



By Aziz Gulbeden  
Mausmi Kotecha

# RUNNING MICROSOFT WINDOWS HPC SERVER 2008 ON DELL POWEREDGE CLUSTERS

The Microsoft® Windows® HPC Server 2008 high-performance computing (HPC) platform is designed to simplify the deployment, configuration, and management of HPC clusters while also integrating multiple performance enhancements. Running this platform on clusters of Dell™ PowerEdge™ servers can provide a highly available, high-performance foundation for HPC applications.

**M**icrosoft Windows HPC Server 2008 is designed to provide a simplified, highly available platform for high-performance computing (HPC) cluster environments. Built on Microsoft Windows Server® 2008 64-bit technology and integrating multiple enhancements to help increase network performance, it includes both Windows Server 2008 HPC Edition as the cluster OS and Microsoft HPC Pack 2008 to provide the necessary cluster utilities for running and managing HPC applications. By deploying this platform on clusters of Dell PowerEdge servers, cluster administrators can minimize the time required for cluster management and focus on running HPC applications productively.

#### Related Categories:

Dell PowerEdge servers  
High-performance computing (HPC)  
Microsoft  
Microsoft Windows HPC Server 2008

Visit [DELL.COM/PowerSolutions](http://DELL.COM/PowerSolutions) for the complete category index.

#### SIMPLIFIED CLUSTER CONFIGURATION AND MANAGEMENT

Microsoft Windows HPC Server 2008 includes HPC Cluster Manager, an integrated cluster management utility that helps simplify cluster deployment, configuration, and management, including key HPC tasks such as setting up head nodes and managing compute nodes. It incorporates wizard-based configuration tools, compute node templates, node monitoring and management tools, job monitoring and management tools, integrated diagnostic and

reporting utilities, and the Windows PowerShell™ command-line shell.

**Wizard-based configuration tools.** Administrators can use wizards to perform many initial configuration tasks in HPC Cluster Manager. When the application launches for the first time, it displays a To-Do List screen showing the wizards available for required configuration steps (see Figure 1). After the cluster is configured, administrators can perform cluster management tasks through corresponding Configuration, Node Management, Job Management, Diagnostics, and Charts and Reports panes of HPC Cluster Manager.

**Compute node templates.** Compute node templates specify the operations to be performed during compute node provisioning. To help simplify compute node management when compute nodes have different hardware configurations or different roles, Windows HPC Server 2008 allows administrators to create multiple compute node templates. These templates do not need to include an OS; administrators can include the OS as part of the template or install it separately, depending on their needs. During OS deployment, the templates specify whether the copy operation is to be performed through multicast, which helps reduce overall network traffic and accelerate

the simultaneous installation of multiple compute nodes.

Administrators can customize templates using the template editor, including installing additional patches or applications or running additional commands during deployment. Deployments are based on Windows Imaging (WIM) images, which administrators can create from the installation media. HPC Cluster Manager also allows administrators to inject drivers into these images. For advanced setups, administrators can also create images using the Microsoft Windows Automated Installation Kit. As updates for compute nodes become available, administrators can apply them using a node template.

**Node monitoring and management tools.** The Node Management pane in HPC Cluster Manager includes different views of the compute nodes for monitoring and lists the available actions administrators can perform on cluster nodes. A heat map allows real-time monitoring of selected metrics on the cluster nodes, which can be grouped and filtered for management or job scheduling.

**Job monitoring and management tools.** Administrators can schedule or monitor jobs using the Job Management pane in HPC Cluster Manager. A separate job scheduler application is available for nodes that are not part of the cluster but will be submitting jobs. Only domain users identified as cluster users are allowed to submit jobs, and only users identified as cluster administrators can perform administrative tasks such as canceling a job submitted by another user.

