HIGH-PERFORMANCE NETWORKING IN WINDOWS COMPUTE CLUSTERS

Eric Lantz (elantz)
Lead Program Manager, Windows HPC Team
Microsoft Corp.

AGENDA

- Microsoft Compute Cluster Server (CCS)
  - How Does Microsoft Describe the HPC Market?
  - What is CCS?
- Networking Options for CCS
  - Optimizing for Performance
- MS Investing in Infiniband
  - HPC Hosted Clusters
  - HPC Team clusters
- A Word about CCSv2

BUSINESS MOTIVATIONS "HIGH PRODUCTIVITY COMPUTING“

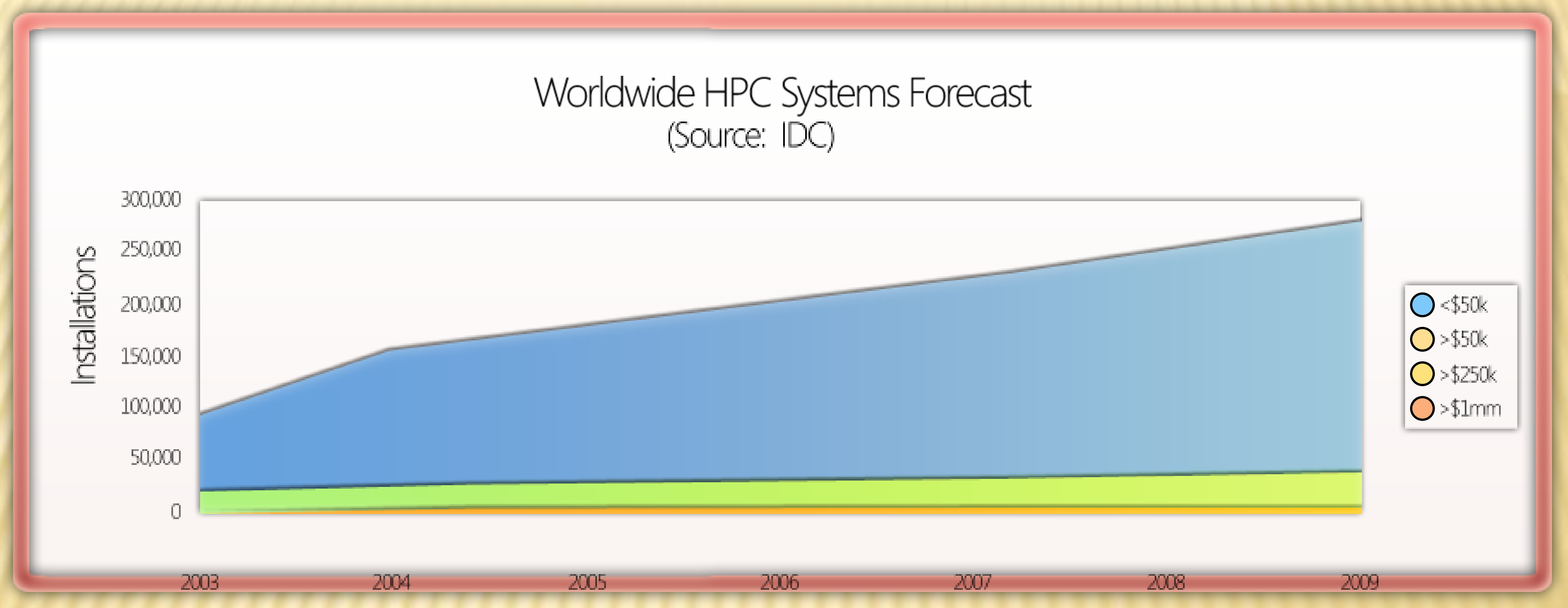
- Application complexity increases faster than clock speed so need for parallelization
- Windows applications users need cluster-class computing
- Make compute cluster ubiquitous and simple starting at the departmental level
- Remove customer pain points for
  - Implementing, managing and updating clusters
  - Compatibility and integration with existing infrastructure
  - Testing, troubleshooting and diagnostics

HPC market is growing- 50% for cluster servers (source IDC 2006). Need for resources such as development tools, storage, interconnects and graphics

MARKET PERSPECTIVE

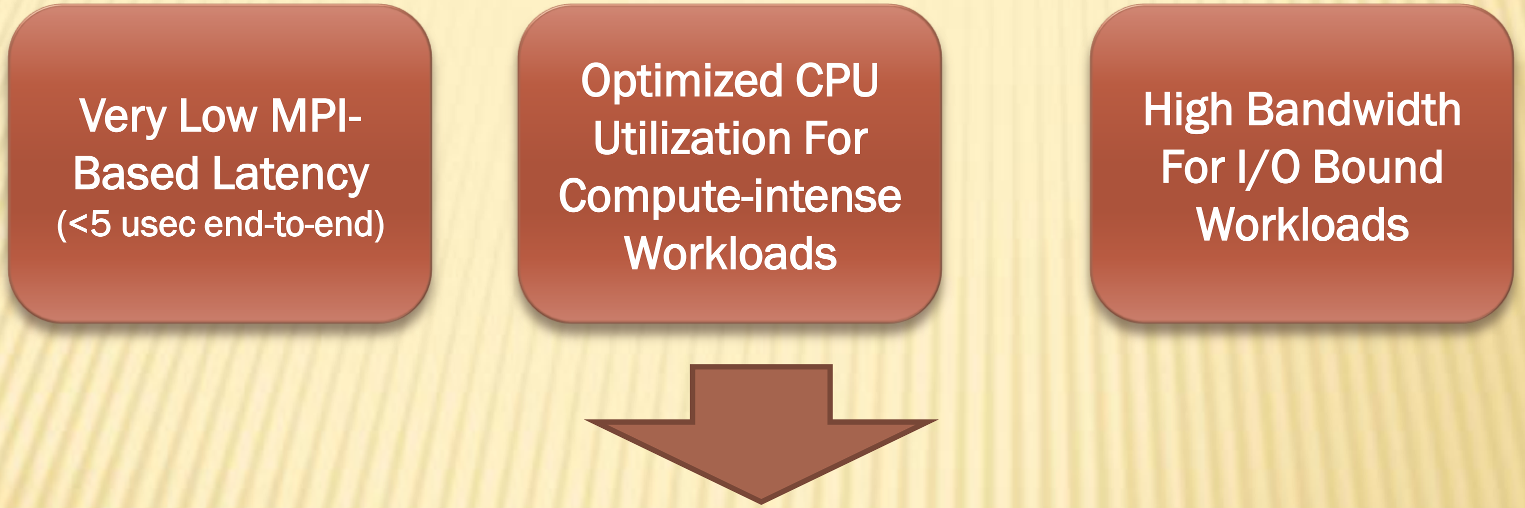
	1991	1998	2005
System	Cray Y-MP C916 	Sun HPC10000 	Small Form Factor PCs
Architecture	16 x Vector 4GB, Bus	24 x 333MHz Ultra-SPARCII, 24GB, SBus	4 x 2.2GHz Athlon64 4GB, GigE
OS	UNICOS	Solaris 2.5.1	Windows Server 2003 SP1
GFlops	**~10**	**~10**	**~10**
Top500 #	1	500	N/A
Price	**$40,000,000**	**$1,000,000 (40x drop)**	**< $4,000 (250x drop)**
Customers	Government Labs	Large Enterprises	Every Engineer & Scientist
Applications	Classified, Climate, Physics Research	Manufacturing, Energy, Finance, Telecom	Bioinformatics, Materials Sciences, Digital Media

HPC GROWTH



- x86 server clusters growing faster than market (15%-20% for HPC, 11.4% for x86 overall). Projected at 850,000 units in 2007.
- Windows CCS is strategic investment, focused on driving volume market for computationally intense applications.

