NOVEMBER 2005 · $12.95

DELL™ POWER SOLUTIONS

THE MAGAZINE FOR DIRECT ENTERPRISE SOLUTIONS

Driving Clusters for High Performance and High Availability

Latest in High-Performance Computing Technology

Planning for Microsoft SQL Server 2005 Deployments

Building Grids with Oracle Database 10*g* RAC

Uptime expert.

It could be you.

Want to achieve a new level of reliability
while increasing server throughput?
Team multi-port Intel® PRO Server Adapters
with onboard connections.

Improved network uptime? Yes.
Increased bandwidth
and balanced traffic? Yes.
Bottlenecks? No way.

Intel® PRO
Network Connections

Whatever your infrastructure needs,
Intel® PRO Server Adapters
can help make network design easier.
Way easier.

Learn more: intel.com/go/adapters

intel®

DELL POWER SOLUTIONS
THE MAGAZINE FOR DIRECT ENTERPRISE SOLUTIONS
NOVEMBER 2005

COVER STORY | PAGE 8

Driving Clusters for High Availability and High Performance

The special clustering section in this issue is packed with expert guidance and best-practices techniques to help ensure high availability and high performance in a wide range of enterprise applications. Addressing concerns of small and growing data centers as well as global IT infrastructures, in-depth articles explore how to plan effectively for Microsoft SQL Server 2005 deployments, how to enable grids with Oracle Database 10*g* RAC, and how to leverage the latest advances in blade server clusters and high-performance computing clusters.

TABLE OF CONTENTS

DELL POWER SOLUTIONS | THE MAGAZINE FOR DIRECT ENTERPRISE SOLUTIONS
WWW.DELL.COM/POWERSOLUTIONS

EDITORIAL

EDITOR-IN-CHIEF | Tom Kolnowski
MANAGING EDITOR | Debra McDonald
SENIOR EDITOR | Liza Graffeo
CONTRIBUTING AUTHORS | Tim Abels; Alex Akinnuoye; Rizwan Ali; Andrew Bachler; Bala Beddhannan; Greg Benson; Bernard Briggs; Alan Brumley; Michael Brundridge; Bill Bryce; Fareed Bukhari; Balasubramanian Chandrasekaran; Alan Daughetee; Randy De Meno; David Detweiler; Narayan Devireddy; Jake Diner; Steve Fagan; Yung-Chin Fang; Anthony Fernandez; John Fruehe; Ron Gidron; Rinku Gupta; Jon Hass; Paul Hoke; Saeed Iqbal, Ph.D.; Simon Jandreski; Monica Kashyap; Zain Kazim; Greg Kincade; Florenz Kley; Kevin Knight; Garima Kochhar; Jukka Kouletsis; Chethan Kumar; Achim Lernhard; Robert Lesieur; Tong Liu; Zafar Mahmood; Douglas McDowell; J. Lance Miller; Farrukh Noman; J. Marcos Palacios, Ph.D.; Keith Priddy; Ramesh Radhakrishnan, Ph.D.; Anusha Ragunathan; Arun Rajan; Amina Saify; Simone Shumate; Sanjeev S. Singh; Jim Skirvin; Thorsten Staerk; Scott Stanford; Wolfgang Trenkle; Bryant Vo; and Brad Winett

ART

ART DIRECTOR | Iva Frank
DESIGNER AND ILLUSTRATOR | Cynthia Webb
COVER DESIGN | Iva Frank

MARKETING

MARKETING MANAGER | Kathy White

ONLINE

WEB PRODUCTION | Brad Klenzendorf

SUBSCRIPTION AND EDITORIAL SERVICES

EDITORIAL ASSISTANT | Amy Hargraves

Subscriptions are free to qualified readers who complete the online subscription form. To sign up as a new subscriber, renew an existing subscription, change your address, or cancel your subscription, submit the online subscription form at **www.dell.com/powersolutions_subscribe**, return the subscription reply form by surface mail, or fax the subscription reply form to +1 512.283.0363. For subscription services, please e-mail **us_power_solutions@dell.com**.

ABOUT DELL

Dell Inc., headquartered in Round Rock, Texas, near Austin, is the world's leading direct computer systems company. Dell is one of the fastest growing among all major computer systems companies worldwide, with approximately 47,800 employees around the globe. Dell uses the direct business model to sell its high-performance computer systems, workstations, and storage products to all types of enterprises. For more information, please visit our Web site at **www.dell.com**.

Dell Power Solutions is published quarterly by the Dell Product Group, Dell Inc. *Dell Power Solutions*, Mailstop 8456, Dell Inc., One Dell Way, Round Rock, TX 78682, U.S.A. This publication is also available online at **www.dell.com/powersolutions**.

November 2005

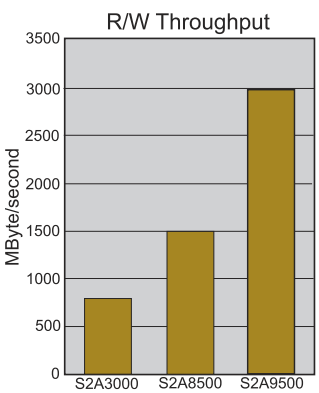
TALK BACK

We welcome your questions, comments, and suggestions. Please send your feedback to the *Dell Power Solutions* editorial team at us_power_solutions@dell.com.

Why Does DataDirect Storage Power 20 of the World's Fastest Computers?

The answer is simple – DataDirect's sixth generation S2A9500 storage controller boasts a world-class architecture with **FC-4** and **InfiniBand** front ends, **Fibre Channel** and **SATA** back-end disks, and ultra-high levels of performance, scalability, and reliability across a broad range of computational, visualization, and nearline environments.

Performance: Pushes the limits of computing with up to 3 GB/second of sustained bandwidth for a single system.



The S2A9500 delivers a staggering 3GB/second of sustained bandwidth to clusters and networked environments.

Scalability: Designed to extend to more than 1,000 disks, enabling hundreds of terabytes of virtualized capacity from a single, managed storage system.

Reliability: Enables multiple levels of redundancy and on-the-fly parity calculations on reads and writes, without adversely affecting data availability or system bandwidth.

DataDirect offers broad **infrastructure** support in the industry, with S2A storage installed behind processors, interconnects, or networks, and optimized for multiple operating systems and clustered file systems.

And our team of **experts** has successfully installed many petabytes of storage across a wide range of the world's most demanding networked environments.

Thousands of terabytes of DataDirect storage are being used in government, academia, and corporate environments worldwide. Visit our Web site for details about how these organizations have benefited from DataDirect products.

Wherever there is a need for speed, capacity, or reliability, DataDirect offers a storage solution to meet the challenge. Contact us today and let us show you how we can make your environment highly efficient, productive, and easy to manage.

DataDirect S2A Storage

Storage Solutions Designed for Performance, Scalability and Reliability

Please Call DataDirect
or your Dell™ Representative to show you why!
www.datadirectnet.com
(818) 700-7607
sales@datadirectnet.com

DataDirect™ NETWORKS

TABLE OF CONTENTS

ADVERTISER INDEX

SEE IT HERE FIRST!

Check the *Dell Power Solutions* Web site for our late-breaking exclusives, how-to's, case studies, and tips you won't find anywhere else. Want to search for specific article content? Visit our Related Categories index online at www.dell.com/powersolutions.

ONLINE EXTRA

DELL POWER SOLUTIONS MAGAZINE ARTICLES AVAILABLE ONLY AT WWW.DELL.COM/POWERSOLUTIONS

WWW.DELL.COM/ POWERSOLUTIONS

Deploying and Managing SQL Server 2005 Across the Scalable Enterprise

By Tim Abels

The first in a series presenting prescriptive architectures and operational techniques designed to realize the scalable enterprise tenets of simplified operations, improved utilization, and cost-effective scaling, this article examines best practices for deploying and managing Microsoft SQL Server 2005 applications. Guidelines for integration with Dell management tools and Microsoft Operations Manager (MOM) are accompanied by practical advice for efficient cluster and storage management.

End-to-End Change Management with the Dell OpenManage Server Update Utility

By Steve Fagan and J. Marcos Palacios, Ph.D.

The Dell OpenManage Server Update Utility is designed to identify, view, and automatically apply BIOS, driver, and firmware updates—helping to streamline the change-management process.

DRAC/MC User Management and Security Configuration

By Anusha Ragunathan and Sanjeev S. Singh

The Dell Remote Access Controller/Modular Chassis (DRAC/MC) is a critical infrastructure component for authenticating and authorizing user access to the Dell Modular Server Enclosure, the chassis that houses Dell blade servers.

Automated BIOS Management Using the Dell OpenManage Deployment Toolkit

By Zain Kazim, Alan Daughetee, and Bala Beddhannan

BIOS management can be a time-consuming and cumbersome task for enterprise IT administrators. The Dell OpenManage Deployment Toolkit enhances operational efficiency by enabling automated and scripted BIOS management for Dell PowerEdge servers—helping administrators perform BIOS and configuration updates simultaneously on multiple systems.

Unattended Installation of Dell OpenManage Server Administrator Using the Microsoft Windows Installer Command-Line Interface

By Alex Akinnuoye and Bernard Briggs

Dell OpenManage Server Administrator (OMSA) uses Microsoft Windows Installer (MSI) technology to perform installations, upgrades, modifications, and uninstallations on Windows platforms. Using MSI engine parameters, administrators can set up and customize OMSA through the OMSA–MSI command-line interface.

Optimizing Console Redirection for Dell PowerEdge Servers Using HyperTerminal and Telnet Clients

By Paul Hoke

Dell PowerEdge servers allow administrators to remotely control and configure key settings using a variety of interfaces. This article describes procedures to configure and operate HyperTerminal and Telnet clients to optimize console redirection features.

But the cheap one was completely inadequate and the expensive one was overkill, so she tried Galaxy Express for her data management software and it was just right. And her small company grew into a major world player and she lived happily ever after.

On her own island.

Galaxy Express data management software is easy to use, easy to afford and easy to scale. Because it was designed to work together, rather than be patched together. Galaxy Express. It's for smaller businesses that have no intention of staying that way. To get it working for you, visit **www.dell.com/galaxy** or contact your Dell sales representative.

CommVault®
Unified Data Management™

Multi-Core Moves Mainstream

If you are looking for evidence that Moore's Law is alive and well, look no further than the 300 mm silicon wafers incorporated into the latest generation of Intel® Pentium® D and Intel Xeon™ processors. It is not an expected die-size reduction or speed increase that is driving these processors to enable what may be a major boost in performance for workstations and servers. It is the number of processor *logic cores* on a single physical processor chip.

The term *multi-core* has now moved into the mainstream IT vernacular. Single-processor, dual-processor, and multi-processor terminology are well known and easily understood: the number of processors equates with the number of processor sockets built into the main printed circuit board. With the advent of multi-core platforms, however, things are not quite so elementary. For example, the single-processor server or workstation has morphed into the single-socket system. And that system may have a single-core processor or a multi-core processor installed in its socket (see Figure 1).

Predictably, the dual-processor server has become the dual-socket system, and it may be equipped with one or two single-core processor chips, or one or two multi-core processor chips. Figure 2 shows a dual-socket system populated with a pair of multi-core processors. Similarly, a quad-socket system can be populated with one to four single-core processors, or one to four multi-core processors. It is important to note, however, that single-core and multi-core processors cannot be intermixed in the same physical server.

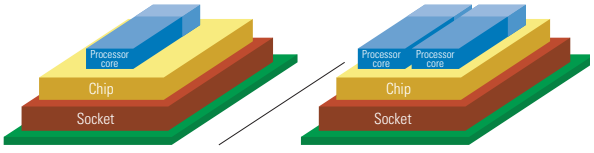


Figure 1. A single-socket system can be equipped with either a single-core processor (left) or a multi-core processor (right)

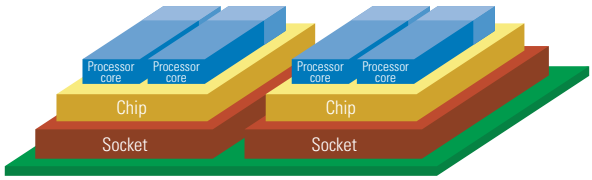


Figure 2. A dual-socket system equipped with multi-core processors

While we have barely scratched the silicon-based surface of multi-core technology here, you can read much more on this topic by turning to page 67. "Realizing Multi-Core Performance Advances in Dell PowerEdge Servers" describes potential performance gains and compatibility you may expect from Dell's multi-core product offerings—and how you may plan for integrating them cost-effectively into your data center environment.

But don't stop there. On page 8 our gateway cover story, "Driving Clusters for High Availability and High Performance," provides an executive overview of the extensive coverage in this issue's special clustering section. Key articles address Microsoft® SQL Server™ 2005, Oracle® clusters, blade server clusters, and the latest advances in high-performance computing clusters—providing technical content that may be indispensable when planning your cluster deployments.

Tom Kolnowski
Editor-in-Chief
tom_kolnowski@dell.com
www.dell.com/powersolutions

Backup Exec 10d.
Designed for disk.

The crowd goes wild. Because everything that made Backup Exec™ the gold standard in backup and recovery comes designed for disk: Backup Exec 10d. Comprehensive data protection that's fast, simple and thoroughly proven in the Windows® environment. Backup Exec 10d eliminates backup windows, improves reliability, and introduces the industry's first web-based file retrieval with continuous disk-based data protection. And Backup Exec 10d is now integrated with Symantec LiveState™ Recovery, a leader in hardware independent, advanced system recovery. Judge for yourself. Call 800-745-6045 or visit symantec.com/backupexec BE FEARLESS.

symantec™

Driving Clusters

for High Availability and High Performance

Best-practices architecture and operational techniques can help IT administrators respond to emerging business opportunities with speed and agility while helping to reduce data center complexity and cost. This special clustering section explores a wide range of high-availability and high-performance solutions, including how to plan effectively for Microsoft® SQL Server™ 2005 deployments, how to enable grids with Oracle® Database 10*g* Real Application Clusters, and how to leverage the latest advances in blade server clusters and high-performance computing clusters.