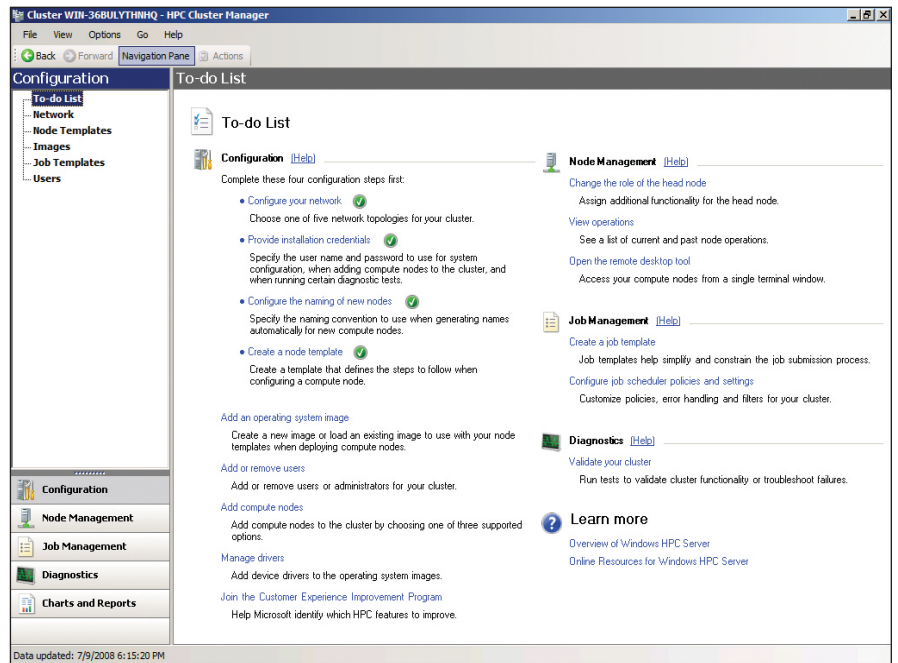


Figure 1. Wizards in Microsoft HPC Cluster Manager help simplify initial configuration tasks

The job scheduler also includes policies for service-oriented architecture applications and adaptive multilevel resource allocation, and allows administrators to launch jobs on a large number of nodes without any additional penalties compared with small jobs.

**Integrated diagnostic and reporting utilities.** HPC Cluster Manager provides a variety of tools for monitoring cluster performance and node health and to help troubleshoot the cluster if problems arise. The tests are categorized as scheduler tests, services tests, connectivity tests, system configuration tests, and service-oriented application tests. The reports monitor cluster metrics such as job throughput, cluster processor usage,

cluster network usage, node availability, and job turnaround.

**Windows PowerShell.** Windows PowerShell offers comprehensive management utilities through a command-line shell, enabling administrators to use commands or scripts to help effectively perform cluster operations.

## PERFORMANCE AND AVAILABILITY ENHANCEMENTS

Microsoft Windows HPC Server 2008 includes multiple enhancements to performance and availability, including Microsoft Message Passing Interface (MS MPI), event tracing, the Network Direct interface, a variety of networking changes and enhancements, and head node high availability.

**Microsoft Message Passing Interface.** Windows HPC Server 2008 comes with MS MPI, which provides the networking interface to the cluster applications and is compatible with the MPICH2 reference implementation. MPI-1 and MPI-2 are standard messaging interfaces defined for cluster applications that implement the required communication operations used by parallel applications. Multi-core

**“By deploying Microsoft Windows HPC Server 2008 on clusters of Dell PowerEdge servers, cluster administrators can minimize the time required for cluster management and focus on running HPC applications productively.”**

processor-based systems can benefit from the enhancements to shared memory communications in MS MPI.

**Event tracing.** To help increase cluster application performance, MS MPI includes integrated event tracing that compilers can use to optimize the code for a specific cluster. Event tracing helps simplify developer tasks by enabling them to tune performance and debug the application if they encounter any errors.

**Network Direct.** Network Direct is a Remote Direct Memory Access (RDMA) networking interface that uses a more direct path than Windows Sockets (Winsock) Direct to support networking hardware. This approach helps increase the performance and efficiency of MPI applications running over high-speed network fabrics.

**Networking enhancements.** Windows Server 2008 includes many networking changes and enhancements. The updated implementation of the TCP/IP stack included in Windows Server 2008 is called the Next Generation TCP/IP stack and is based on current networking technologies.<sup>1</sup> Key features of this stack include the following:

- **Receive Window Auto-Tuning:** The receive window size sets the maximum number of packets that have been received but not acknowledged. Receive Window Auto-Tuning allows the OS to dynamically adjust the receive

window to an optimal size based on the network condition, enabling increased network throughput between cluster nodes and increased network utilization during data transfer.