HPC NETWORKING REQUIREMENTS



Very Low MPI-Based Latency
(<5 usec end-to-end)

Optimized CPU Utilization For Compute-intense Workloads

High Bandwidth For I/O Bound Workloads

- Windows Compute Cluster Server requires RDMA as core networking technology.
- Tier 1 OEMs estimate RDMA-enabled fabrics included in
  - ~20% of units in 2007
  - ~40% of units in 2008-2009

WINDOWS COMPUTE CLUSTER SERVER 2003

Mission: **Deliver the easiest to deploy and most cost effective solution for solving scaled-out business, engineering and scientific computational problems.**

Windows Server 2003, Compute Cluster Edition + Compute Cluster Pack = Microsoft Windows Compute Cluster Server 2003

Windows Server 2003, Compute Cluster Edition

- Support for high performance hardware (x64bit architecture)
- RDMA support for high performance interconnects (Infiniband, Myrinet, and others)

Compute Cluster Pack

- Support for Industry Standards MPI2
- Integrated Job Scheduler
- Cluster Resource Management Tools
- CCS SDK
    - Scheduler
    - Parallel Programming

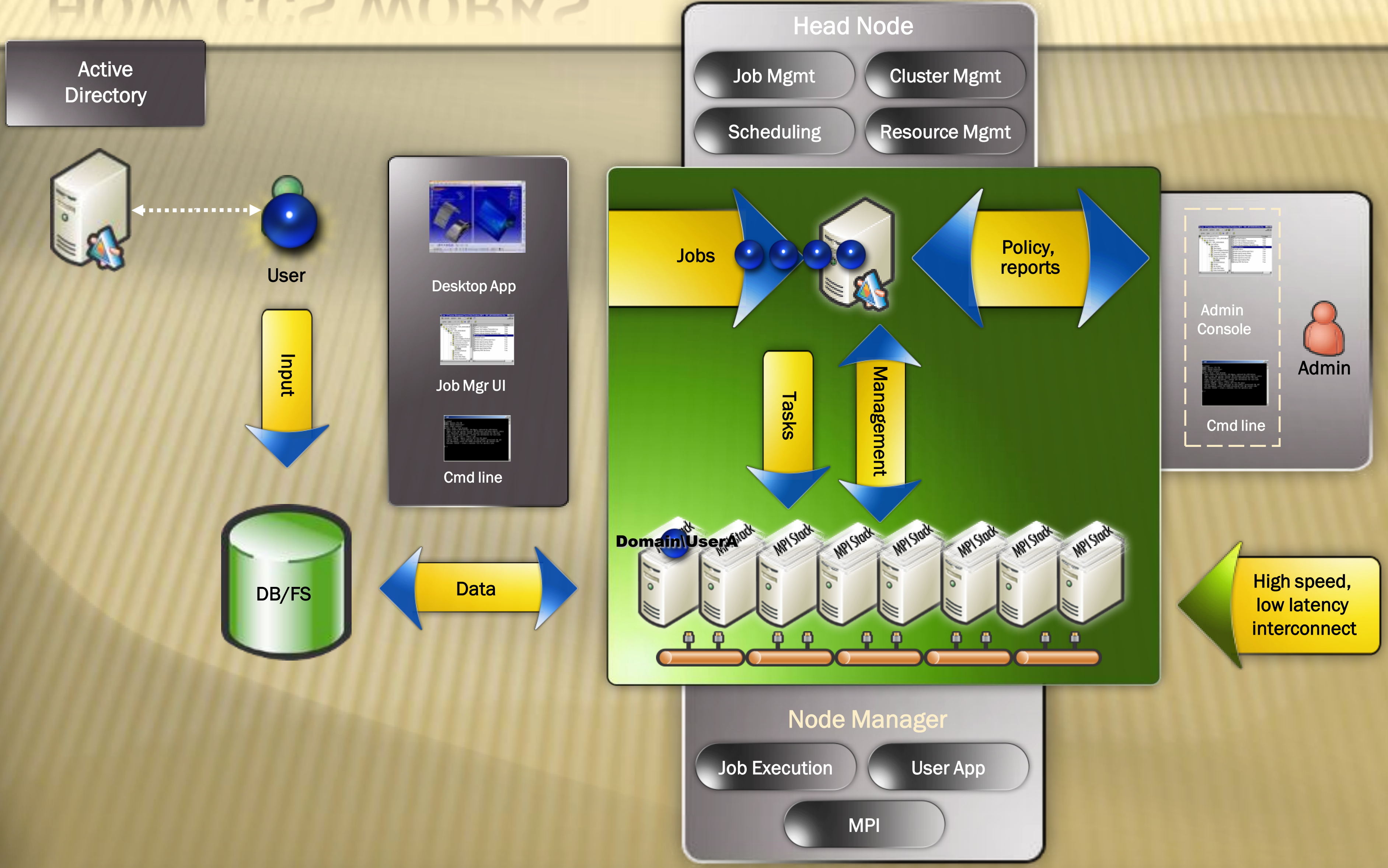
Microsoft Windows Compute Cluster Server 2003

- Integrated Solution out-of-the-box
- Leverages investment in Windows administration and tools
- Makes cluster operation easy and secure as a single system

CCS KEY FEATURES

- Easier node deployment and administration
  - Task-based configuration for head and compute nodes
  - UI and command line-based node management
  - Monitoring with Performance Monitor (Perfmon), Microsoft Operations Manager (MOM), Server Performance Advisor (SPA), and 3rd-party tools
- Extensible job scheduler
  - Simple job management, similar to print queue management
  - 3rd-party extensibility at job submission and/or job assignment
  - Submit jobs from command line, UI, or directly from applications
- Integrated Development Environment
  - OpenMP Support in Visual Studio, Standard Edition
  - Parallel Debugger in Visual Studio, Professional Edition
  - MPI Profiling tool