From simple two-node configurations running high-availability enterprise applications to high-performance computing (HPC) networks crunching unimaginably complex data sets across thousands of nodes, clusters have successfully invaded the territory once dominated by costly proprietary systems. Standards-based data center components—including servers, storage, network interconnects, and open source software—allow organizations of all sizes to cluster cost-effective computing power to strategic advantage.

Best-practices architecture and operational techniques can help administrators transform their existing IT framework into a scalable enterprise in practical, cost-effective phases. For example, administrators can start with consolidation and automation of cluster components, which lays the foundation for enterprise-wide resource management—leading to the ultimate goal of responding to emerging business opportunities with speed and agility. At every stage, Dell's scalable enterprise approach can help cluster architects and administrators to simplify operations, improve resource utilization, and scale cost-effectively.

Microsoft SQL Server 2005 clusters

For many organizations, high availability is not simply an option—it is a requirement. Enterprises depend on 24/7 availability for clusters running applications such as order processing and tracking, inventory control, transaction processing, customer support, and electronic commerce. Preparing properly for the transition to Microsoft SQL Server 2005 can help administrators to provide top-flight availability for application environments running on multi-node server clusters. "Reap the Benefits of SQL Server 2005" (page 51) explains advanced features in the latest Microsoft SQL Server release, and how Microsoft SQL Server 2005 is designed to provide exceptional data availability and manageability, hardened security, and the capability to scale from handheld mobile devices to multi-terabyte data warehouses.

In addition, Dell and Microsoft have teamed up to provide a comprehensive SQL Server 2005 solution that is designed to deliver exceptional performance and value

for critical database applications. Dell engineers tested the dual-socket Dell™ PowerEdge™ 2800 server running SQL Server 2005 on the TPC-C benchmark from the Transaction Processing Performance Council (TPC). The PowerEdge 2800 with one dual-core Intel® Xeon™ processor at 2.8 GHz achieved a price/performance ratio of US$0.99/transactions per minute (tpmC), becoming the first system to shatter the US$1/tpmC mark.[1]

The Dell and Microsoft solution includes high-availability clusters configured with Dell PowerEdge servers and Dell/EMC Fibre Channel–based or Dell PowerVault™ SCSI-based storage, as well as backup and recovery components. In the area of systems management, Dell's open standards approach and tight integration with Microsoft Operations Manager 2005 and Microsoft Systems Management Server 2003 enable efficient management of deployed SQL Server 2005 applications. And flexibility in SQL Server 2005 licensing, which ranges from prepackaging with Dell PowerEdge servers to volume enterprise licensing arrangements, can add particular value when coupled with Dell's SQL Server installation services, SQL Server 2000 migration services, or Enterprise Support services, which are designed to meet the unique needs of SQL Server 2005–based environments.

Oracle Database 10*g* clusters

In the past, scaling has often meant replacing hardware with faster, more expensive equipment. However, Oracle 10*g* Real Application Clusters (RAC) heralds an emerging model for high-availability application scaling. Instead of scaling up by replacing existing hardware with more powerful systems, Oracle 10*g* RAC allows organizations to scale out simply by adding computing power or storage to an existing database grid. Using Oracle 10*g* RAC with Dell PowerEdge servers and Dell/EMC Fibre Channel–based storage, enterprises can create a shared-storage database cluster in which cluster nodes have equal access to the same set of database objects. "Oracle 10*g* Real Application Clusters: Building and Scaling Out a Database Grid on Dell PowerEdge Servers and Dell/EMC Storage" (page 55) examines how to use Dell servers and Dell/EMC storage with Oracle 10*g* RAC to scale out database grids.

For small and growing enterprises, Dell and Oracle offer cost-effective, preconfigured RAC-based solutions on Microsoft Windows Server™ 2003, Standard Edition, with Service Pack 1. "Exploring Dell-Supported Configurations for Oracle Database 10*g* Standard Edition with RAC" (page 60) discusses how these highly available clusters can be deployed as two-node direct attach and Fibre Channel configurations using low-cost storage based on Dell/EMC AX100 and CX300 arrays.

Blade server clusters

For high-density computing environments, Dell's blade server architecture enhances availability while economizing on data center space with shared system components inside the Dell Modular Server Enclosure. Designed to work in a storage area network or with direct attach storage, blade server clusters can be preconfigured by Dell Services to help enterprises balance the need for performance, availability, and cost. In "High-Availability Blade Server Clustering with the Dell PowerEdge Cluster FE555W" (page 63), Dell engineers discuss high-availability configuration options for the PowerEdge 1855 blade server, which supports 10 removable server blades, fabric switches, and Ethernet switches in the 7U Dell Modular Server Enclosure.

High-performance computing clusters

Today, massively parallel HPC clusters are harnessed to solve computing problems that were once the exclusive domain of proprietary symmetric multiprocessing systems—problems as diverse as decoding the human genome, mapping the world's underground oil reserves, and digitizing the planet's music library. Several articles in this issue explain solutions that are designed to simplify HPC cluster deployment and systems management as well as enhance cluster utilization, thereby helping to reduce data center cost and complexity.

In "Dell OpenManage Tools for High-Performance Computing Cluster Management" (page 40), Dell engineers describe how to enhance cluster management with the Dell OpenManage™ remote management software suite. In addition, "Platform Rocks: A Cluster Software Package for Dell HPC Platforms" (page 29) demonstrates the use of Platform Rocks to help deploy, maintain, and manage HPC clusters. And "The Dell Life-Cycle Approach to Implementing HPC Cluster Services" (page 46) explains how Dell Services follows a comprehensive life-cycle approach to help organizations plan, deploy, and manage HPC clusters.

By understanding best-practices architecture and operational techniques, administrators can build upon a foundation of standards-based cluster components to manage change efficiently—transforming their existing IT framework into a scalable enterprise in practical, cost-effective phases.

[1] TPC-C performance results are based on benchmark tests performed by Dell labs in September 2005 on a PowerEdge 2800 server with one dual-core Intel Xeon processor at 2.8 GHz and 2 MB level 2 cache per processor core, 8 GB of error-correcting code (ECC) double data rate 2 (DDR2) SDRAM, an 800 MHz frontside bus, a Peripheral Component Interconnect (PCI) Express bus, one 36 GB SCSI drive, Windows Server 2003, and SQL Server 2005—resulting in TPC-C performance of 38,622 tmpC, price/tpmC of US$0.99, and a system availability date of November 8, 2005. Actual performance will vary based on configuration, usage, and manufacturing variability. The top TPC-C price/performance results are available at www.tpc.org/tpcc/results/tpcc_price_perf_results.asp.

Configuring the BMC and BIOS

on Dell Platforms in HPC Cluster Environments

High-performance computing clusters continue to grow dramatically—a 256-node cluster is now considered the average size and large clusters can range into thousands of nodes. Deploying, maintaining, and monitoring large clusters can become problematic. To help reduce cluster deployment time, administrators can use Dell™ OpenManage™ Server Administrator in conjunction with Platform Rocks to configure each node.

BY GARIMA KOCHHAR, RIZWAN ALI, AND ARUN RAJAN

Related Categories:

Baseboard management controller (BMC)

Cluster management

Clustering

Dell PowerEdge servers

High-performance computing (HPC)

Remote management

System deployment

Systems management

Visit www.dell.com/powersolutions for the complete category index.

High-performance computing (HPC) clusters, also known as Beowulf clusters, combine cost-effective, standards-based symmetric multiprocessing (SMP) systems together with high-speed interconnects to achieve the raw computing power of traditional supercomputers. Twice a year, the TOP500 Supercomputer Sites[1] list of the 500 most powerful computer systems is compiled. From June 2000 to June 2005, the percentage of clusters in this list grew from 2.2 to 60.8 percent.[2] In the same time period, the typical size of these clusters grew from a few hundred processors to thousands of processors. Today, the largest Intel® processor–based cluster on the list consists of more than 10,000 processors.[3] These deployments indicate that the challenges for implementing HPC clusters no longer relate to feasibility and acceptance but rather to the scale of the clustered system.

All typical server configuration, management, and maintenance activities are magnified in the HPC environment. In clusters, any manual administrative interaction for these activities can be prohibitive in terms of time and effort. As clusters continue to grow, connecting a keyboard, monitor, and mouse to each node in the cluster—even for a small configuration step—is not an option.

Three key components affect the management of Dell-based clusters: the baseboard management controller (BMC), the Intelligent Platform Management Interface (IPMI),[4] and the Dell OpenManage software suite.[5] Platform Rocks,[6] which is based on National Partnership for Advanced Computational Infrastructure (NPACI) Rocks[7] developed by the San Diego Supercomputer Center (SDSC), is a comprehensive cluster management toolkit that is designed to simplify the deployment and management of large-scale Linux® OS–based clusters. Developed by Platform Computing Inc., Platform Rocks can be used along with Dell OpenManage to configure cluster nodes for remote management via IPMI and console redirection.

[1] The TOP500 project was started in 1993 to provide a reliable basis for tracking and detecting trends in high-performance computing; for more information, visit www.top500.org.

[2] Results for the June 2000 list are available at www.top500.org/lists/2000/06; results for the June 2005 list are available at www.top500.org/lists/2005/06.

[3] This cluster is NASA's Columbia Supercomputer; for more information, visit www.top500.org/sublist/System.php?id=7288.

[4] For more information about IPMI, visit www.intel.com/design/servers/ipmi/index.htm.

[5] For more information about Dell OpenManage systems management software, visit www1.us.dell.com/content/topics/global.aspx/solutions/en/openmanage?c=us&cs=555&l=en&s=biz.

[6] For more information about Platform Rocks, visit www.platform.com/products/Rocks.

[7] For more information about NPACI, visit www.rocksclusters.org.

Understanding cluster management

HPC clusters, by definition, comprise several standards-based components: servers, switches, interconnects, and storage. The capability to streamline the management of a large number of parts is essential to help ensure efficient, cost-effective cluster operation.

Dell PowerEdge™ servers are equipped with software and hardware that allow for easy manageability. Each eighth-generation Dell PowerEdge server has a BMC, which is an on-board microcontroller that complies with IPMI 1.5. It allows administrators to monitor components and environmental conditions such as fans, temperature, and voltage inside the system chassis and enables access to the platform even when a server is powered down or the OS is hung.[8]

Dell servers also include the Dell OpenManage software suite. This suite of tools is designed to help simplify the management of server and storage hardware. One tool in this suite, Dell OpenManage Server Administrator (OMSA), allows administrators to monitor the health of a system, access asset and inventory information, analyze logs, update firmware and BIOS, and diagnose problems. OMSA also provides a Linux-based command-line interface (CLI) utility, omconfig, which allows configuration of the BMC and the BIOS of the server. This utility can be used to configure cluster nodes for remote management. By configuring each node's BMC and BIOS appropriately, administrators can set up the nodes for IPMI traffic and console redirection. In addition, the Platform Rocks cluster deployment package can use omconfig to configure the cluster nodes during deployment.

Using Platform Rocks to deploy and maintain HPC clusters

NPACI Rocks is an open source, Linux-based software stack for building and maintaining Beowulf clusters.[9] It is designed to make clusters easy to deploy, manage, maintain, and scale, and it is built on standard and open source components. Platform Rocks is a comprehensive cluster solutions package that is based on NPACI Rocks and includes drivers for Dell hardware and Red Hat® Enterprise Linux in addition to other features.

The Rocks software stack provides a mechanism to produce a customized distribution for a cluster node. This distribution defines the complete set of software and configuration for a particular node.

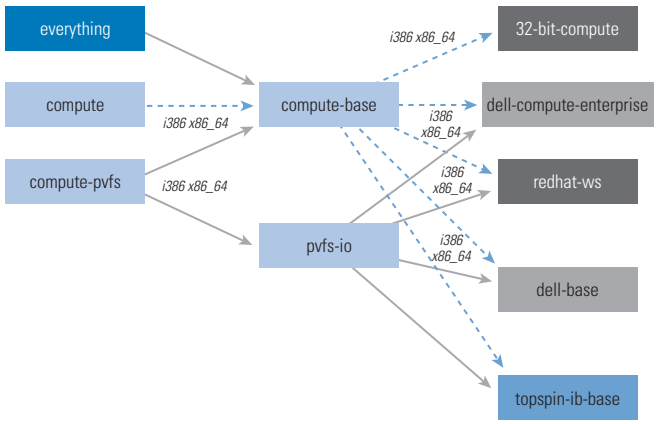


Figure 1. Portion of sample Platform Rocks kickstart graph

A cluster may require several node types, including compute nodes, I/O nodes, and monitoring nodes. Within a distribution, node types are defined with a system-specific Red Hat kickstart file, made from a Rocks kickstart graph. A Red Hat kickstart file is a text-based description of the software packages and software configuration to be deployed on a node. The Rocks kickstart graph is an XML-based tree structure used to define Red Hat kickstart files.[10]

Rocks generates kickstart files for compute nodes dynamically using the kickstart graph, which consists of the edges and nodes, both described in an XML language.[11] Each node specifies a service and its configuration, while edges between nodes specify membership in a kickstart file. Figure 1 shows a portion of a kickstart graph. The kickstart file for a compute node type, for example, is generated from the kickstart graph as follows:

- Use the configuration from the compute.xml file present in the compute node of the kickstart graph
- Follow all the outgoing edges from the "compute" node (see the dashed line from "compute" to "compute-base" in Figure 1)
- Include all contents of the child node's XML files (in this example, that would be compute-base.xml)
- Follow all the outgoing edges of the child node (see the dashed lines going out from "compute-base" in Figure 1)

[8] For more information about the BMC and IPMI, see "Remote Management with the Baseboard Management Controller in Eighth-Generation Dell PowerEdge Servers" by Haihong Zhuo; Jianwen Yin, Ph.D.; and Anil V. Rao in *Dell Power Solutions,* October 2004, www.dell.com/downloads/global/power/ps4q04-20040110-Zhuo.pdf; and "Managing and Monitoring High-Performance Computing Clusters with IPMI" by Yung-Chin Fang, Garima Kochhar, and Randy DeRoeck in *Dell Power Solutions,* October 2004, www.dell.com/downloads/global/power/ps4q04-20040138-Fang.pdf.