- **Compound TCP:** Compound TCP tunes the amount of data sent at a time in a packet by monitoring the bandwidth-delay product, network variations, and packet losses, with an increase in packet size that helps accelerate data transfer. Together, Receive Window Auto-Tuning and Compound TCP can help increase network utilization and optimize network performance for HPC applications.
- **Explicit Congestion Notification (ECN):** Network congestion can potentially cause dropped packets. ECN support for TCP enables routers experiencing congestion to mark their packets, with TCP peers then reducing their transmission rate on receiving marked packets to help avoid packet loss.
- **Server Message Block (SMB) 2.0:** Windows uses SMB for file sharing. Many HPC applications use file shares to communicate input, output, or intermediate result files. SMB 2.0 supports a larger buffer size than SMB 1.0, allows multiple SMB commands to be sent at once, and has increased limits on the number of concurrent open file handles.
- **TCP Chimney Offload:** TCP Chimney Offload delegates TCP traffic processing to TCP/IP Offload Engine (TOE)-capable

network adapters, helping free processor cycles for other application tasks.

- **Receive-side scaling:** Receive-side scaling distributes the processing of incoming traffic among multiple processors, enabling multi-core processor-based servers to handle incoming traffic more quickly than they could otherwise.

**Head node high availability.** Windows HPC Server 2008 enables administrators to set up the head node in a failover cluster configuration for high availability. If the cluster head node fails, the standby server in the failover cluster becomes active and then serves as the head node, enabling submitted jobs to continue running. Other head node tasks also migrate automatically to the standby server after a short pause. This high-availability configuration requires that the head node be running the Microsoft Windows Server 2008 Enterprise Edition or Datacenter Edition OS as well as the Microsoft SQL Server® 2005 database platform.<sup>2</sup>

## NETWORKING PERFORMANCE ON DELL POWEREDGE SERVERS

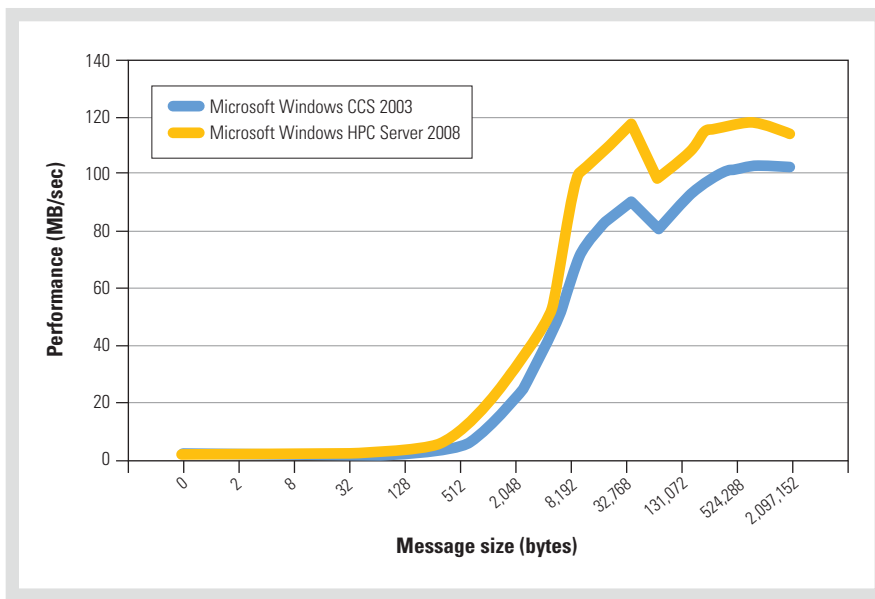
Dell supports Microsoft Windows HPC Server 2008 on Intel® Xeon® processor-based Dell PowerEdge 1950 III and PowerEdge 2950 III servers as head nodes and on Intel Xeon processor-based PowerEdge 1950 III, PowerEdge 2950 III, and PowerEdge M600 servers as compute nodes. Dell also supports Windows HPC Server 2008 on AMD Opteron™ processor-based PowerEdge 2970 servers as head nodes and on AMD Opteron processor-based PowerEdge SC1435, PowerEdge M605, PowerEdge M805, and PowerEdge M905 servers as compute nodes. For cluster storage, Dell PowerVault™ MD1000, PowerVault MD3000, and PowerVault MD1120 RAID arrays in direct attach storage mode can be attached to the head node.<sup>3</sup>

**“Using Microsoft Windows HPC Server 2008 on clusters of Dell PowerEdge servers can enable organizations to deploy highly manageable, highly available, high-performance clusters in their HPC environments.”**

<sup>1</sup> For more information about TCP/IP networking enhancements, visit [technet.microsoft.com/en-us/library/bb726965.aspx](http://technet.microsoft.com/en-us/library/bb726965.aspx).