HOW CCS WORKS



Active Directory

User

Input

DB/FS

Data

Desktop App

Job Mgr UI

Cmd line

Head Node

Job Mgmt

Cluster Mgmt

Scheduling

Resource Mgmt

Jobs

Policy, reports

Tasks

Management

Domain\UserA

MPI Stack

Admin Console

Cmd line

Admin

High speed, low latency interconnect

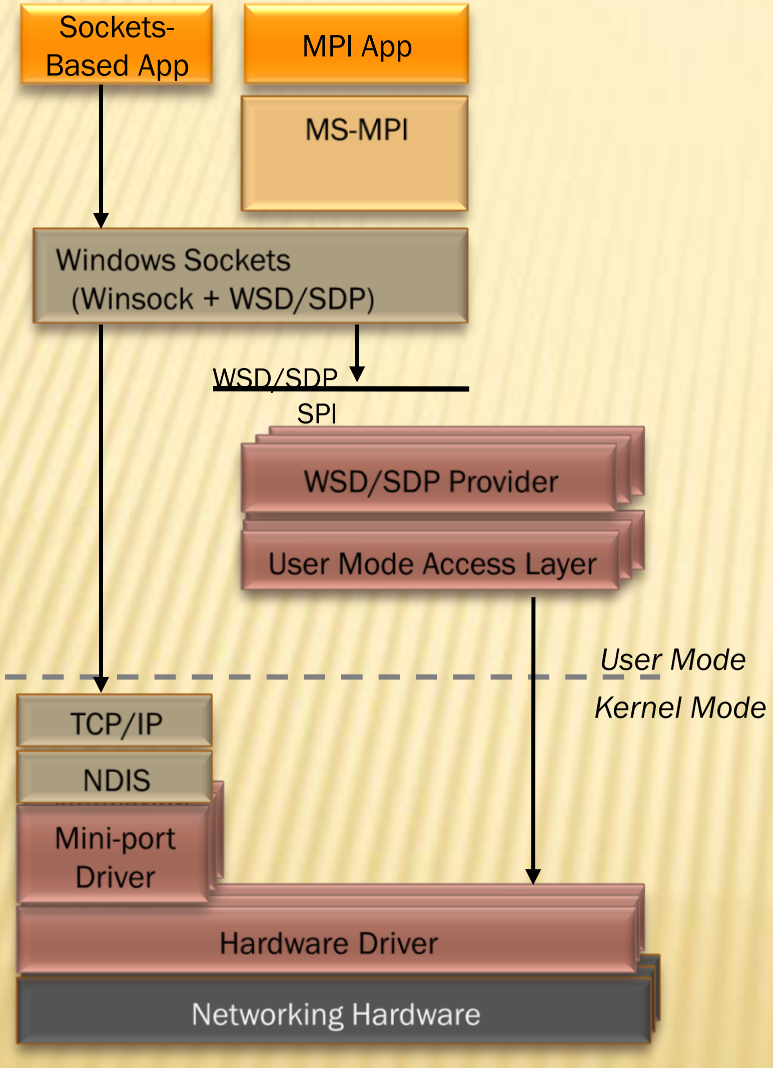
Node Manager

Job Execution

User App

MPI

MS-MPI BUILT ON WINSOCK DIRECT



- MS-MPI Uses Winsock Direct
  - Lower Latency than NDIS path
  - Increased flexibility for users to upgrade their network gear *without* rebuilding their application

- Microsoft Compute Cluster Server (CCS)
  - How Does Microsoft Describe the HPC Market?
  - What is CCS?
- Networking Options for CCS
  - Optimizing for Performance
- MS Investing in Infiniband
  - HPC Hosted Clusters
  - HPC Team clusters
- A Word about CCSv2

NETWORKING OPTIONS FOR CCS

OPTIMIZING PERFORMANCE ON WINDOWS

- Network Congested?
  - Depending on switching, All-To-All & similar operations can drop connections via timeout
    - HKLM\SYSTEM\CurrentControlSet\Services\Tcpip\Parameters\TcpMaxDataRetransmissions = 20 (default=5)
- Hammering One of Your Nodes?
  - All-To-One & similar operations can trigger Syn Attack Protection
  - Shut off Syn Attack monitoring on compute nodes (but NOT head node)
    - HKLM\SYSTEM\CurrentControlSet\Services\Tcpip\Parameters\SynAttackProtect = 0 (default = 1)
- A Couple Patches You Should Apply
  - >4 processors? Then you NEED this one, but all should apply
    - KB914784: Update for Kernel patch protection
  - Using Winsock Direct?
    - KB927620: Resolve performance issues experienced when using Winsock Direct (WSD)
  - Both patches are included in Windows Server 2003 SP2
- Can set processor affinity via either of 2 methods
  - Tag affinity onto an executable's PE area with IMAGECFG.EXE tool
  - At the command line with start's /affinity argument
- Whitepaper has detailed IB info including use of OpenFabric driver/tools.
  - And detailed perf measurement procedure for Windows clusters
  - And use of Windows Perfmon with recommended counters for HPC use
    - http://www.microsoft.com/downloads/details.aspx?FamilyID=40cd8152-f89d-4abf-ab1c-a467e180cce4&DisplayLang=en