[9] For more information about NPACI Rocks, see "Streamlining Beowulf Cluster Deployment with NPACI Rocks" by Rinku Gupta, Yung-Chin Fang, and Munira Hussain in *Dell Power Solutions,* February 2005, www.dell.com/downloads/global/power/ps1q05-20040176-Gupta.pdf; and "Platform Rocks: A Cluster Software Package for Dell Platforms" by Rizwan Ali, Rinku Gupta, Garima Kochhar, and Bill Bryce in *Dell Power Solutions,* November 2005, www.dell.com/downloads/global/power/ps4q05-20040227-Ali.pdf.

[10] For more information about the kickstart file and kickstart graph, see www.rocksclusters.org/rocks-documentation/3.3.0/introduction.html; and "Rolls: Modifying a Standard System Installer to Support User-Customizable Cluster Frontend Appliances" by Greg Bruno, Mason J. Katz, Federico D. Sacerdoti, and Philip M. Papadopoulos from the *2004 IEEE International Conference on Cluster Computing,* July 2, 2004, www.rocksclusters.org/rocks-documentation/3.3.0/papers/cluster2004-roll.pdf.

[11] For more information, see www.rocksclusters.org/rocks-documentation/reference-guide/3.2.0/kickstart-xml.html, www.rocksclusters.org/rocks-a-palooza-slides/roll-development-basics.pdf, and www.rocksclusters.org/rocks-a-palooza-slides/rolls-overview.pdf.

This process repeats itself until the leaf nodes are reached. In the example scenario shown in Figure 1, the leaf node XML files would include:

- **dell-base.xml and dell-compute-enterprise.xml:** These include Dell-specific driver updates.
- **redhat-ws.xml:** This includes all Red Hat Package Manager (RPM™) packages in Red Hat Enterprise Linux WS.
- **32-bit-compute.xml:** This includes 32-bit RPM packages for 32-bit applications running on Intel Extended Memory 64 Technology (EM64T) installations.
- **topspin-ib-base.xml:** This includes Topspin InfiniBand drivers.

Note: This example scenario is based on the kickstart graph generated by Rocks 3.3.0. In Rocks 4.0 and later, the kickstart generation differs.

```
## Installing OMSA, configuring BMC, configuring
   BIOS and grub.conf for console redirection.
   Logs captured in /root/dell-omsa.log

## Create directory for OMSA /opt/omsa
mkdir /opt/omsa &gt;&gt; /root/dell-omsa.log
   2&gt;&amp;1
cd /opt/omsa &gt;&gt; /root/dell-omsa.log
   2&gt;&amp;1

## Retrieve OMSA tar Ball and config-omsa input
   file from the frontend
wget -nv https://<<frontend-ip address>>/install/
   rocks-dist/lan/enterprise/3/en/os/<<arch>>/
   RedHat/RPMS/om4.4.tgz &gt;&gt; /root/
   dell-omsa.log 2&gt;&amp;1
wget -nv https://<<frontend-ip address>>/install/
   rocks-dist/lan/enterprise/3/en/os/<<arch>>/
   RedHat/RPMS/config-omsa &gt;&gt; /root/
   dell-omsa.log 2&gt;&amp;1

## Install OMSA - Server Administrator CLI in /
   opt/dell and start OMSA     services
/bin/tar xvfz om4.4.tgz &gt;&gt; /root/
   dell-omsa.log 2&gt;&amp;1
cd linux/supportscripts &gt;&gt; /root/
   dell-omsa.log 2&gt;&amp;1

## Note: We pass the config-omsa input file for
   inputting our responses to the prompts the
   srvadmin-install.sh script makes.
./srvadmin-install.sh &lt; /opt/omsa/config-omsa
   &gt;&gt; /root/dell- omsa.log 2&gt;&amp;1

## BMC IP will be set to: 20.255.$addr where $addr
   is the last two octets of the compute node's OS
   level eth0 IP.
## The LAN Alert Destination will be set to $bmc_gw

gige_if=eth0
bmc_gw=20.255.1.1
bmc_mask=255.255.0.0
addr=`ifconfig $gige_if |grep "inet addr" |awk
   '{print $2}' |awk -F : '{print $2}' |awk -F .
   '{print $3"."$4}'`

##Configure the BMC
## In the BMC, enable IPMI over LAN
omconfig chassis bmc config=nic enable=true
## In the BMC, set the BMC IP address, netmask and
   gateway
omconfig chassis bmc config=nic ipsource=static
   ipaddress=20.255.$addr subnet=$bmc_mask
   gateway=$bmc_gw
## In the BMC, set the LAN Alert Destination
omconfig system pedestinations index=1
   ipaddress=$bmc_gw destenable=true
## In the BMC, enable the platform alert traps
omconfig system platformevents alertsenable=true

##Configure the BIOS for Console redirection and
   set the failsafe baud rate to 19200
omconfig chassis biossetup attribute=fbr
   setting=19200

##for PE1850 (For other Dell PowerEdge servers you
   may need to modify accordingly.)
## In the BIOS, set console redirection on and
   through the Serial Port
omconfig chassis biossetup attribute=serialport1
   setting=bmcnic
omconfig chassis biossetup attribute=conredirect
   setting=enable

##Set up grub.conf and /etc/inittab for OS-level
   console redirection.
/bin/cat /etc/grub.conf|sed -e "s/
   noexec=off/noexec=off console=tty0,19200
   console=ttyS0,19200/g" &gt;grub.conf
/bin/cp -f grub.conf /etc/rocks.conf &gt;&gt; /
   root/dell-omsa.log 2&gt;&amp;1
echo "co:2345:respawn:/sbin/agetty -h -L 19200
   ttyS0 ansi" &gt;&gt; /etc/inittab
```

Figure 2. Sample script to be added to post-installation section of extend-compute.xml

To alter default node types, administrators can extend or replace default nodes—for example, the "compute" default node could be extended or replaced by creating corresponding extend-compute.xml or replace-compute.xml files, respectively. To extend a default node, the contents of the default node XML file must be concatenated with the contents of its "extend" node XML file. While the kickstart file is created for a compute node, the kickstart graph traversal follows the compute.xml file to the extend-compute.xml file, allowing administrators to broaden the compute node definition.

Using an "extend" XML file is one way to add customizations to the standard Rocks node definition. Another method is to add a new type of node to the kickstart graph and to set up edges from predefined node types to this new node based on the customization required. This second method is used in Platform Rocks 4.0.0-1.1 and Platform Rocks 3.3.0-1.2, which include a new node called "dell-bmcbios-setup." The XML file for this node encapsulates the BMC and BIOS configuration to enable IPMI traffic and to set up console redirection. Following instructions in the Dell readme file distributed with Platform Rocks, administrators can connect this new node to client.xml (Platform Rocks 4.0.0-1.1), allowing all appliance types to inherit properties of dell-bmcbios-setup.xml.

Automating BMC and BIOS configuration with Dell OpenManage

The ability to extend a compute node definition can also be used to include Dell OpenManage RPM packages. Administrators can then use the tools provided by Dell OpenManage to configure the BMC and BIOS of the cluster node during the post-installation phase. A single installation and a single reboot can help ensure that the cluster nodes are installed with the OS as well as configured for IPMI traffic and console redirection.

Platform Rocks 4.0.0-1.1 and Platform Rocks 3.3.0-1.2 already include this configuration feature using a predefined dell-bmcbios-setup.xml file. Administrators should follow the readme file instructions included with Platform Rocks to enable this feature for a cluster. The following steps describe an alternative method for configuring the BMC and BIOS using Platform Rocks 3.3.0.x:

1. From the Dell OpenManage Server Administrator CD,[12] create a tarball that includes the OMSA Linux RPM packages and the OMSA installation scripts, and put this tarball in the appropriate directory on the front-end node following the instructions available at www.rocksclusters.org/rocks-documentation/3.3.0/customization-adding-packages.html. In the example scenario, all the subdirectories on the OMSA CD under /cdrom/srvadmin/linux were collected to create om4.4.tgz, and this tarball was placed under /home/install/contrib/enterprise/3/public/*arch*/RPMS on the front-end node, where *arch* refers to the architecture (i386 or x86_64) of the cluster.
2. Following the instructions cited in the previous step, create an extend-compute.xml script on the front-end node. During the post-installation of each compute node, this script copies the OMSA tarball to the compute node, installs OMSA, and then configures the BMC and BIOS of the server.

   Use the srvadmin-install.sh script included on the OMSA CD to install OMSA. This interactive script requires user input to configure and install OMSA correctly. Because the compute nodes are identical, the information that needs to be provided is identical. For this purpose, create a config-omsa file, which contains answers to the questions that the srvadmin-install.sh script asks.

   In the post-installation section of the extend-compute.xml file, add the script shown in Figure 2. This figure shows a sample script to set up the BMC and BIOS configuration.
3. Rebuild the distribution following the instructions cited in step 1.

The cluster is now ready to be installed following the regular Rocks cluster installation steps. Once the compute nodes are installed, they will boot up with the BMC configured, the BIOS configured for console redirection, and the OS-level files modified to allow OS-level console redirection.

Managing HPC environments with Platform Rocks and Dell OpenManage

Using Platform Rocks, cluster administrators can automate the configuration of the BMC and BIOS of compute nodes in an HPC cluster, significantly reducing deployment time. With the omconfig utility in Dell OpenManage Server Administrator, administrators can configure other components of the BIOS. The configuration method described in this article can provide administrators with the flexibility to perform any desired customizations to a cluster node.

Garima Kochhar is a systems engineer in the Scalable Systems Group at Dell. She has a B.S. in Computer Science and Physics from Birla Institute of Technology and Science (BITS) in Pilani, India, and an M.S. in Computer Science from The Ohio State University.

Rizwan Ali is a systems engineer in the Scalable Systems Group at Dell. He has a B.S. in Electrical Engineering from the University of Minnesota.

Arun Rajan is a systems engineer in the Scalable Systems Group at Dell. He has a B.S. in Electronics and Communications Engineering from the National Institute of Technology, Tiruchirappalli, in India and an M.S. in Computer and Information Science from The Ohio State University.

[12] To download Dell OpenManage Server Administrator, visit support.dell.com/support/downloads/type.aspx?c=us&cs=555&l=en&s=biz&SystemID=PWE_1850&category=36&os=LE30&osl=en&deviceid=2331&devlib=36.

Achieving Continuous Availability for HPC Clusters with the

IBRIX Fusion Cluster File System

The IBRIX Fusion file system is a highly available cluster file system that can help fulfill I/O subsystem throughput and scalability requirements for high-performance computing (HPC) clusters. To help ensure availability and data consistency of file system I/O services even when file system components fail, the IBRIX Fusion file system offers flexible and scalable high-availability functions based on its innovative segmented architecture. In this article, Dell engineers demonstrate how the IBRIX Fusion file system maintained data access availability in three failure scenarios: HBA failure, segment server failure, and NIC failure.

BY TONG LIU; YUNG-CHIN FANG; RAMESH RADHAKRISHNAN, PH.D.; AND AMINA SAIFY

Related Categories:

Characterization

Clustering

Dell PowerEdge servers

High-performance computing (HPC)

IBRIX file system

Linux

Parallel systems

Performance

Visit www.dell.com/powersolutions for the complete category index.

A high-performance computing (HPC) cluster composed of cost-effective, industry-standard systems running a Linux® OS has become a widely accepted architecture for solving large-scale scientific computation problems that require massive computational power. As problem complexity and data-set size grow, increasing amounts of data must be processed by these scientific applications. However, I/O subsystems within HPC clusters typically cannot keep pace with today's escalating performance requirements. The traditional Network File System (NFS) architecture is limited by single-server bandwidth and can be a single point of failure—a vulnerability that can dramatically constrain scientific applications by degrading performance, scalability, and reliability. To break through the I/O bottleneck and enhance the I/O capabilities of cost-effective, industry-standard hardware in HPC clusters, IT organizations can deploy parallel file systems.

IBRIX Fusion, a cluster file system developed by IBRIX, Inc., is designed to provide file system capabilities well beyond the challenges of today's I/O-intensive requirements.[1] The I/O nodes in an IBRIX Fusion environment—known as *segment servers*—maintain the metadata and lock the files that are stored in their segments. The IBRIX Fusion file system pulls together the file systems and storage capacity of each segment server into a pool of storage resources that can be easily managed by a central management

[1] For more information, see "Achieving Scalable I/O Performance in High-Performance Computing Environments" by Amina Saify; Ramesh Radhakrishnan, Ph.D.; Sudhir Srinivasan, Ph.D.; and Onur Celebioglu in *Dell Power Solutions,* February 2005, www.dell.com/downloads/global/power/ps1q05-20040171-Saify.pdf.

interface known as IBRIX FusionManager. IT administrators can build a single IBRIX Fusion file system to support multiple petabytes of data accessible by tens of thousands of nodes—with throughput designed to deliver hundreds of gigabytes per second. In addition, the IBRIX Fusion file system can tolerate component failures, allowing segment servers to be temporarily unavailable without disrupting running applications. Data reading and writing tasks can run successfully across the IBRIX Fusion file system even when failovers and failbacks occur, and the IBRIX Fusion file system is designed to retain data integrity after failovers.

Achieving high availability with the IBRIX Fusion file system

As shown in Figure 1, the IBRIX Fusion file system provides several high-availability (HA) features designed to detect and recover quickly from various hardware failures and thus to help keep data accessible continuously.

The IBRIX Fusion file system includes the following failover components:

- Segment servers
- Host bus adapters (HBAs)
- Network interfaces
- File system segment access
- IBRIX FusionManager configuration

Segment servers

IBRIX Fusion segment servers can be configured to provide standby service for one another in standard 1 × 1, *N* × 1, and *N* × *M* modes, where *N* is the number of segment servers and *M* is the number of backup segment servers. In those configurations, all segment servers serve the file system simultaneously and confirm their current status with IBRIX FusionManager, which regularly verifies that all segment servers are operational and accessible. If a segment server misses five consecutive status updates, FusionManager automatically initiates failover operation—all the segment management duties belonging to the failed server will be gracefully migrated to its standby segment server. In this scenario, the failed segment server will be automatically powered down by FusionManager through a programmable power source to avoid a "split-brain" scenario in which the failed segment server and its standby server cannot communicate with each other. When the failed segment server is fixed, administrators can initiate failback procedures through the FusionManager graphical user interface (GUI) or command-line interface (CLI) to switch segment ownership from the standby server back to the original server.