<sup>2</sup> For more information, see the Windows HPC Server 2008 documentation.

<sup>3</sup> For more information on supported configurations, visit [DELL.COM/HPCC](http://DELL.COM/HPCC).



**Figure 2.** Microsoft Windows HPC Server 2008 provides increased network performance over Windows Compute Cluster Server 2003

To evaluate the networking performance enhancements in Windows HPC Server 2008, in July 2008 Dell engineers tested an example cluster consisting of 16 Dell PowerEdge 1950 servers, each configured with two quad-core Intel Xeon E5450 processors at 3.00 GHz, 4 GB of RAM, and Gigabit Ethernet interconnects. The tests were designed to compare networking performance on these nodes when running the Windows Compute Cluster Server (CCS) 2003 platform and when running Windows Server 2008 Standard Edition as the OS with Microsoft HPC Pack Release Candidate 1 (RC1)<sup>4</sup> as the cluster package.

The tests were based on the SendRecv parallel transfer benchmark in the Intel MPI Benchmarks 3.0 suite,<sup>5</sup> which is based on MPI\_SendRecv and is well suited for measuring bidirectional bandwidth. The processes in this benchmark form a virtual ring, with each process sending messages to the process on the right and receiving messages from the process on the left in the communication chain. For each process, the turnover

count is two messages per sample (one send and one receive). The test runs with varying message lengths starting at 0 bytes, then increasing to 1 byte, then doubling up to 4,194,304 bytes, with timings averaged over multiple samples.

Figure 2 shows the results, which demonstrate that Windows HPC Server 2008 performed better than Windows CCS 2003 for all message sizes. Performance did dip at around the 128 KB message size, which was attributable to the MPI protocol changing from the Eager protocol to the Rendezvous protocol. The Eager protocol transfers the MPI header and the message without waiting for the receiver to be ready, and is typically suitable for small message sizes. As message size increases, however, MPI switches to the Rendezvous protocol, in which an initial handshake occurs and the sender waits for the receiver to be ready with a buffer. Administrators can modify the point at which this switch occurs using the MPICH\_SOCKET\_EAGER\_LIMIT environment variable; the default value for this variable is 128 KB.

## SIMPLIFIED, HIGHLY AVAILABLE COMPUTE CLUSTER PLATFORM

Microsoft Windows HPC Server 2008 is designed to greatly simplify cluster management, helping reduce the time administrators spend managing a cluster and enabling them to focus on productively running HPC applications. Using this platform on clusters of Dell PowerEdge servers can enable organizations to deploy highly manageable, highly available, high-performance clusters in their HPC environments. [u](#)

**Aziz Gulbeden** is a systems engineer in the Scalable Systems Group at Dell. His current areas of focus include Microsoft Windows OS-based computer clusters and scalable file and storage systems. He has a B.S. in Computer Engineering from Bilkent University and an M.S. in Computer Science from the University of California, Santa Barbara.

**Mausmi Kotecha** is an engineering analyst in the Enterprise Solutions Group at the Dell Bangalore Development Center. Her areas of interest include HPC clustering packages, performance analysis of parallel applications, and cluster file systems. Mausmi has a bachelor's degree in Computer Science from Atmiya Institute of Technology and Science.

**MORE**  
**ONLINE**  
[DELL.COM/PowerSolutions](http://DELL.COM/PowerSolutions)

**QUICK LINKS**

**Dell HPC cluster solutions:**  
[DELL.COM/HPCC](http://DELL.COM/HPCC)

**Microsoft Windows HPC Server 2008:**  
[www.microsoft.com/hpc](http://www.microsoft.com/hpc)

<sup>4</sup> Available at [connect.microsoft.com](http://connect.microsoft.com).

<sup>5</sup> Available at [www3.intel.com/cd/software/products/asm-na/eng/219848.htm](http://www3.intel.com/cd/software/products/asm-na/eng/219848.htm).