IB TRICKS ON A CCS CLUSTER

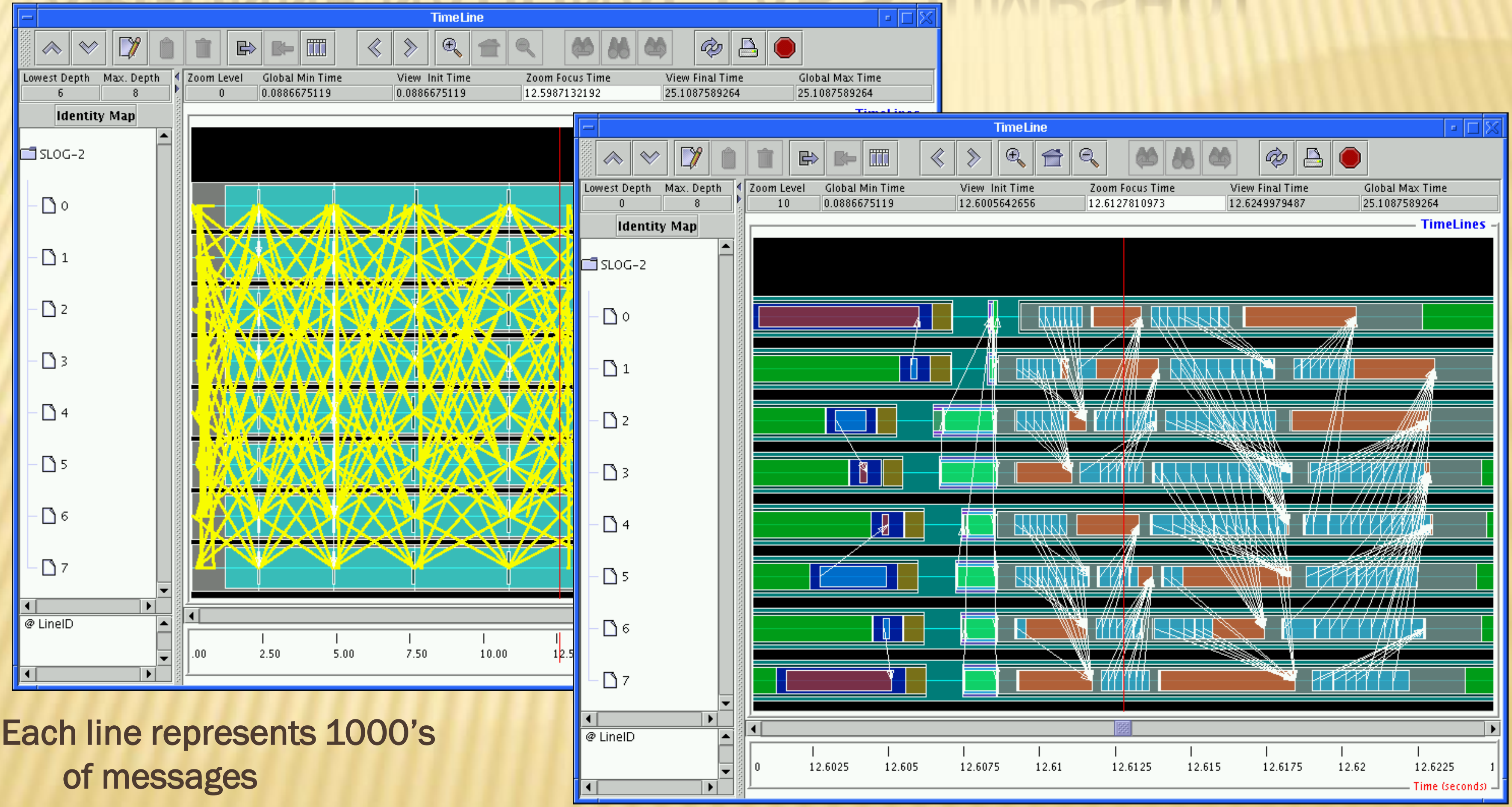
- Determine your brand of IB network adapter (HCA) without opening the computer.
  - Use the vstat utility and check the first 3 bytes of the node_guid against the list of vendor organizationally unique identifiers (OUIs) at http://standards.ieee.org/regauth/oui/index.shtml.
- Ensure the IB adapter is enabled on all nodes.
  - You should get the same count as you have nodes when you run the following command.
  - clusrun /all c:\drivers\openib\mft\mst status | find /c “mt25208_pciconf0”
- Remotely update the IB HCA’s firmware on all nodes
  - clusrun /all c:\drivers\openib\mft\flint.exe –y nofs –d mt25208_pciconf0 –i <the right firware.bin> burn
- Determine the number of PCI devices found on each ready node
  - clusrun /readynodes \\headnode\share\devcon findall pci* | find "matching“
- Determine the number of Mellanox cards found across all nodes
  - clusrun /all \\headnode\share\devcon findall pci* | find /c "Mellanox"

MS-MPI TRICKS FOR INFINIBAND

- Environment Variables To Configure MS-MPI For Use With WSD-Enabled Infiniband

Variable	**Setting**
MPICH_SOCKET_SBUFFER_SIZE	0 (no copy on send) Significantly greater bandwidth at the expense of higher CPU utilization. **NOTE:** Use *only* when compute nodes are fitted with a WSD-enabled driver.
MPICH_DISABLE_SHM	1 (do not use shared memory within a local computer) Disable shared memory when aggressively polling with a WSD provider (for example, using InfiniBand's *IBWSD_POLL* environment variable); otherwise, two threads simultaneously poll for network completions, which significantly slows your application on a multiprocessor compute node.

PARALLEL EXECUTION VISUALIZATION WITH ARGONNE NATIONAL LAB’S JUMPSHOT



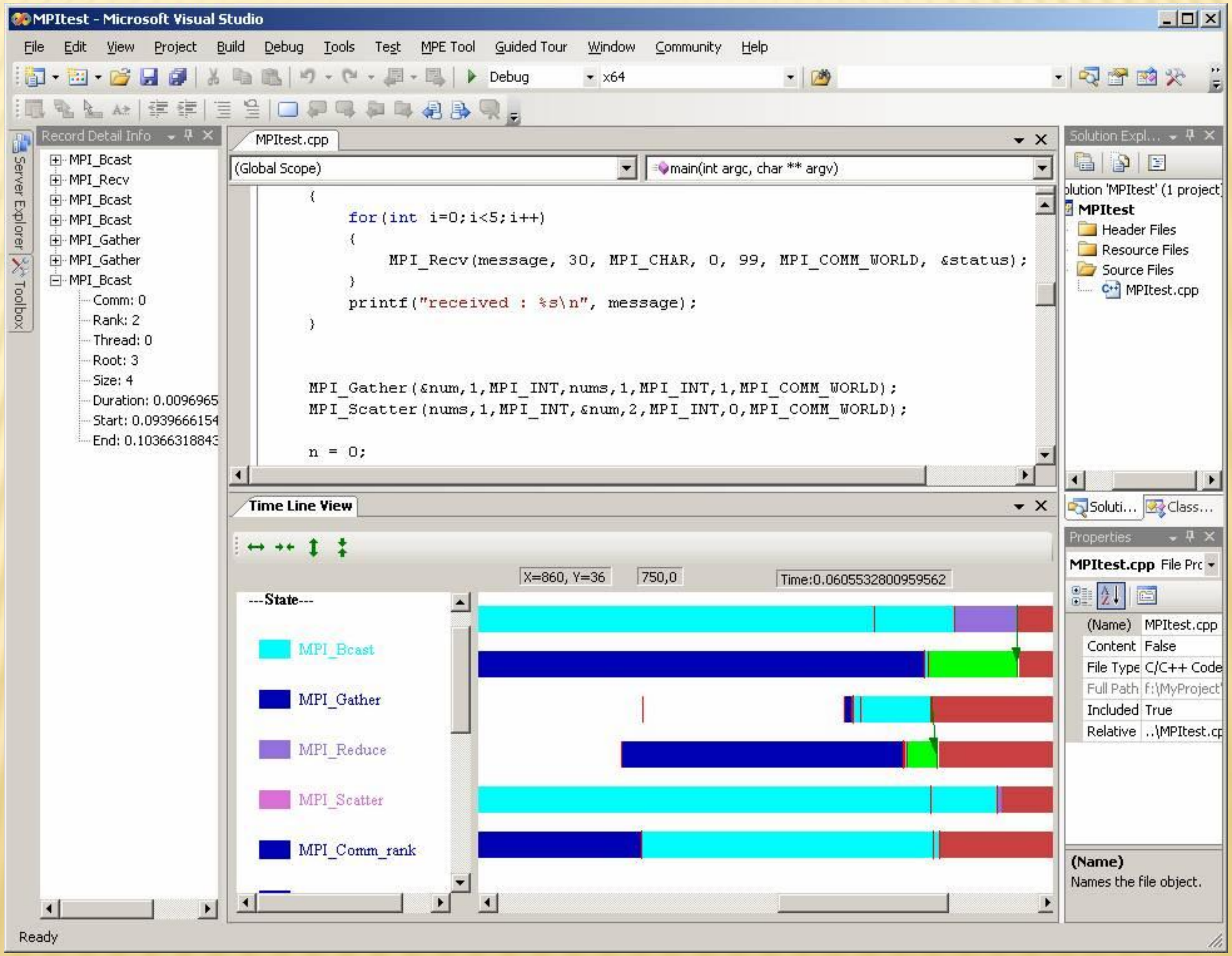
Lowest Depth	Max. Depth	Zoom Level	Global Min Time	View Init Time	Zoom Focus Time	View Final Time	Global Max Time
6	8	0	0.0886675119	0.0886675119	12.5987132192	25.1087589264	25.1087589264

