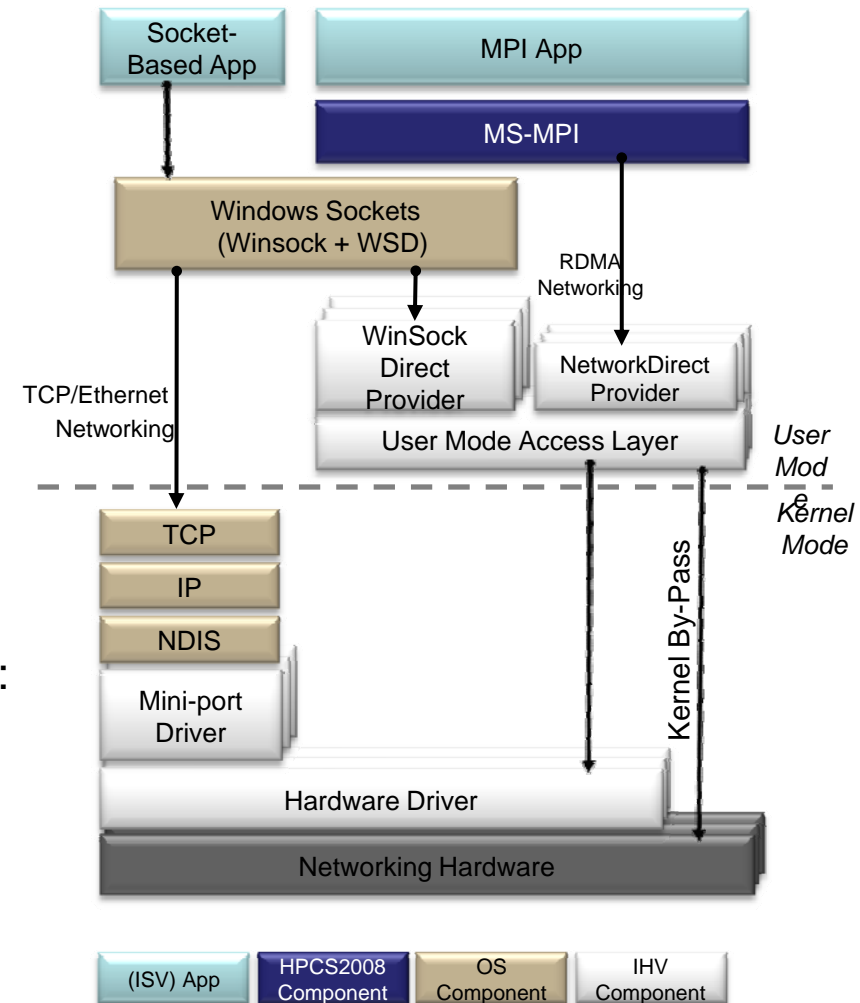
- Infiniband vendors (3)
- 10GigE vendors (2)

NetworkDirect

A new RDMA networking interface built for speed and stability



- Verbs-based design for close fit with native, high-perf networking interfaces
- Equal to Hardware-Optimized stacks for MPI micro-benchmarks
- NetworkDirect drivers for key high-performance fabrics:
 - Infiniband [available now!]
 - 10 Gigabit Ethernet (iWARP-enabled) [available now!]
 - Myrinet [available soon]
- 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
 - New NetworkDirect interface



Version Comparison



Feature	Windows Compute Cluster Server 2003	Windows HPC Server 2008
Operating system	Windows Server 2003 SP1	Windows Server 2008 HPC Edition, Standard, Enterprise, Datacenter
Processor Type	X64 (AMD64 or Intel EM64T)	X64 (AMD64 or Intel EM64T)
Memory	32 GB (Compute Cluster Edition)	128 GB (HPC Edition)
Node Deployment	Remote Installation Services(RIS)	Windows Deployment Services
Head Node Availability	N/A	Windows Failover Clustering and SQL Server Failover Clustering
Management	Basic node and job management	Integrated node and job management, grouping, monitoring at-a-glance, diagnostics
Network Topology	Network Configuration Wizard	Improved Network Configuration Wizard
MS-MPI	Winsock Direct-based	Network Direct-based. New shared memory implementation for multicore processors
Scheduler	Command line or GUI	Integrated in management console, with full support for Windows PowerShell scripting and legacy command-line UI scripts from v1. Greatly improved speed and scalability
Programmability	Support for Batch or MPI based jobs	Added support for interactive Service Oriented Applications (SOA) using the Windows Communication Foundation (WCF)
Reporting	N/A	Integrated into Management console
Monitoring	Rely on Windows. No cluster specific support.	Heat map on cluster or node group. Per node charts. Cluster-wide performance overview
Diagnostics	N/A	In the box verification tests and performance tests. Store, filter, and view test results and history

Resources

- Microsoft HPC Web site – Evaluate Today!
 - <http://www.microsoft.com/hpc>
- Windows HPC Community site
 - <http://www.windowshpc.net>
- Windows HPC Techcenter
 - <http://technet.microsoft.com/en-us/hpc/default.aspx>
- HPC on MSDN
 - <http://code.msdn.microsoft.com/hpc>
- Windows Server Compare website
 - <http://www.microsoft.com/windowsserver/compare/default.mspix>