If a segment server misses five consecutive status updates, FusionManager automatically initiates failover operation—all the segment management duties belonging to the failed server will be gracefully migrated to its standby segment server.

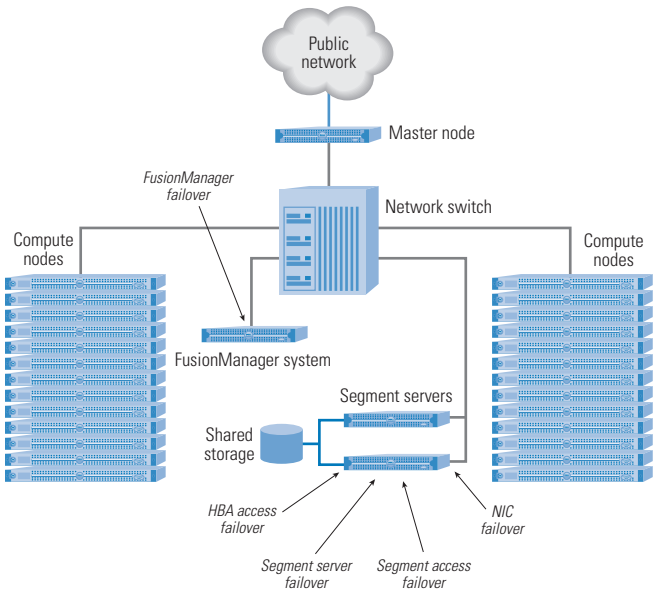


Figure 1. High-availability configuration using the IBRIX Fusion file system

Host bus adapters

The IBRIX Fusion file system supports two HA configurations for HBAs on segment servers: active/standby and segment-server failover. In both modes, a segment server reports a failure to IBRIX FusionManager when it cannot communicate with an HBA. If the segment server's HBAs are configured in active/standby mode and the HBAs have the embedded function to provide standby switching, then connections to the failed HBA will be automatically transferred to the standby HBA. If no functioning HBAs reside on the segment server or if the HBAs do not support standby switching, FusionManager will trigger segment-server failover—transferring the workload to a different server—to maintain the data access provided by the original server.

Network interfaces

The IBRIX Fusion file system can be exported for access via NFS. In this scenario, network interface card (NIC) failure will disrupt all connections to the IBRIX Fusion file system for NFS clients. To supply continuous data access for NFS users, the IBRIX Fusion file system allows users to set up a network interface with a standby network interface—virtual or physical—on another segment server that

has access to the same file system segments. When the monitored network interface is not functioning properly, IBRIX FusionManager will move that interface's IP address to the configured standby network interface on another segment server. NFS network traffic will then be directed to the standby network interface.

File system segment access

In the IBRIX Fusion file system, segments are building blocks used to allocate storage capacity. As long as segments can be accessed by clients through the segment servers, the file system can accept data requests. To help protect against file system failure, the IBRIX Fusion file system provides standby servers for segments, enabling each segment to be reached by multiple segment servers. Migrating ownership of the segments to the assigned standby server is the basic method used by the IBRIX Fusion file system to help safeguard data availability. If a segment server fails, IBRIX FusionManager enables the failed server's segments to be accessed through an alternative segment server without interruption.

IBRIX FusionManager configuration

The failure of IBRIX FusionManager may not impede file system operations because FusionManager's metadata management function is distributed across multiple segment servers. However, FusionManager holds the configuration information for the file system and manages the file system components. Therefore, FusionManager failure may interrupt file-system management functionality. IBRIX provides a detailed backup and restore solution for easily recovering FusionManager with the appropriate backup database.

Configuring the test environment

To demonstrate the failover capability of the IBRIX Fusion file system, a team of Dell engineers conducted a test on an IBRIX Fusion environment in June 2005. The team used the Platform Rocks toolkit—which Platform Computing Inc. developed based on the widely used National Partnership for Advanced Computational Infrastructure (NPACI) Rocks toolkit from the San Diego Supercomputer Center (SDSC)—to install the test cluster. Platform Rocks supports installation using Rocks Roll software packages just as NPACI Rocks does. In this test, the IBRIX Fusion File System Roll provided by IBRIX, Inc., was used to build the IBRIX Fusion file system during the initial cluster deployment. This Roll helps simplify the file system installation by building FusionManager, segment servers, and IBRIX clients with the user-friendly Rocks wizard.

In the test setup of the IBRIX Fusion file system, one FusionManager system and two segment servers, named ibrix-ss-0-0 and ibrix-ss-0-1, were installed on Dell™ PowerEdge™ 1850 servers. Figure 2 shows the hardware details for this configuration. Both segment servers were connected to the same Dell/EMC CX700 storage array, and ibrix-ss-0-0 was configured to be the standby server for ibrix-ss-0-1 in all the test

Cluster components	Hardware
FusionManager system	Dell PowerEdge 1850 with 2 Intel® Xeon™ processors at 3.2 GHz, 800 MHz frontside bus, and 2 GB of RAM
Segment servers (2)	Dell PowerEdge 1850 with 2 Intel Xeon processors at 3.2 GHz, 800 MHz frontside bus, and 4 GB of RAM
Compute nodes (8)	Dell PowerEdge 1850 with 2 Intel Xeon processors at 3.2 GHz, 800 MHz frontside bus, and 4 GB of RAM
HBA (1 per segment server)	QLogic QLA2340
Storage system	Dell/EMC CX700 with 73 GB 15,000 rpm drives
Ethernet controller (1 per compute node)	Intel Gigabit Ethernet* NIC

**This term does not connote an actual operating speed of 1 Gbps. For high-speed transmission, connection to a Gigabit Ethernet server and network infrastructure is required.*

Figure 2. Hardware used in the test environment

scenarios. The Dell team created six segments across the file system and assigned three segments to each segment server. Eight compute nodes were dedicated to performing parallel data reading and writing tasks. To observe the capability of the IBRIX Fusion file system to maintain continuous data access and data integrity for files with different sizes, the team created a 5 GB file to serve as a large data size example as well as hundreds of small files with a total size of 3 GB to be used as an example for small-data-size operations. In this test, the Dell team used a script developed in-house, in which all clients simultaneously read various-size saved sample files and duplicate those files eight times with different names on the same file system.

Examining IBRIX Fusion file system availability

To effectively observe the HA feature of the IBRIX Fusion file system, the Dell team configured only the segments on ibrix-ss-0-1, which was the monitored segment server, with a preferred segment allocation policy to accept reading and writing requests from clients. Thus, any failure on ibrix-ss-0-1 that did not spawn a successful failover function would interrupt the job immediately. The team simulated the following failure scenarios to investigate the IBRIX Fusion file system's capabilities to provide continuous data availability.

HBA failure

In this scenario, the Dell team planned to test the HA configuration of HBAs in which failure of the HBA should trigger failover of the segment server to its standby server. First, the team enabled IBRIX Fusion monitoring for the HBA on ibrix-ss-0-1 and assigned ibrix-ss-0-0 as its standby server. Then, the team started the script from the eight client nodes to repeatedly copy the 5 GB file and the 3 GB of small files to another directory on the IBRIX Fusion file system with different names. After a short time, the Dell engineers unplugged the fabric cable from the monitored HBA, which was connected to the Dell/EMC CX700 storage array.

As expected, data copying was not corrupted and the job resumed after a short halt. This result indicates that ibrix-ss-0-0

successfully took over ownership of segments belonging to ibrix-ss-0-1. At the end of the test, all the files were successfully generated on the IBRIX Fusion file system as scripted. The team found no differences when comparing those files with the original files using the Linux `diff` command. For this test, the overall running time was slightly longer compared with normal execution time (without HBA failure). The main reason for this overhead is that the IBRIX Fusion file system spends five minutes waiting on five consecutive failed heartbeat messages before triggering failover.

Segment server failure

In the IBRIX Fusion file system, client requests must go through segment servers to reach segments on the file system. This process requires IBRIX FusionManager to migrate segments from the failed segment server to its standby server. To examine this capability, the Dell team simulated segment server failure by powering down the monitored segment server (ibrix-ss-0-1). Similar to the HBA failure test, all jobs were temporarily halted for about five minutes to wait for failover completion. After the jobs resumed reading and writing files, the Dell engineers powered up ibrix-ss-0-1 and initiated the failback function on FusionManager.

The entire failback process completed within one minute, and no job failed during this operation. Once all the clients finished the script execution, the team compared the output with the output of a non-failed scenario and observed no differences. Thus, the IBRIX Fusion file system demonstrated the ability to maintain data access availability during segment server failover and failback.

NIC failure

In the final test, the Dell team demonstrated NIC failover capability when the IBRIX Fusion file system is exported by NFS. The team used FusionManager to set up a standby NIC on ibrix-ss-0-0 for eth0 on ibrix-ss-0-1. A directory within the IBRIX Fusion file system was exported by NFS. All the clients executed the test script over the NFS mount directory. While the test script was generating files on the NFS exported directory, the Dell engineers unplugged the network cable from eth0 on ibrix-ss-0-1.

The failover status of ibrix-ss-0-1 was reported in IBRIX FusionManager. By the time the status changed to "failed over," the NFS network traffic had been transparently redirected to the standby NIC on ibrix-ss-0-0, and all the jobs had resumed working on the clients. The Dell team checked the final results by using the `diff` command to help ensure that all the files were created accurately.

Enhancing HPC cluster availability using the IBRIX Fusion file system

Standards-based, cost-effective HPC clusters are now an established architecture for parallel and distributed computing. Cluster file systems are widely used in HPC cluster environments to help alleviate the bottleneck introduced by disparities between I/O and processor performance. Cluster file systems have been widely implemented on HPC clusters in downtime-sensitive environments that require fault tolerance and high availability.

Besides its contributions to an HA cluster management system, an HA file system can be a key component for building a reliable HPC cluster. The IBRIX Fusion file system enables promising I/O throughput and balanced resource utilization by providing a global view of heterogeneous networked storage. Furthermore, it offers a flexible and scalable HA file system with component-level, server-to-server, and active-active failover features. The IBRIX Fusion file system can monitor a variety of hardware and software elements, detect failures, and automatically switch control to another server with sustained data access. Constructing HPC clusters with Platform Rocks and the IBRIX Fusion file system can be a highly effective approach for HPC environments.

Tong Liu is a systems engineer in the Scalable Systems Group at Dell. His current research interests are HPC cluster management, high-availability HPC clusters, and parallel file systems. Tong serves as a program committee member of several conferences and working groups on cluster computing. Before joining Dell, he was an architect and lead developer of High Availability Open Source Cluster Application Resources (HA-OSCAR). Tong has an M.S. in Computer Science from Louisiana Tech University.

Yung-Chin Fang is a senior consultant in the Scalable Systems Group at Dell. He specializes in HPC systems, advanced HPC architecture, and cyberinfrastructure management. Yung-Chin has published more than 30 conference papers and articles on these topics. He also participates in HPC cluster–related open source groups as a Dell representative.

Ramesh Radhakrishnan, Ph.D., is a member of the Scalable Systems Group at Dell. His interests include performance analysis and characterization of enterprise-level applications. Ramesh has a Ph.D. in Computer Engineering from The University of Texas at Austin.

Amina Saify is a member of the Scalable Systems Group at Dell. Amina has a bachelor's degree in Computer Science from Devi Ahilya University (DAVV) in India and a master's degree in Computer and Information Science from The Ohio State University.

FOR MORE INFORMATION

IBRIX Fusion file system:
www.ibrix.com

NPACI Rocks:
www.rocksclusters.org/Rocks

Platform Rocks:
www.platform.com/products/Rocks

High-Performance Computing and the SMASH Initiative

To enhance management interoperability and help reduce total cost of ownership across heterogeneous nodes in high-performance computing (HPC) clusters, IT organizations can implement systems that comply with the Systems Management Architecture for Server Hardware (SMASH) initiative. The SMASH initiative is a suite of specifications designed to standardize management interfaces for heterogeneous computing environments and to provide an architectural framework that includes unified interfaces, resource discovery, resource addressing, and data model profiles. In this way, SMASH not only addresses complicated administrative challenges, but it also enables HPC clusters to enhance resource utilization and system uptime.

BY YUNG-CHIN FANG AND JON HASS

Related Categories:

Cluster management

Clustering

Dell PowerEdge servers

High-performance computing (HPC)

Parallel systems

Remote management

Standards

Systems management

Visit www.dell.com/powersolutions for the complete category index.

High-performance computing (HPC) clusters are widely used in industry, research, and academic segments for geophysics, semiconductor, telecommunication, database, digital content creation, weather and climate research, automotive, software, finance, and other research and development purposes. On the June 2000 TOP500 Supercomputer Sites list, only 2.2 percent of the supercomputer systems were cluster-based. By 2005, 60.8 percent of the supercomputer systems on the TOP500 list were cluster-based, while the percentage of massively parallel processing (MPP) and constellation architectures dropped rapidly from five years earlier.[1] Figure 1 shows the increasing prevalence of cluster architectures among the TOP500 Supercomputer Sites from 2000 to 2005.

The need for manageability

Besides gaining popularity among architecture types, HPC clusters are also scaling out. All the systems listed on the June 2005 TOP500 Supercomputer Sites list are equipped with at least 200 processors and some systems are equipped with thousands of processors. The direct total cost of ownership (TCO) for an HPC cluster is in proportion to the scale of the data center, which includes computing resource depreciation, facility rent and maintenance costs, utility bills, system administrator costs, and other financial overhead. When an HPC cluster is equipped with hundreds or thousands of processors, such factors can contribute considerably to TCO. A well-designed, comprehensive remote management framework

[1] For more information, see the TOP500 Supercomputer Sites Web site at www.top500.org. Results are available at www.top500.org/lists/2000/06 for the June 2000 list and at www.top500.org/lists/2005/06 for the June 2005 list.