Lowest Depth	Max. Depth	Zoom Level	Global Min Time	View Init Time	Zoom Focus Time	View Final Time	Global Max Time
0	8	10	0.0886675119	12.6005642656	12.6127810973	12.6249979487	25.1087589264

Each line represents 1000’s of messages

Detailed view shows opportunities for optimization

OR USE A VISUAL STUDIO INTEGRATED TOOL FROM THE CCP TOOLPACK



- MPI_Bcast
- MPI_Recv
- MPI_Bcast
- MPI_Bcast
- MPI_Gather
- MPI_Gather
- MPI_Bcast
  - Comm: 0
  - Rank: 2
  - Thread: 0
  - Root: 3
  - Size: 4
  - Duration: 0.0096965
  - Start: 0.0939666154
  - End: 0.10366318843

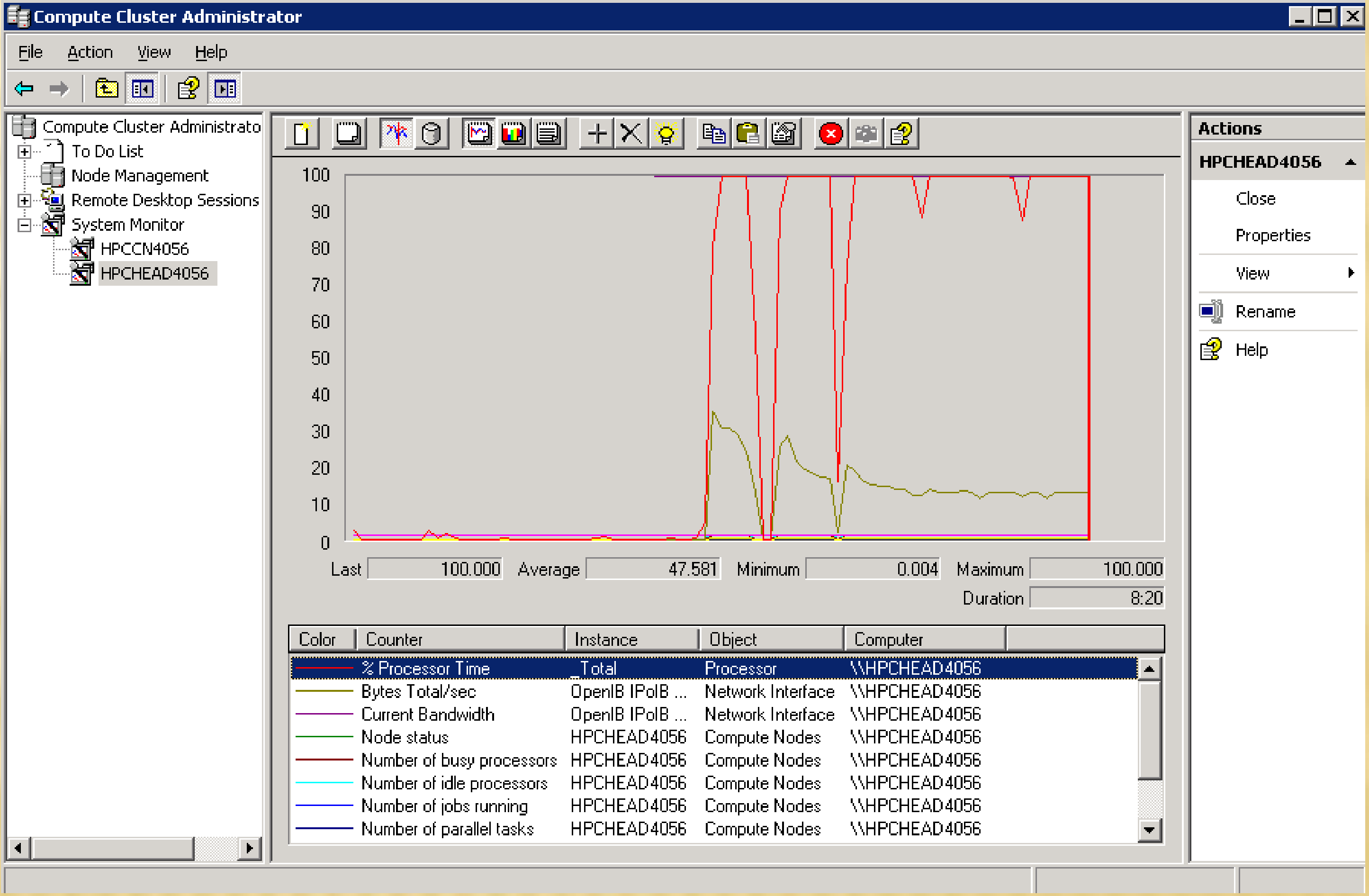
```
{
    for(int i=0;i<5;i++)
    {
        MPI_Recv(message, 30, MPI_CHAR, 0, 99, MPI_COMM_WORLD, &status);
    }
    printf("received : %s\n", message);
}


MPI_Gather(&num,1,MPI_INT,nums,1,MPI_INT,1,MPI_COMM_WORLD);
MPI_Scatter(nums,1,MPI_INT,&num,2,MPI_INT,0,MPI_COMM_WORLD);

n = 0;
```

(Name)	MPItest.cpp
Content	False
File Type	C/C++ Code
Full Path	f:\MyProject'
Included	True
Relative	..\MPItest.cp

(Name)
Names the file object.

WINDOWS MONITORING GOES “CLUSTER”

- Built-In Perfmon Access for Simple Monitoring



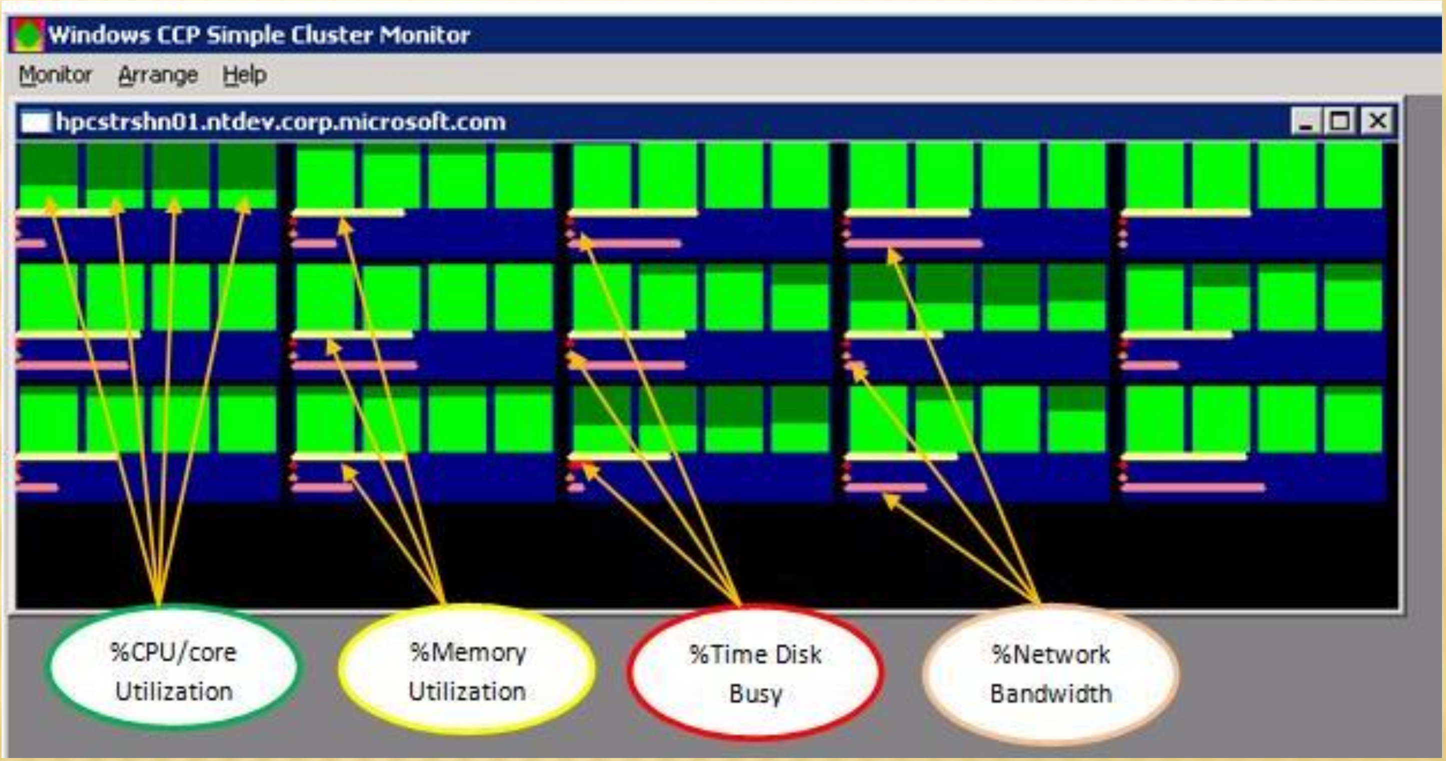
Last 100.000 Average 47.581 Minimum 0.004 Maximum 100.000

Duration 8:20

Color	Counter	Instance	Object	Computer
—	% Processor Time	_Total	Processor	\\HPCHEAD4056
—	Bytes Total/sec	OpenIB IPoIB ...	Network Interface	\\HPCHEAD4056
—	Current Bandwidth	OpenIB IPoIB ...	Network Interface	\\HPCHEAD4056
—	Node status	HPCHEAD4056	Compute Nodes	\\HPCHEAD4056
—	Number of busy processors	HPCHEAD4056	Compute Nodes	\\HPCHEAD4056
—	Number of idle processors	HPCHEAD4056	Compute Nodes	\\HPCHEAD4056
—	Number of jobs running	HPCHEAD4056	Compute Nodes	\\HPCHEAD4056
—	Number of parallel tasks	HPCHEAD4056	Compute Nodes	\\HPCHEAD4056

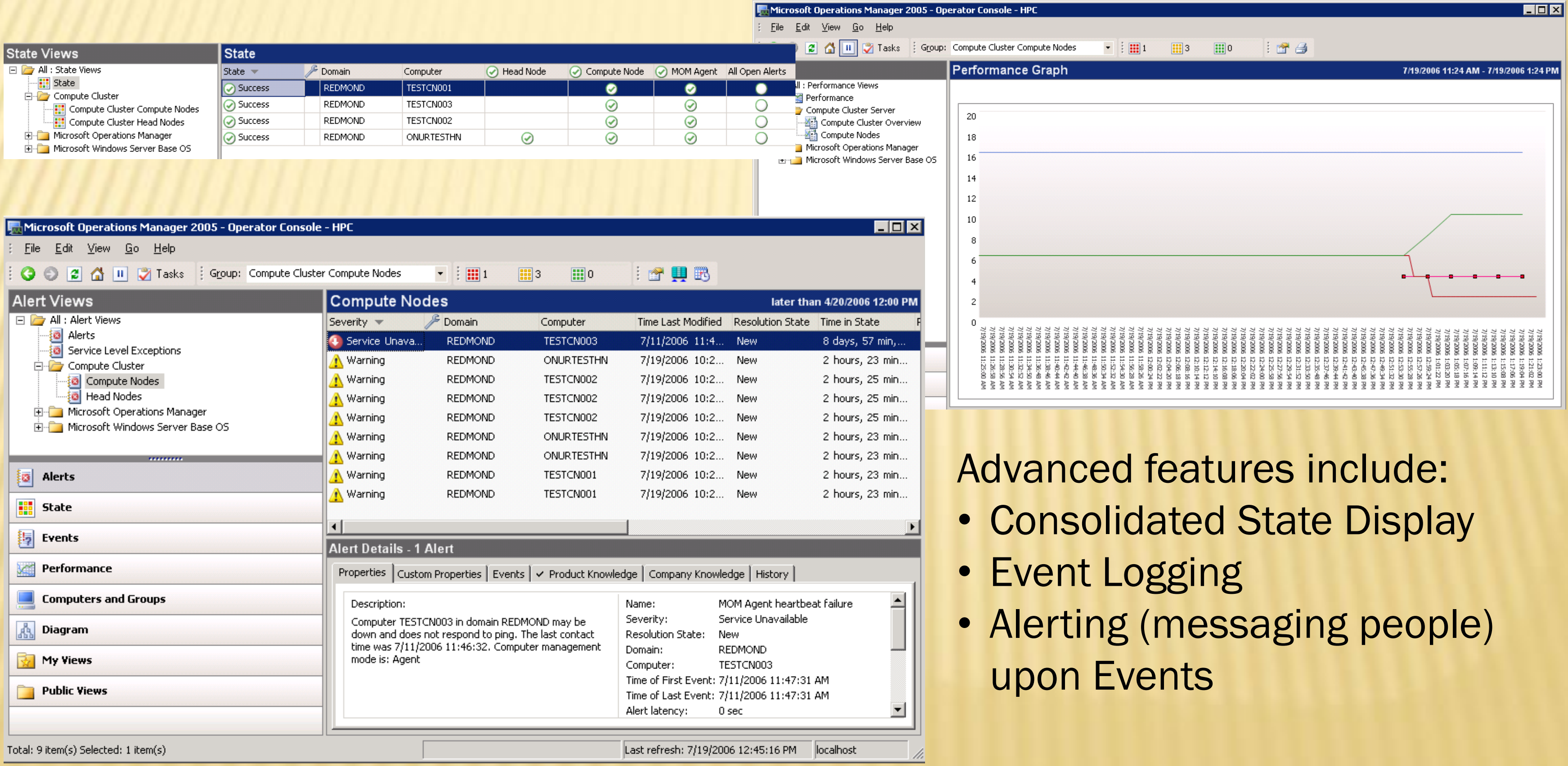
WINDOWS MONITORING GOES "CLUSTER"