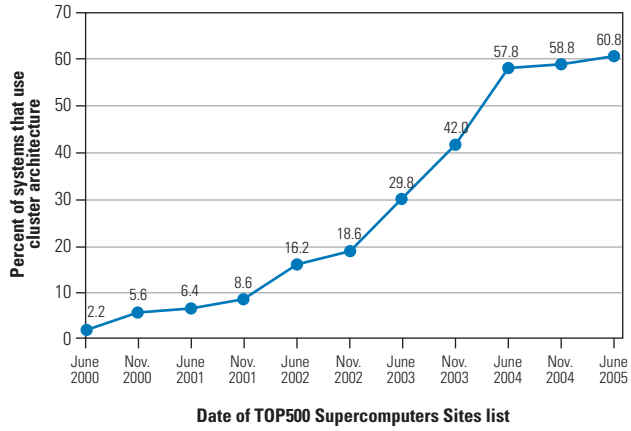


Figure 1. Growing use of cluster architecture among TOP500 Supercomputer Sites

can help reduce TCO by enhancing system uptime, resource utilization, and yield; streamlining administrative operations; and monitoring the health of cluster components to help address warning conditions before they result in system failure.

The need for interoperability

TCO not only includes direct costs, but it also includes indirect costs such as support, training, and retooling. Data centers commonly operate generations of heterogeneous hardware. Each platform comes with a management framework, which typically contains a rich set of tools and utilities. In many cases, these tools are specialized and adapted to an individual environment, installation, and product in the data center.

Usually, system administrators must learn all the different management frameworks in a heterogeneous environment before using the corresponding remote management frameworks to perform one task—such as remote power cycling of all nodes. Multiple management frameworks require more administrator and user training, higher support costs, and longer training time than a single management framework. Such factors contribute to increased TCO. Many organizations are finding that a unified, interoperable management interface and framework has become a necessity to meet enterprise requirements to scale out HPC clusters quickly, flexibly, and cost-effectively.

The SMASH specifications suite

In January 2004, the Server Management Working Group (SMWG) was established in the Distributed Management Task Force (DMTF). The SMWG initiated the Systems Management Architecture for Server Hardware (SMASH)[2] initiative to address the interoperable manageability requirements of small to large-scale heterogeneous computing environments. The DMTF has more than 3,000 active participants across industries and leads the development of management standards and integration technology for enterprise and Internet environments. The DMTF governs several specifications including System Management BIOS (SMBIOS),[3] Common Information Model (CIM),[4] Desktop Management Interface (DMI),[5] Web-Based Enterprise Management (WBEM),[6] and Alert Standard Format (ASF).[7] SMWG members include Dell, HP, IBM, Intel, Newisys, OSA Technologies, Sun, and others. Dell has committed more than 50 professionals to participate in the DMTF and has made significant contributions to the SMASH initiative as well.

SMASH comprises a suite of specifications that deliver architectural semantics, standardized server management protocols, and hardware data model profiles designed to help unify data center management. SMASH includes the Server Management (SM) Command Line Protocol (CLP) specification, the SM Managed Element Addressing specification, the SM CLP-to-CIM Mapping specification, the SM CLP Discovery specification using the Services Locator Protocol (SLP), the scripting specification, and several dozen system and component data model profile specifications.

System administrators can use a consistent SMASH command-line interface (CLI) to monitor and manage heterogeneous cluster hardware resources remotely, update firmware, and perform inventory. The CLI can be used to monitor and remotely manage the health status of components in large heterogeneous clusters to help overcome differences in hardware architecture, OS dependencies, and issues stemming from different command sets and utilities in existing management frameworks from different vendors. Through the CLI, administrators can tailor and automate computer- and data center–specific management tasks such as remotely changing the cluster's BIOS boot order, remotely power cycling hung nodes, and remotely updating firmware in parallel. A SMASH-compliant implementation is designed to improve interoperability and manageability—enabling optimal resource utilization and uptime while helping to reduce the indirect costs of training, support, and retooling. In this way, SMASH can help minimize TCO while enhancing reliability, availability, and manageability.

SM CLP. This specification defines the syntax and semantics of a small set of verbs that act consistently on heterogeneous hardware

[2] For more information about SMASH, visit www.dmtf.org/standards/smash.

[3] For more information about SMBIOS, visit www.dmtf.org/standards/smbios.

[4] For more information about CIM, visit www.dmtf.org/standards/cim.

[5] For more information about DMI, visit www.dmtf.org/standards/dmi.

[6] For more information about WBEM, visit www.dmtf.org/standards/wbem.

[7] For more information about ASF, visit www.dmtf.org/standards/asf.

systems and components represented by CIM-based data models. The CLP can be implemented in different ways, including in band, out of band, and via proxy. The CLP and the SMASH architecture are designed to be independent of machine state, OS state, server system architecture, and access method. The variety of ways the CLP can be implemented can facilitate existing local and remote management components without requiring additional memory and CPU resources on compute nodes. The unified command protocol is designed to be user-friendly and simple on both existing and future clusters.

SM Managed Element Addressing. This specification defines the formulation of command target addresses that resemble the naming conventions of hierarchical file systems. It specifies the user-friendly class tags and implied association classes that may be used to construct paths to address any managed element appearing within the scope of the manageability access point (MAP). The MAP is a collection of system services that provide management in accordance with specifications published under the DMTF SMASH initiative. An important aspect of MAP operations management is the capability to support simultaneous sessions through the MAP, thus unleashing the potential of remote parallel management functionality.

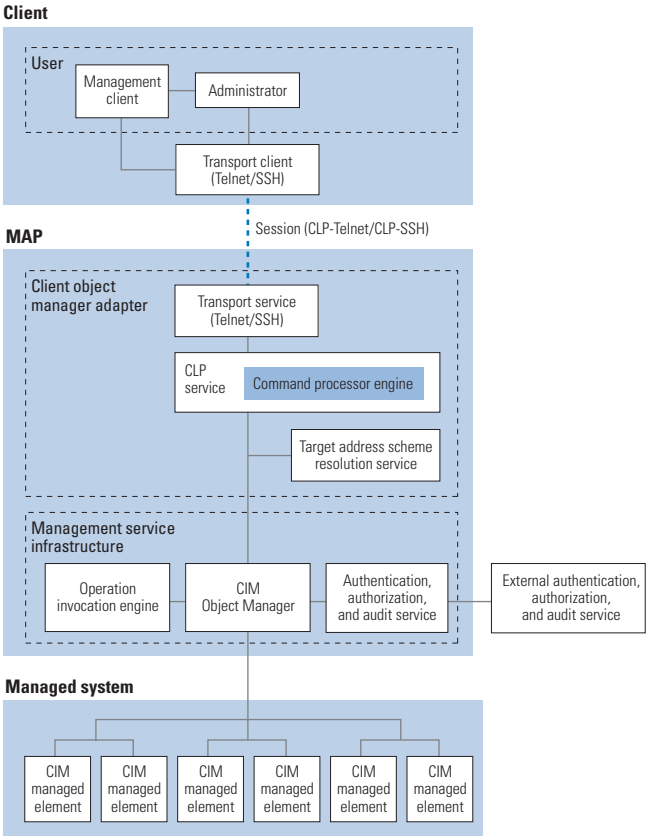


Figure 2. Example SMASH architecture

SM CLP-to-CIM Mapping. CIM provides a common definition of management information for systems, components, networks, applications, and services, and it allows for vendor extensions. CIM common definitions enable vendors to exchange semantically rich management information between systems throughout management fabrics. The CLP-to-CIM Mapping specification details how the CLP command verbs manipulate or act on the CIM, thus enabling a WBEM/CIM-compliant interface to the hardware-level management provided by the Intelligent Platform Management Interface (IPMI), ASF, and other device-level hardware management interfaces. The mapping also enables the CLP to potentially apply to existing CIM-based management frameworks.

SM CLP Discovery using SLP. This specification, leveraging SLP, addresses three aspects of discovery: how a client discovers which managed elements the MAP manages, the discovery of the capabilities of the MAP itself, and the discovery of the service access points of the MAP. The MAP is a network-accessible service for managing a system. A MAP can be instantiated by a management process, a management processor, a service process, or a service processor.

SMASH profiles. Server management profiles provide the object model definitions for manageability content and the architecture models for mapping computer hardware to fully connected association graphs in a manner that is consistent between different implementations. Profiles describe the legal classes and associations that can be used to model system and component hardware. The profiles can be reused and combined in different combinations to help ensure that all instances in the system are implemented in a consistent manner across multiple vendor architectures and offerings. User-friendly views based on profiles are defined to help simplify managing system boot, power, storage, firmware change management, system configuration, and hardware asset inventory. The Boot Control Profile is an example SMASH profile; it can be used to define boot order and other configuration aspects of boot devices.

Example SMASH architecture

Administrators issue standardized CLP commands to either a management client or a MAP directly. The MAP includes a CLP command/response protocol and processor engine: a text command message is transmitted from the administrator over the transport service, such as Telnet or Secure Shell (SSH), to the MAP; the MAP receives the command, authenticates the use/command privilege, and processes the command by converting it to a CIM model manipulation (see Figure 2). The command target address is resolved to the appropriate instance of the CIM object representing the component being manipulated. When the model is manipulated, a CIM provider is engaged and it translates the manipulation to a native hardware command. The response from a managed system or component is then transmitted from the MAP back to the client/administrator in

CLP verb	Definition
help	Retrieves context-sensitive help (same as the -help option with the addition of help for targets)
cd	Sets the current default target (navigates the target address space of the MAP)
show	Shows the values of a property or the contents of a collection/target
set	Sets a property or a group of properties to a specific value
exit	Terminates a CLP session
reset	Resets the target
start	Starts the target
stop	Stops the target

Figure 3. Example SMASH CLP command verbs

human-readable text or XML format. Figure 3 shows a few examples of SMASH CLP command verbs.

The CLP command/response message output can be in human-readable text, or for further processing, XML or comma-separated value (CSV) format. For example, the following command shows the current sensor reading of power supply 3 in system 1:

```
Prompt> show -display associations -o
   format=clpxml /system1/powersup3
```

The assigned output format is XML. System administrators can then direct the XML information to other applications for further processing to meet specific needs.

The implication of SMASH for HPC clusters

In the HPC cluster deployment phase, the CLP can be scripted to boot up an additional HPC cluster remotely—and properly, to avoid a power surge that could damage the system. Administrators can use the Boot Control Profile to define the boot order. For example, the boot order can specify that the first boot should be from the network via the Preboot Execution Environment (PXE) to remotely deploy the OS and predefined software stack to all nodes, while the second boot can be from a local hard drive to complete the necessary cluster configuration. Because HPC clusters usually consist of a large number of nodes, remote diagnostic services can also be invoked in this phase to examine cluster-wide hardware health status. Configuration activities can be invoked to stage cluster hardware and to stabilize the cluster operating state. The Power Control Profile, Boot Control Profile, and Diagnostics Profile can help reduce deployment time. The number of days saved during deployment time can translate into additional days of production time and fewer days of overall hardware depreciation time—thereby enhancing cost-effectiveness.

In the HPC cluster operational phase, SMASH can be used to monitor and remotely manage the hardware health status of heterogeneous HPC clusters and grids to help prevent hardware failure and the need to rerun parallel jobs as well as to help reduce recovery time. For example, when a SMASH-compliant, one-to-many management console reports an unusual memory-bit error count or a nonfatal SMART (Self-Monitoring, Analysis, and Reporting Tools) error in the hard drive, system administrators can respond by launching runtime job migration or check-pointing to preserve the current computing process, suspend the job, swap out the potential problem hardware, and restart the job. SMASH also can be used to reduce cluster maintenance time by facilitating activities such as remote, parallel firmware updates of heterogeneous clusters and remote power management for post-OS updates and upgrades. The CLP also can be tightly integrated with existing job schedulers to enhance overall hardware utilization, because new job scheduling schemes such as temperature- and power-aware scheduling or run-time environment–sensitive scheduling become achievable.

Enhanced HPC cluster deployment, management, and operations

SMASH-compliant implementations are designed to help solve many of the management difficulties prevalent in heterogeneous HPC clusters. By enhancing the efficiency of an HPC cluster's deployment and operational phases, SMASH-compliant systems can help minimize hardware depreciation and operational costs while enabling administrators to enhance resource utilization and system uptime—responding quickly, flexibly, and cost-effectively to business-critical scale-out requirements. In addition, administrators who do not have to learn and use multiple management frameworks can spend less time training and more time producing results.

Yung-Chin Fang is a senior consultant in the Scalable Systems Group at Dell. He specializes in HPC systems, advanced HPC architecture, and cyberinfrastructure management. Yung-Chin has published more than 30 conference papers and articles on these topics. He also participates in HPC cluster–related open source groups as a Dell representative.

Jon Hass is a software architect with the Dell Systems Management Architecture and Standards team. He is vice-chair of the DMTF CIM Core Model Working Group and is the Dell representative on the DMTF Technical Committee. He is also the chair of the IPMI CIM Mapping Committee of the IPMI Forum.

FOR MORE INFORMATION

SMASH specifications:
www.dmtf.org/standards/smash

Using PCI Express Technology

in High-Performance Computing Clusters

Peripheral Component Interconnect (PCI) Express is a scalable, standards-based, high-bandwidth I/O interconnect technology. Dell™ HPC clusters use PCI Express–based products including InfiniBand interconnects and Dell PowerEdge™ Expandable RAID Controller (PERC) storage. To demonstrate the benefits of PCI Express, a team of Dell engineers compared the performance of a Dell HPC cluster using PCI Express–based InfiniBand and PERC components with a cluster using previous-generation PCI Extended (PCI-X) technology.

BY RINKU GUPTA; SAEED IQBAL, PH.D.; AND ANDREW BACHLER

Related Categories:

Benchmarks

Clustering

Dell PowerEdge servers

High-performance computing (HPC)

InfiniBand

Interconnects

PCI Express

Performance

Visit www.dell.com/powersolutions for the complete category index.

High-performance computing (HPC) clusters comprising industry-standard servers, storage, and interconnect components are designed to provide high performance at a relatively low cost compared to traditional monolithic supercomputers. As technology advances, IT organizations have more options regarding the choice of the most suitable cluster components. For example, fast interconnects such as InfiniBand and Myricom Myrinet are now available for HPC clusters. Server technology is transitioning to dual-core processors, and storage technology is moving toward Serial ATA (SATA) and Serial Attached SCSI (SAS). SAS is designed to replace SCSI parallel cable with a much smaller from factor; it uses point-to-point connectivity that promises many advantages over its parallel predecessor (SCSI). Along with these developments, Peripheral Component Interconnect (PCI) Express technology has been created to provide higher I/O bus throughput in servers than previous-generation PCI technologies.

The use of PCI Express technology can affect the choice of other components within an HPC cluster. In particular, using PCI Express with a cluster interconnect such as InfiniBand and with storage-based RAID cards such as the Dell PowerEdge Expandable RAID Controller (PERC) can offer certain advantages.

The history of PCI, PCI-X, and PCI Express

In the 1980s, a typical PC used Instruction Set Architecture (ISA), which was a 16-bit wide I/O bus that operated at 4.77 MHz and had a bandwidth of 9 MB/sec. Subsequently, ISA was enhanced and new implementation technology was developed—for example, the Extended ISA (EISA) was 32 bits wide and operated at 8 MHz, and the VESA local bus (VL-bus) was 32 bits wide. A major issue that limited scalability in these I/O buses was the possibility of interference among communicating devices when more than two devices where connected simultaneously.

Developed by Intel in 1992, the PCI bus standard combined and improved upon the features of ISA and VL-bus. PCI introduced a bridge between the I/O devices and the CPU via the frontside bus (FSB), enabling higher scalability than had been possible using previous I/O technology. With these innovations, up to five devices could be connected to the PCI bus. The PCI bus was

32 bits wide, operated at 33 MHz, and had a bandwidth of 132 MB/sec. Since its introduction, PCI has been universally accepted as the standard I/O interconnect in servers. In addition, several design and technology improvements have increased the performance of PCI to 64 bits at 133 MHz in the PCI Extended (PCI-X) specification, which has a maximum bandwidth of 1 GB/sec.[1]

Because the upper limit of PCI-X bandwidth is 1 GB/sec, further improvement in performance is not economical. Several computer manufacturers have proposed new bus technologies in response to the numerous requirements of I/O devices in use today. In general, the trend is to move away from shared-bus technology toward point-to-point (direct connection) technology among communicating devices.

PCI Express architecture

Today, microprocessors and I/O devices—such as 10 Gigabit Ethernet, 10 Gbps Fibre Channel, InfiniBand, SATA, and SAS—demand more bandwidth than PCI-X can provide. PCI-X is an extension of PCI, and the basic architectures of PCI and PCI-X are very similar (parallel shared buses). In the PCI and PCI-X architectures, I/O devices are connected to the memory controller through an I/O bridge. The I/O bridge limits scalability because the bus is shared among all connected devices. In addition, only one I/O device can be connected in a point-to-point configuration to the PCI-X 2.0 bus.

The PCI Express architecture is very different from the PCI-X architecture. PCI Express is based on high-speed point-to-point technology, which uses serial interfaces to connect devices. The point-to-point architecture improves scalability by allowing multiple lanes of data at each PCI Express slot. The PCI Express architecture has a host bridge with several end points. Devices are connected to these end points, and traffic is routed through the host bridge. To add more devices, IT administrators can add a switch to the end points in the host bridge. Several devices can then be connected to the switch. Traffic can be routed through the switch from one device to another in a peer-to-peer configuration without going through the host bridge.

The PCI Express architecture is designed to improve performance substantially by directly connecting I/O devices to a memory controller via PCI Express links. Each PCI Express link can have multiple lanes, with each lane capable of 250 MB/sec of bidirectional bandwidth. Thus, an 8x PCI Express channel (8 lanes per link) can achieve 2 GB/sec in each direction. The basic PCI Express slot is 1x, which means one lane carries data. Different slots may be 2x, 4x, 8x, or 16x, depending on how many lanes are designed in them.

PCI Express has a layered architecture, which enhances scalability and eases backward compatibility. For example, PCI Express is compatible with PCI, and components based on previous versions of PCI can be used in PCI Express slots. Similarly, silicon technology can be improved without altering other layers, or a technology such as Fibre Channel can be used to implement the physical layer.

The PCI Express architecture consists of five layers. The lowest layer is the physical layer, which offers pairs of low-voltage differential, high-bandwidth dedicated serial channels. Each channel is dual-simplex—that is, it can transmit and receive signals simultaneously.[2] The link layer is above the physical layer and adds packet sequencing and error detection to help ensure data integrity during transmission. Above the link layer is the transaction layer, which receives read and write requests from the layers above it and creates packets according to these requests; the transaction layer supports 32-bit and 64-bit addressing. The transaction layer also offers several standard packet-switching features. On top of the transaction layer is the software layer, which is primarily used for the driver software. The software layer generates read and write requests to I/O devices. The topmost layer is the configuration/OS layer, which uses the PCI plug-and-play specification to initialize, enumerate, and configure any connected device. It does this with cooperation from the software layer below it; the result is robust device initialization and system configuration. System architects can choose from various mechanical form factors currently available for PCI Express connectors.[3]

PCI Express has a layered architecture, which enhances scalability and eases backward compatibility. For example, PCI Express is compatible with PCI, and components based on previous versions of PCI can be used in PCI Express slots.

PCI Express offers several advanced features, but the most important are advanced power management, support for real-time traffic, hot swapping and hot plugging, data integrity, and error handling. PCI Express can adjust power consumption when the bus is not in use to save power. Some multimedia devices require guaranteed processing in real time, which is supported under PCI Express by implementing virtual channels. PCI Express has native support for hot plugging and hot swapping of I/O devices, which can help minimize required server downtime.

[1] For more information about PCI Express, visit www.intel.com/technology/pciexpress/devnet.

[2] For more information, visit www.express-lane.org.

[3] For more information, visit www.PCI-sig.org.

PCI Express and cluster interconnects in Dell HPC clusters

Interconnects play an important role in clusters because they connect industry-standard computing components. To help communication-intensive applications achieve high performance on clusters, various high-speed interconnects have been developed over the last few years. While many of these interconnects are proprietary (such as Myrinet and Quadrics QsNet), there has been an initiative to develop a low-latency, high-bandwidth interconnect based on industry standards. InfiniBand is the result of such an initiative led by the InfiniBand Trade Association (IBTA),[4] a consortium of hardware and software vendors.

The InfiniBand architecture is a point-to-point, switched fabric architecture that connects various end points, where each end point can be a storage controller, a network interface card (NIC), or an interface to a host system. A host channel adapter (HCA) provides the host interface. This HCA has traditionally been connected to the host processor through a standard PCI or PCI-X bus. Because most Dell servers now support the PCI Express interface, this bus has become the primary choice for connecting HCAs to the host processor.

InfiniBand architecture defines different link speeds: 1X, 4X, and 12X. These links speeds are designed to yield data rates of 2.5 Gbps, 10 Gbps, and 30 Gbps, respectively. At the physical layer, InfiniBand uses 8B/10B encoding. In September 2004, the IBTA completed the InfiniBand 1.2 specification, which specifies double data rate (DDR) and quad data rate (QDR) modes of operation. These modes define increased signaling rates over existing 1X, 4X, and 12X InfiniBand links and are designed to effectively double or quadruple bandwidth. While current 4X InfiniBand HCAs have specified speeds of 10 Gbps, these speeds cannot be achieved via the current PCI-X bus. PCI-X buses, which have a maximum bidirectional throughput of 1 GB/sec, can act as a major bottleneck and limit the performance achievable by the InfiniBand 4X cards. The PCI Express bus, in contrast, is designed to eliminate this bottleneck, helping achieve the full bandwidth potential of InfiniBand cards. Dell partners with Topspin Communications, Inc., to incorporate InfiniBand hardware and software into Dell HPC clusters.

Dell HPC clusters can range from 8 to 256 nodes and include PCI Express–based components. Servers such as the Dell PowerEdge 1850 and PowerEdge 1855 support PCI Express. The PowerEdge 1855 can accommodate up to 10 server blades and includes an InfiniBand daughtercard; the PowerEdge 1850 uses a PCI Express slot–based HCA. For further information, refer to www.dell.com/hpcc.

InfiniBand performance analysis

In November 2004, a team of engineers from the Scalable Systems Group at Dell tested Dell HPC clusters to demonstrate the advantages of PCI Express–based InfiniBand HCAs as compared to PCI-X-based HCAs. The test environment comprised a cluster of 16 identically configured Dell PowerEdge 1850 servers running the Red Hat® Enterprise Linux® 3 OS, interconnected with InfiniBand HCAs and switches. Each PowerEdge 1850 server had two Intel® Xeon™ processors running at 3.2 GHz and 4 GB of RAM operating on an 800 MHz FSB.

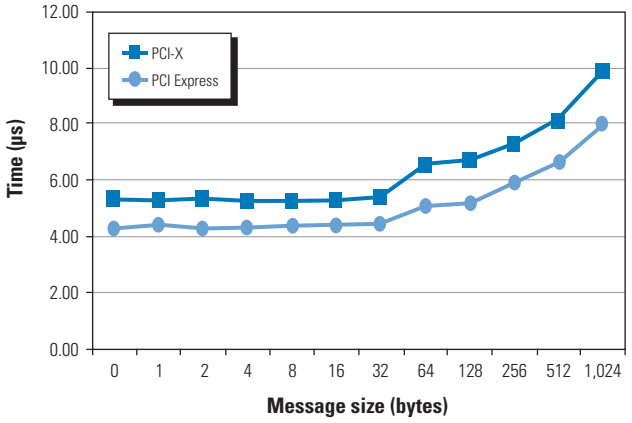


Figure 1. Pallas Ping-Pong latency test results for small message sizes

For the PCI-X systems, each server was equipped with a PCI-X riser consisting of two PCI-X slots operating at 100 MHz and 133 MHz. The InfiniBand card was inserted in the 133 MHz slot. For the PCI Express systems, each server used a PCI Express riser, which had 4x and 8x PCI Express slots. The InfiniBand card was inserted into the 8x slot. The PCI-X and PCI Express InfiniBand components used in the study were obtained from Dell partner Topspin Communications. Each PCI-X and PCI Express HCA had dual 4x ports.

The Dell test team used two popular benchmarks: the Pallas Message Passing Interface (MPI) Benchmarks (PMB) 2.2.1 and the NASA Advanced Supercomputing (NAS) Parallel Benchmarks (NPB) 2.4.

Testing performance with Pallas benchmarks

To determine the advantages of PCI Express–based InfiniBand over PCI-X-based InfiniBand, the test team used the Pallas Ping-Pong test, which provides point-to-point bandwidth and latency measurements between two nodes; and the Pallas Send-Receive test, which provides bidirectional bandwidth between two nodes. Figures 1 and 2 show the point-to-point link performance results from these tests. Figure 1 shows that the latency for small messages for the PCI Express–based InfiniBand was approximately 4.3 microseconds (µs) as compared to the 5.32 µs obtained for PCI-X-based InfiniBand. Figure 2 shows that the bandwidth for

[4] For more information about InfiniBand and the IBTA, visit www.infinibandta.org.

PCI-X-based InfiniBand peaked at approximately 700 MB/sec, while PCI Express–based InfiniBand achieved up to 915 MB/sec for large message sizes.

In the Pallas Send-Receive benchmark, the nodes in a group first send a message to the node on the right and receive a message from the node on the left. The total number of messages with respect to each node is two: one sending and one receiving. The benchmark is based on the MPI_SendRecv primitive implementation in MPI; the Pallas Send-Receive test between two nodes serves as a bandwidth test and consists of two nodes sending and receiving messages from each other simultaneously. As shown in Figure 3, for a PCI-X-based InfiniBand HCA, peak bidirectional bandwidth was limited to approximately 813 MB/sec; for PCI Express, this bandwidth scaled up to approximately 1,763 MB/sec. These test results indicate that PCI Express–based InfiniBand can leverage the additional bandwidth made available by PCI Express buses and can help provide a major performance boost to applications.

Testing performance with the NPB suite

The NPB suite consists of eight programs derived from computational fluid dynamics (CFD) code. These programs measure overall cluster performance, and results are measured in millions of operations per second (MOPS). Different classes of these programs represent different problem sizes. For the Dell tests, the team used Class C, which represents the largest problem size. Figure 4 shows the percentage improvement shown by PCI Express–based HCAs as compared to the PCI-X-based HCAs (which served as a baseline) for six of the programs in the NPB suite. The test team used six of the eight NPB programs: the kernel programs FT (Fast Fourier Transform), MG (Multigrid), CG (Conjugate Gradient), and IS (Integer Sort) emulate the computational core of different numerical methods used by CFD applications; BT (Block Tridiagonal) and SP (Scalar Pentadiagonal) are simulated CFD applications.