- Monitoring At-A-Glance with clusmon
  - Free in the CCP Toolpack



WINDOWS MONITORING GOES "CLUSTER"

- Scale Up to Enterprise-Class Monitoring with Microsoft Operations Monitoring (MOM) & The CCS Mom-Pack



State Views

- All : State Views
  - State
  - Compute Cluster
    - Compute Cluster Compute Nodes
    - Compute Cluster Head Nodes
  - Microsoft Operations Manager
  - Microsoft Windows Server Base OS

State

State	Domain	Computer	Head Node	Compute Node	MOM Agent	All Open Alerts
✓ Success	REDMOND	TESTCN001		✓	✓	○
✓ Success	REDMOND	TESTCN003		✓	✓	○
✓ Success	REDMOND	TESTCN002		✓	✓	○
✓ Success	REDMOND	ONURTESTHN	✓	✓	✓	○

Microsoft Operations Manager 2005 - Operator Console - HPC

Group: Compute Cluster Compute Nodes

- All : Performance Views
  - Performance
  - Compute Cluster Server
    - Compute Cluster Overview
    - Compute Nodes
  - Microsoft Operations Manager
  - Microsoft Windows Server Base OS

Performance Graph 7/19/2006 11:24 AM - 7/19/2006 1:24 PM

Microsoft Operations Manager 2005 - Operator Console - HPC

Group: Compute Cluster Compute Nodes

Alert Views

- All : Alert Views
  - Alerts
  - Service Level Exceptions
  - Compute Cluster
    - Compute Nodes
    - Head Nodes
  - Microsoft Operations Manager
  - Microsoft Windows Server Base OS

Compute Nodes later than 4/20/2006 12:00 PM

Severity	Domain	Computer	Time Last Modified	Resolution State	Time in State
Service Unava...	REDMOND	TESTCN003	7/11/2006 11:4...	New	8 days, 57 min,...
Warning	REDMOND	ONURTESTHN	7/19/2006 10:2...	New	2 hours, 23 min...
Warning	REDMOND	TESTCN002	7/19/2006 10:2...	New	2 hours, 25 min...
Warning	REDMOND	TESTCN002	7/19/2006 10:2...	New	2 hours, 25 min...
Warning	REDMOND	TESTCN002	7/19/2006 10:2...	New	2 hours, 25 min...
Warning	REDMOND	ONURTESTHN	7/19/2006 10:2...	New	2 hours, 23 min...
Warning	REDMOND	ONURTESTHN	7/19/2006 10:2...	New	2 hours, 23 min...
Warning	REDMOND	TESTCN001	7/19/2006 10:2...	New	2 hours, 23 min...
Warning	REDMOND	TESTCN001	7/19/2006 10:2...	New	2 hours, 23 min...

Alert Details - 1 Alert

Description:

Computer TESTCN003 in domain REDMOND may be down and does not respond to ping. The last contact time was 7/11/2006 11:46:32. Computer management mode is: Agent

Name: MOM Agent heartbeat failure
Severity: Service Unavailable
Resolution State: New
Domain: REDMOND
Computer: TESTCN003
Time of First Event: 7/11/2006 11:47:31 AM
Time of Last Event: 7/11/2006 11:47:31 AM
Alert latency: 0 sec

Total: 9 item(s) Selected: 1 item(s)

Last refresh: 7/19/2006 12:45:16 PM

localhost

Advanced features include:

- Consolidated State Display
- Event Logging
- Alerting (messaging people) upon Events

- Microsoft Compute Cluster Server (CCS)
    - How Does Microsoft Describe the HPC Market?
    - What is CCS?
- Networking Options for CCS
    - Optimizing for Performance
- MS Investing in Infiniband
    - HPC Hosted Clusters
    - HPC Team clusters
- A Word about CCSv2

MS INVESTING IN INFINIBAND

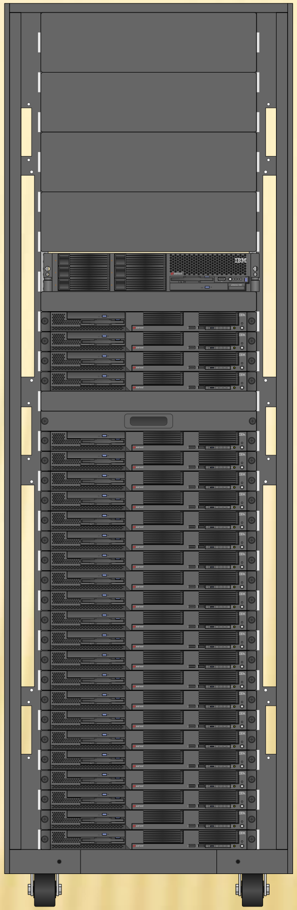
MS HPC INCUBATION & OPS TEAM

- Understand the value proposition of hosted clusters for end users and service providers
- Derive the cost of operating a hosted cluster environment
- Provide a user context for CCS team that is a source for short term and long term product feedback as well as best practices

OPS TEAM DOES HOSTED CLUSTERING

- Hardware
  - Facilities Planning (power & cooling)
  - Growth Management & Forecasting
  - Spares & Servicing
  - Deployment
- Systems Management & Monitoring
  - Alerting
  - Patching
  - Data and Applications Management
- Resource Allocation & Accounting (“Bill Bucks” only)
- User Docs & Help Desk Support

WORKGROUP CONFIGURATION: "MYRTO"



Purpose:

- Prototyping
- Dogfood
- Staging environment for Dept. Cluster
- Ops Experience

Configuration:

- 1 Head node
- 1 IIS server
- 18-23 compute nodes
- File server is on a machine separate from head
- Private Gb-E network for compute nodes
- Each compute node has dual-core AMD Opteron® 252, 2.6 Ghz, 2GB RAM

Currently used by:

- MS Research
  (Machine Learning)
- HPC Incubation Team

Statistics:

- Availability is ~99%
- Approx 3000-5000 jobs/month

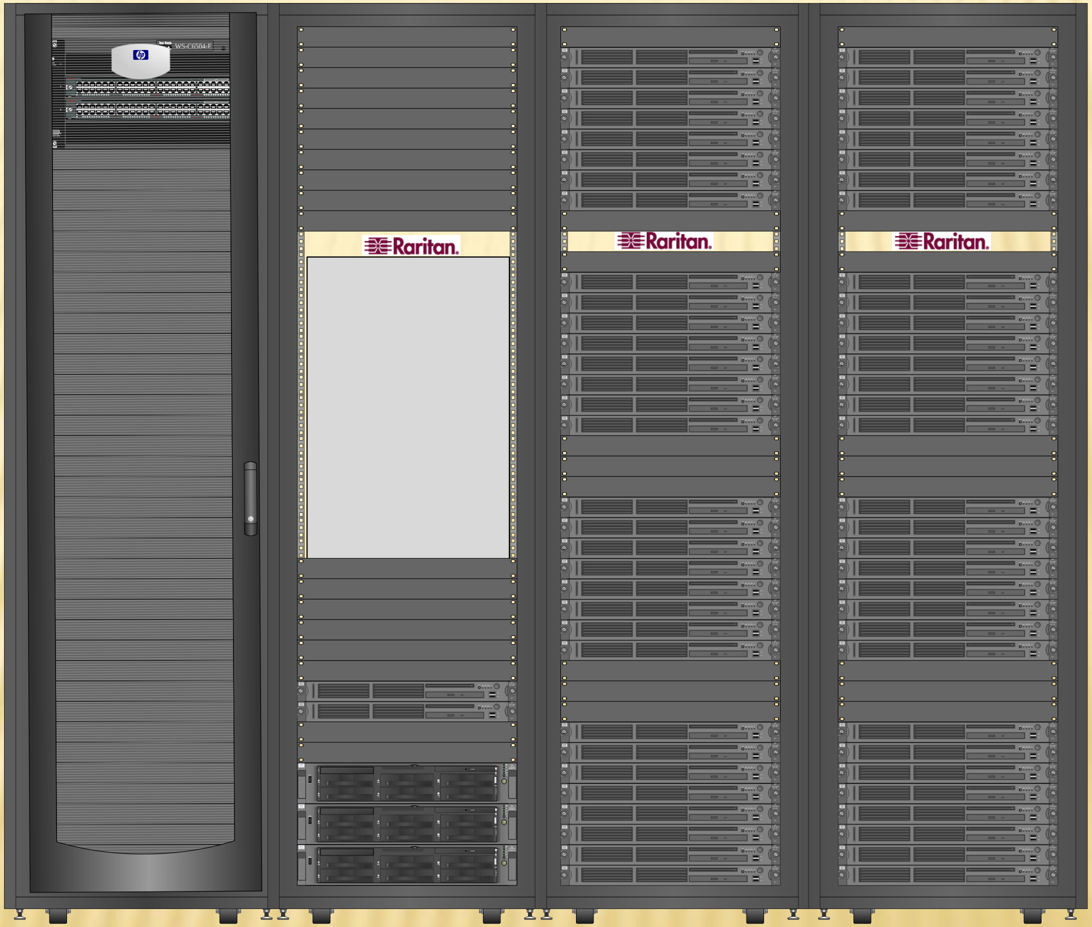
DEPARTMENTAL CONFIGURATION: "ATHENA"

Purpose:

- External Access
- Prototyping
- Ops Experience

Configuration:

- HP Servers
- 1 Head node
- 64 compute nodes
- 1 IIS server
- 1 File Server
- App/MPI: Myrinet
- Private: Gb-E
- Public: Gb-E
- Each compute node has two dual-core AMD Opteron™275, 2.2 Ghz, 8GB RAM



Users:

- HPC Incubation Team
- Partners

Location:

- Microsoft Partner Solutions Center (MPSC) – Building 25

ENTERPRISE CONFIGURATION: "RAINIER"

Purpose:

- External Access
- Prototyping at Scale
- ISV App testing at Scale
- Ops Experience

Configuration:

- 260 Dell Blade Servers
- 1 Head node
- 256 compute nodes
- 1 IIS server
- 1 File Server
- App/MPI: Infiniband
- Private: Gb-E
- Public: Gb-E
- Each compute node has two quad-core Intel 5320 Clovertown, 1.86GHz, 8GB RAM
- 34 Cisco SFS7000P SDR IB Switches in leaf & node configuration

- Total
  - 2080 Cores
  - 2+TB RAM

Users:

- MS Incubation Team
- ISV Partners
- MS Product team

Location:

- Microsoft Tukwila Data center (22 miles from Redmond campus)

MICROSOFT'S HPC TEAM USE INFINIBAND DAILY

Size	Usage	
9 nodes	MPI Development	Dual core, Dual proc
10 nodes	Test Automation Development	Dual core, Dual proc
6 nodes	MPI Test	
7 nodes	Test	
8 nodes	Performance Test	IB, GigE, and Myrinet cards on each node
16 nodes	ISV App Test	
260 nodes	Scale-Out Test	Rainer: 2080 cores

- Basis Of Weekly CCS Performance Benchmarking
  (We Track Perf Changes As We Code)
- Now Adding IB To The Clusters Used For Daily Build Verification
- Use Openfabrics Windows Drivers Exclusively On All Nodes
  (Go Openfabrics!!)

- Microsoft Compute Cluster Server (CCS)
  - How Does Microsoft Describe the HPC Market?
  - What is CCS?
- Networking Options for CCS
  - Optimizing for Performance
- MS Investing in Infiniband
  - HPC Hosted Clusters
  - HPC Team clusters
- A Word about CCSv2

A WORD ABOUT CCS V2

CCS NETWORKING ROADMAP

2008+

- Future version based on Windows Server codenamed “Longhorn”
  - Networking Mission: Scale
- MSMPI improvements
  - Low-latency, better tracing, multi-thread
- Network management
  - Driver and hardware settings configuration, deployment and tuning from new UI
  - ‘Toolbox’ of scripts and tips

2006

- CCS v1 networking based on Windows Server 2003
  - MSMPI and Winsock API
  - Both using Winsock Direct to take advantage of RDMA hardware mechanisms

LINKS

- Tuning Whitepaper
  - Windows Compute Cluster Server 2003: Performance Tuning White Paper
  - http://www.microsoft.com/downloads/details.aspx?FamilyID=40cd8152-f89d-4abf-ab1c-a467e180cce4&DisplayLang=en
- Winsock Direct QFE for Windows Server 2003 Networking
  - Only install the latest- QFEs are cumulative
  - Latest as of 04/15/07: KB924286
- CCS v1 SP1 released
  - Compatible with WinServer 2003 SP2 which includes all QFEs

LINKS (CON'T)

- Compute Cluster Server Case studies
  - http://www.microsoft.com/casestudies/
  - Search with keyword HPC
- Microsoft HPC web site (evaluation copies available)
  - http://www.microsoft.com/hpc/
- Microsoft Windows Compute Cluster Server 2003 community site
  - http://www.windowshpc.net/
- Windows Server x64 information
  - http://www.microsoft.com/64bit/
  - http://www.microsoft.com/x64/
- Windows Server System information
  - http://www.microsoft.com/wss/

THANK YOU