The IS benchmark from the NPB suite tests both integer computation speed and communication performance. It is a parallel program that is used in particle method codes. The IS benchmark involves no floating-point arithmetic but does have intense data communication. When run on PCI Express–based InfiniBand, the IS program achieved nearly a 12 percent performance improvement compared to PCI-X. As indicated in Figure 4, other benchmark programs show improvement depending upon the amount of communication that takes place within them. Applications with less communication can suffice with PCI-X-based InfiniBand cards or possibly a slower interconnect such as Gigabit Ethernet.[5]

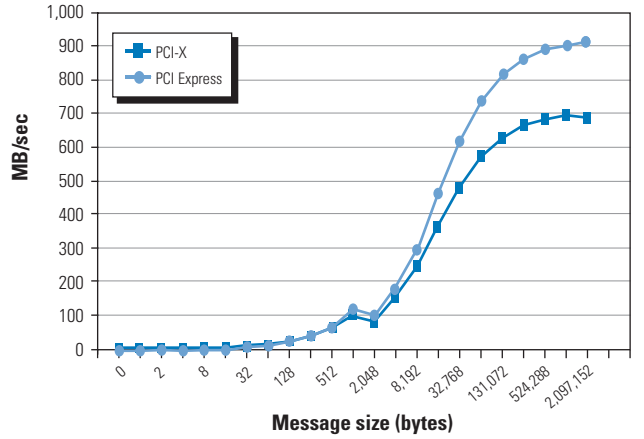


Figure 2. Pallas Ping-Pong bandwidth test results

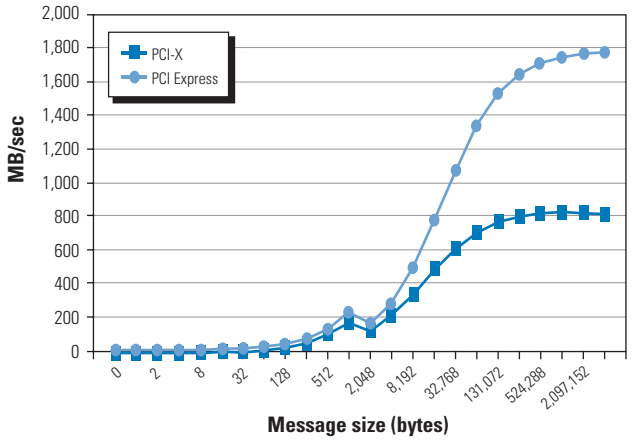


Figure 3. Pallas Send-Receive test results between two nodes

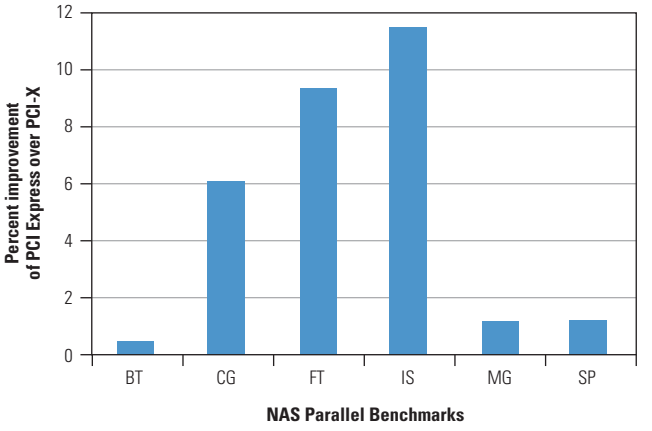


Figure 4. Relative percentage improvement in NAS parallel benchmarks of PCI Express versus baseline PCI-X-based InfiniBand HCAs

[5] This term does not connote an actual operating speed of 1 Gbps. For high-speed transmission, connection to a Gigabit Ethernet server and network infrastructure is required.

PCI Express and storage in Dell HPC clusters

HPC clusters are commonly built with low-cost SCSI RAID storage. In the Dell tests, the team also compared a PCI-X PERC 4, Dual Channel (PERC 4/DC) SCSI adapter with a new-generation PCI Express PERC 4, Extended Dual Channel (PERC 4 e/DC) SCSI adapter. These two RAID adapters were connected to an external Dell PowerVault™ 220S storage system fully populated with fourteen 73 GB 10,000 rpm drives in RAID-0 and RAID-5 configurations.

The PERC 4/DC and the PERC 4 e/DC each has a 128 MB read/write cache that supports up to 14 drives per channel with a maximum of 28 drives. Both adapters can also handle up to 40 logical drives in RAID-0, RAID-1, RAID-5, RAID-10, and RAID-50 configurations.

To evaluate the performance comparisons, the Dell test team used a public domain benchmark utility called IOzone, which is widely used in testing file system I/O performance for sequential and random read/write processes. The testing was performed on a Dell PowerEdge 2850 server configured with 8 GB of RAM and connected to a Dell PowerVault 220S storage system using a PERC 4/DC. The OS was Red Hat Enterprise Linux AS 4 using Intel Extended Memory 64 Technology (EM64T).

The first test compared the PCI-X PERC 4/DC with the PCI Express PERC 4 e/DC utilizing both external channels and all 14 drives in a RAID-0 configuration for a direct baseline comparison. Running IOzone on the server, the test team used a test file size of 12 GB—one and a half times the available memory, which saturated the 8 GB of total system memory to simulate the storage load of an HPC environment. The PCI Express PERC 4 e/DC achieved on average a performance increase of more than 150 percent for writes and more than 100 percent for reads when compared with the PCI-X PERC 4/DC (see Figure 5).

The ability to aggregate bandwidth through the use of wide ports and expanders will help provide the performance scalability required by next-generation enterprise servers and storage systems.

The second test used the same hardware, but instead of a RAID-0 configuration, a seven-drive RAID-5 configuration was implemented to help ensure redundancy in the event of a drive failure. Again, the team used a 12 GB test file, and the PCI Express system performed significantly better—write performance increased more than 200 percent and read performance increased more than 50 percent—when compared with the PCI-X PERC 4/DC adapter (see Figure 5).

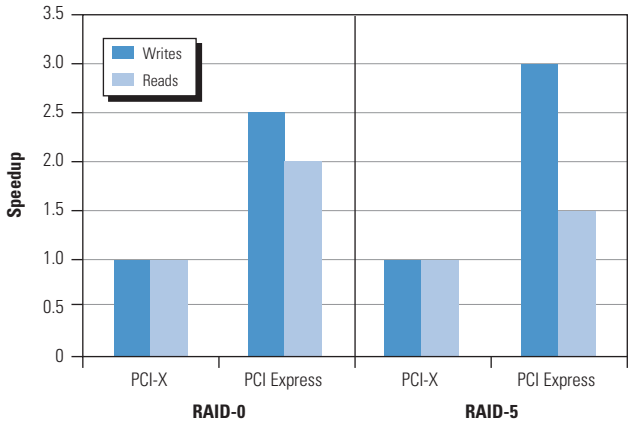


Figure 5. IOzone benchmark results comparing read/write performance for PCI-X and PCI Express in RAID-0 and RAID-5 configurations

The role of PCI Express in HPC clusters

PCI Express technology has helped achieve the improved performance that has been a feature of interconnects such as InfiniBand and has served as an impetus for the development of communication devices that can take advantage of this improved performance. On the InfiniBand side, the industry is moving toward memory-free PCI Express HCAs; on the storage side, SAS technology is the next-generation storage interface for SCSI. The ability to aggregate bandwidth through the use of wide ports and expanders will help provide the performance scalability required by next-generation enterprise servers and storage systems.

Rinku Gupta is a systems engineer and advisor in the Scalable Systems Group at Dell. Her current research interests are middleware libraries, parallel processing, performance, and interconnect benchmarking. Rinku has a B.E. in Computer Engineering from Mumbai University in India and an M.S. in Computer Information Science from The Ohio State University.

Saeed Iqbal, Ph.D., is a systems engineer and advisor in the Scalable Systems Group at Dell. His current work involves evaluation of resource managers and job schedulers used for standards-based clusters. He is also involved in performance analysis and system design of clusters. Saeed has a B.S. in Electrical Engineering and an M.S. in Computer Engineering from the University of Engineering and Technology in Lahore, Pakistan. He has a Ph.D. in Computer Engineering from The University of Texas at Austin.

Andrew Bachler is a systems engineer in the Scalable Systems Group at Dell. He has an associate's degree in Electronic Engineering and 12 years of experience with UNIX® and Linux platforms.

Platform Rocks: A Cluster Software Package for Dell HPC Platforms

Administrators can employ cluster solution packages such as Platform Rocks to help deploy, maintain, and manage high-performance computing (HPC) clusters. Based on the Linux® OS and NPACI Rocks, Platform Rocks includes drivers and other features that can provide comprehensive cluster management tools for Dell™ HPC platforms.

BY RIZWAN ALI, RINKU GUPTA, GARIMA KOCHHAR, AND BILL BRYCE

Related Categories:

Cluster management

Clustering

Dell PowerEdge servers

High-performance computing (HPC)

NPACI Rocks

Platform Computing

Systems management

Visit www.dell.com/powersolutions for the complete category index.

Twice a year, a group of manufacturers, computational scientists, experts in high-performance computing (HPC), and members of the Internet community compile and disseminate a list of sites that operate the 500 most powerful computer systems in the world. In the last five years, the percentage of clusters in the TOP500 Supercomputer Sites list has grown from 2.2 percent to 60.8 percent.[1] During the same time frame, the typical size of these clusters has grown from a few hundred processors to thousands of processors. Today, the largest Intel® processor–based cluster comprises more than 10,000 processors.[2]

As clusters expand, cluster deployment, maintenance, monitoring, and management can become complex and time-consuming processes. Before an application can tap the tremendous computing power available to an HPC cluster, administrators must install and configure the cluster for monitoring and management. Manually deploying even a small HPC cluster is no small task. The job demands that each cluster node and its OS be identically configured and installed, including proper drivers, networks, parallel libraries, and management software.

A cluster solution package is a semiautomated tool that helps administrators accomplish deployment, maintenance, and management tasks on an HPC cluster.[3] NPACI Rocks[4,5] is an open source, Linux-based software stack for building and maintaining HPC clusters. Platform Rocks[6] is an enterprise-level version of NPACI Rocks—developed by Platform Computing Inc.—that has been validated and verified on Dell PowerEdge™ servers by the Dell HPC Cluster team and Platform Computing.[7]

[1] For more information, see the TOP500 Supercomputer Sites Web site at www.top500.org. Results are available at www.top500.org/lists/2000/06 for the June 2000 list and at www.top500.org/lists/2005/06 for the June 2005 list.

[2] This cluster is NASA's Columbia Supercomputer; for more information, visit www.top500.org/sublist/System.php?id=7288.

[3] For more information, see "Felix, Scali, and Rocks: An Introduction to Cluster Computing Solutions" by Baris Guler; Rinku Gupta; Saeed Iqbal, Ph.D.; and Monica Kashyap in *Dell Power Solutions,* October 2004, www.dell.com/downloads/global/power/ps4q04-20040139-Gupta.pdf.

[4] For more information, see "Streamlining Beowulf Cluster Deployment with NPACI Rocks" by Rinku Gupta, Yung-Chin Fang, and Munira Hussain in *Dell Power Solutions,* February 2005, www.dell.com/downloads/global/power/ps1q05-20040176-Gupta.pdf.

[5] For more information about NPACI Rocks or to download Rocks, visit www.rocksclusters.org/Rocks.

[6] For more information about Platform Rocks or to download Platform Rocks, visit www.platform.com/products/Rocks.

[7] For more information, visit www.dell.com/hpcc.

Introduction to NPACI Rocks

The National Partnership for Advanced Computational Infrastructure (NPACI) designed the NPACI Rocks toolkit in November 2000 to help simplify the building and management of clusters. Rocks—a simple, self-contained, cluster-aware management system that is scalable and upgradeable—is slowly becoming the de facto cluster package. In October 2005, registered users of NPACI Rocks reported computational power that totals 149 trillion floating-point operations per second (TFLOPS).[8]

NPACI Rocks version 4.0.0 appeared in June 2005. Previous versions of Rocks contained two components—a base CD and Roll CDs. The base CD included minimal components for installation, including a recompiled Red Hat® Enterprise Linux OS.

Rolls are add-on components that provide specific capabilities to a cluster. Administrators can create Rolls to contain domain-specific software and applications; third-party vendors can customize Rolls to answer specific needs and requirements. Rather than using the base CD and assorted Rolls, the latest Rocks release employs a Roll-based framework. With Rocks 4.0.0, even core software packages for building the cluster are furnished in the form of a Roll—for example, Base Roll, Kernel Roll, HPC Roll, OS Roll, and so forth—and are considered essential for an HPC Rocks installation. An OS Roll separate from the core Rocks components allows organizations to create and use their preferred OS, such as Red Hat Enterprise Linux or Scientific Linux, to install their clusters. Apart from the OS Roll, Rocks 4.0.0 also allows administrators to use regular OS CDs to install the cluster.

Rocks—a simple, self-contained, cluster-aware management system that is scalable and upgradeable—is slowly becoming the de facto cluster package.

Platform Rocks functionality

Platform Rocks is designed to improve upon NPACI Rocks by providing additional functionality and enhanced support. The Dell HPC Cluster team and Platform Computing have tested Platform Rocks on Dell hardware,[9] and both Platform Computing and Dell provide support for enterprise customers.

As with NPACI Rocks, Platform Rocks has "base" components and optional modules packaged in "Rolls." Similarly, the base components of Platform Rocks *Standard* edition are designed as an open source software stack to automate and streamline cluster installation management. The *Standard* edition is designed for the Community Enterprise Operating System (CentOS)[10] and has no support option. However, it features the same set of tools, options, and configurations as the *Enterprise* edition.

The Platform Rocks *Enterprise* edition comes packaged with Red Hat Enterprise Linux AS on the front-end node and Red Hat Enterprise Linux WS on the compute nodes. It also provides a support option through Annual Cluster Care, which is a subscription service available from Platform Computing.

The *Enterprise* edition of Platform Rocks features a set of cluster tools, including the following:

- Message Passing Interface (MPI) libraries and drivers
- Cluster management tools and a workload manager
- Interconnect drivers and libraries to support interconnects such as Gigabit Ethernet, Topspin InfiniBand, and Myricom Myrinet
- Drivers for Dell hardware

Rocks management tools include Anaconda and kickstart for hosting images; Ganglia for monitoring; Cluster Top for sensing the precise activity of cluster nodes; and the 411 user management system.

Integrated with Platform Rocks as a Roll is an entry-level workload management tool called Platform Lava. This tool can be a critical component for operational management of the entire cluster, enabling job execution, management, and metrics for demanding enterprise environments in a simple, user-friendly package. Platform Lava is compatible with Platform Load Sharing Facility (LSF) HPC, providing organizations with a clear migration path from cluster workload management to enterprise workload management and enterprise grid deployments.

Platform Rocks 4.0.0 introduces several features that facilitate the administration and maintenance of a Platform Rocks cluster. Some enhancements include:

- Inclusion of Red Hat Enterprise Linux 4
- Ability to add and remove Rocks Rolls
- Integration with Red Hat Network (RHN)
- A database pre-population tool
- National Center for Supercomputing Applications (NCSA) Cluster Monitoring (Clumon) Roll[11]
- Platform Lava Web-based graphical user interface (GUI)

[8] For a registry of Rocks clusters, see www.rocksclusters.org/rocks-register.

[9] For specific Dell hardware, visit www.dell.com/hpcc.

[10] For more information about CentOS, visit www.centos.org.

[11] For more information about NCSA and Clumon, visit clumon.ncsa.uiuc.edu.

Integration with Red Hat Enterprise Linux

NPACI Rocks relies on CentOS to maintain compatibility with Red Hat Enterprise Linux. However, many enterprises require full Red Hat licenses and support. Platform Rocks contains Red Hat Enterprise Linux AS edition on the front-end node and Red Hat Enterprise Linux WS on the compute nodes. Adding Red Hat Enterprise Linux to Platform Rocks enables support for a complete stack: The hardware can be supported by Dell, the OS can be supported by Red Hat, and the Rocks framework can be supported by Platform Computing.

Ability to add and remove Rocks Rolls

Earlier versions of NPACI Rocks lacked a way to add or remove a Roll. Installing a new Roll on a cluster required rebuilding the cluster, including the front-end node, from scratch. This was consistent with the philosophy of NPACI Rocks that compute nodes are stateless. Although statelessness helps ensure compute-node consistency, cluster administrators may want to add Rolls to a cluster without having to reinstall the front end. Version 4.0.0 of NPACI Rocks introduced the capability to add Rolls to the cluster, but this version still does not allow Rolls to be removed—which can make ongoing cluster maintenance difficult.

The capability to add and remove Rolls requires an administrative command on the front end. A command called `rollops` in Platform Rocks 4.0.0 builds upon the kroll utility provided by NPACI Rocks. This command enables Rocks administrators to control Rolls on the front-end and compute nodes. Using `rollops`, an administrator can:

- Add Rolls to the cluster, without reinstalling the front-end node
- Remove Rolls, provided that the Rolls do not have kernel dependencies
- List Rolls currently installed on the front-end node
- Manage permissions for installations and uninstallations through a configuration file

The `rollops` command leverages the NPACI Rocks `kroll` command to install Rolls. It also provides a file called /opt/rocks/etc/rollopsrc, through which administrators can block access to key Rolls to prevent accidental overwriting.

Platform Lava Web-based GUI

Platform Rocks 4.0.0 includes a Web-based GUI for all Platform Lava users. The Web-based GUI is an alternative to the command-line interface and allows simple job submission, control, and monitoring for end users and administrators. This GUI can be used as a portal on the Platform Rocks front-end node and provides access to the cluster from JavaScript-enabled Web browsers.

Annual Cluster Care

The Annual Cluster Care program for Platform Rocks is available to subscribers at a per-node fee. It includes support, maintenance, upgrades, fixes, access to resources, and other services to help organizations manage their Platform Rocks clusters.

Platform Rocks and Red Hat Network

NPACI Rocks regularly updates the Rocks distribution with CentOS; Red Hat Enterprise Linux OS updates are provided by Red Hat through RHN. Platform Rocks 4.0.0 introduces the `rocks-update` command, which updates a Platform Rocks cluster using either the `up2date` (for the Platform Rocks *Enterprise* edition) or `yum` (for the Platform Rocks *Standard* edition) command. The `up2date` command obtains updated Red Hat Package Manager (RPM™) packages from RHN, and the `yum` command obtains updates from a Yum repository such as the Fedora repository.

The `rocks-update` command helps ensures that the Red Hat Enterprise Linux distribution installed by Platform Rocks remains up-to-date and has the latest security patches downloaded. It also maintains a list of RPM modules that should not be automatically updated because doing so could cause problems with the Platform Rocks framework. The `rocks-update` command connects to RHN and downloads the latest packages to /var/spool/rpms. However, this command does not automatically update the front-end and compute nodes. To do this, the administrator must run `% rocks-update --patch-compute` and `% rocks-update --patch-frontend`.

Downloading from RHN changes the build number of the installed version. For example, if the current version number is 4.0.0.0, the new version number advances to 4.0.0.1. In any distributed system such as Platform Rocks, the possibility always exists that some compute nodes will not update properly. The `rocks-update` command helps identify these compute nodes by generating a `rocks-update -list` option that displays the compute nodes and the current installed version. Sample output generated by entering `%rocks-update -list` is shown in Figure 1.

```
Current Repository Version: 4.0.0.2
Current Frontend Install Version: 4.0.0.1

Appliance Install Version

compute-0-0 4.0.0.1
compute-0-1 4.0.0.2
compute-0-2 4.0.0.1
compute-0-3 4.0.0.1
compute-0-4 4.0.0.1
compute-0-5 4.0.0.2-incomplete
compute-0-6 4.0.0.1
```

Figure 1. Sample output generated by rocks-update -list command

In the output shown in Figure 1, the current repository version represents the Rocks version available for installation on the front-end or compute nodes. The currently installed version on the front-end node is displayed as 4.0.0.1 and the versions for each compute node are listed. Some compute nodes have version 4.0.0.1 and some have 4.0.0.2. This means that not all compute nodes have been patched or reinstalled. One compute node bears an "incomplete" tag, meaning that a reimage of the compute node failed. The recommended recovery from this situation is to issue a `shoot-node` command, which reinstalls the compute node.

Rocks database pre-population tool

NPACI Rocks traditionally relies on Preboot Execution Environment (PXE) booting to add compute nodes to a cluster, but in NPACI Rocks 4.0.0, the `insert-ethers` command has additional options that can be used to update the database directly with a new node's IP and Media Access Control (MAC) address. Platform Rocks extends this concept with a Platform Rocks tool, `add-hosts`, which compiles an XML configuration file for all nodes in the cluster and pre-populates the Rocks database with the required information. This makes it easier for administrators to plan a Rocks deployment and determine in advance the IP addresses, subnets, node names, and rack locations for the new compute nodes. For example, if an administrator needs to install 256 nodes in two subnets, the XML configuration file /opt/rocks/etc/add-hostsrc shown in Figure 2 must be created.

Once the configuration file has been compiled and the administrator has run the `add-hosts` command, the new compute node values will enter the Platform Rocks database. This process does not install Rocks on the compute nodes; when each node is turned on, however, Rocks is designed to recognize the node by MAC address and install Rocks on the node automatically.

Optional Rolls

Optional Rolls extend the capability of Platform Rocks and provide administrators with the flexibility to choose desired modules to operate their cluster. The Platform Rocks Rolls should integrate with any Rocks implementation that does not modify the original

```
<?xml version="1.0" standalone="yes"?>
<add-hosts>
    <!-- the MAC addresses for hosts are contained in /opt/rocks/etc/mac-addr-list -->
    <mac_addr_file value="/opt/rocks/etc/mac-addr-list" />
    <!-- order_by_rack -->
    <order_by_rack value="no" />
    <!-- netmasks for all compute nodes will be 255.255.255.0 -->
    <netmask value="255.255.255.0" />
    <subnet>
        <!-- in the first subnet start naming the nodes with prefix node-, the base ip address is
        10.1.2.1, there are 128 nodes in this subnet and they are all compute appliances -->
        <host_prefix value="node-"/>
        <baseip value="10.1.2.1" />
        <num_hosts_in_subnet>128</num_hosts_in_subnet>
        <appliance>Compute</appliance>
    </subnet>
    <subnet>
        <!-- in the second subnet name the nodes with prefix 'node-', the base ip is 10.1.3.1,  there are
        128 nodes in this subnet and they are all compute appliances.-->
        <host_prefix value="node-"/>
        <baseip value="10.1.3.1" />
        <num_hosts_in_subnet>128</num_hosts_in_subnet>
            <appliance>Compute</appliance>
    </subnet>
</add-hosts >
```

Figure 2. XML configuration file for populating the Platform Rocks database

The New McDATA 4 Gb/s Switches

Four Degrees of Flexibility

Today's SAN is more diverse than ever before—and it is constantly changing. To meet your data needs and protect your investment, you need a very flexible fabric switch. But flexibility is a matter of degree.

Introducing the 4 Gb/s McDATA FlexPort 4400/4700 for Dell, the most versatile products on the market today for small to mid-sized companies or the enterprise edge. Your switch must deliver versatility in speed, affordability, interoperability and SAN design. With the 8-32 port Sphereon switches, you get the flexibility you need to meet the needs of your company's essential data today and as your SAN grows.

Speed: 1,2 or 4 Gb/s SANs

Affordability: Pay as you grow with FlexPort

Interoperability: Supports a range of multi-vendor products

SAN Design: FICON or open systems, stand-alone or data center edge

Visit *www.mcdata.com/4gig/dell* to find out more

McDATA®

NPACI source code. Rolls currently available from Platform Computing include:

- Clumon Roll
- Platform Lava Roll, which provides the freeware Platform Lava workload management tool
- Platform LSF HPC Roll, which provides the proprietary Platform LSF workload manager
- Intel Tools Roll, which provides compilers to optimize x86 and Intel Extended Memory 64 Technology (EM64T) environments[12]
- IBRIX Fusion File System Roll
- Topspin InfiniBand Roll

Figure 3 shows the Platform Rocks framework, including the different types of Rolls.

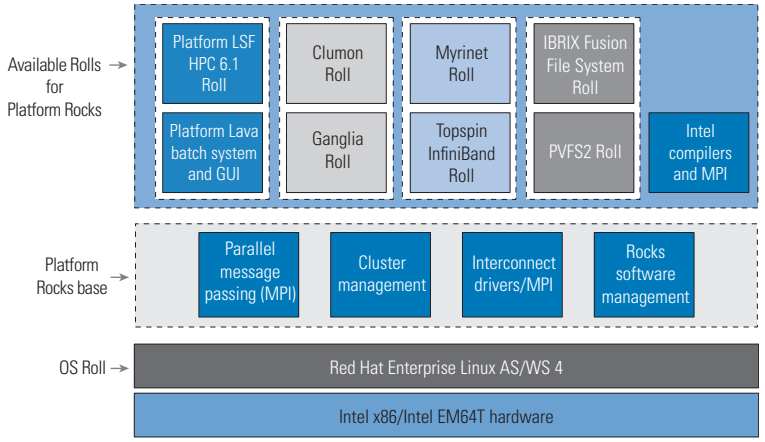


Figure 3. Platform Rocks framework and available Rolls

NCSA Clumon Roll

NPACI Rocks uses Ganglia to provide system information and metrics that help monitor the status and the load of compute nodes in the Rocks cluster. Ganglia permits easy viewing of summarized information for the entire cluster. The NCSA, located at the University of Illinois at Champaign-Urbana, developed the Clumon performance monitoring system for Linux-based clusters. Clumon displays cluster information for each node so concisely that the status and load of each node can be easily displayed on one screen, even for large clusters. This simple method for viewing a cluster provides an instant survey of the status of each node and lets administrators examine the circumstances when the state of a node visibly changes. Platform Computing is working with the developers of Clumon at NCSA to augment the functionality of Clumon and create a Clumon Roll, which is integrated with the Platform LSF and Platform Lava workload managers. Enhancements to Clumon are open source.

Comprehensive, efficient HPC cluster management

Dell and Platform Computing have collaborated to provide a turnkey HPC cluster solution that enables IT organizations to efficiently deploy and administer a cluster. Platform Rocks has been validated and is supported on 8- to 256-node Dell HPC cluster bundles for monolithic, rack-mountable servers and on 10- to 260-node Dell HPC cluster bundles for blade servers.[13] The comprehensive cluster management capabilities provided by Platform Rocks can help make HPC cluster environments an attractive architecture for the enterprise data center.

Rizwan Ali is a systems engineer with the Scalable Systems Group at Dell. His current research interests include performance benchmarking, cluster architecture, parallel applications, and high-speed interconnects. Rizwan has a B.S. in Electrical Engineering from the University of Minnesota.

Rinku Gupta is a systems engineer and advisor in the Scalable Systems Group at Dell. Her current research interests are middleware libraries, parallel processing, performance, and interconnect benchmarking. Rinku has a B.E. in Computer Engineering from Mumbai University in India and an M.S. in Computer Information Science from The Ohio State University.

Garima Kochhar is a systems engineer in the Scalable Systems Group at Dell. She has a B.S. in Computer Science and Physics from Birla Institute of Technology and Science (BITS) in Pilani, India, and an M.S. in Computer Science from The Ohio State University, where she worked on job scheduling.

Bill Bryce is the senior product manager for Platform Rocks at Platform Computing, where he has worked for the past 10 years. His interests include distributed computing, parallel programming, communications protocols, and operating systems. Bill has a B.S.C. in Computer Science from the University of Waterloo.

FOR MORE INFORMATION

Dell HPC clusters:
www.dell.com/hpcc

Platform Rocks:
www.platform.com/products/Rocks

[12] For more information about Intel software development products and cluster tools, visit www.intel.com/software/products/index.htm?iid=HPAGE+low_prod_software&.

[13] For more information about these bundles, visit www1.us.dell.com/content/topics/global.aspx/solutions/en/clustering_hpcc?c=us&cs=555&l=en&s=biz&~tab=4.

Workload Management and Job Scheduling on

Platform Rocks Clusters

Platform Lava, a free and fully functional entry-level workload manager for Platform Rocks, is becoming popular in the high-performance computing community. Platform Lava enables organizations to upgrade to advanced workload managers easily. By understanding the features and benefits of advanced workload managers, administrators can make informed decisions about the migration process—and help advance the dual goal of fair access to and enhanced utilization of shared computing resources.

BY SAEED IQBAL, PH.D.; YUNG-CHIN FANG; KEITH PRIDDY; AND BILL BRYCE

Related Categories:

Beowulf clusters

Cluster management

Clustering

High-performance computing (HPC)

Job scheduling

Platform Computing

Resource allocation

Workload management

Visit www.dell.com/powersolutions for the complete category index.

High-performance computing (HPC) clusters based on industry-standard computing, storage, and network interconnect components offer attractive price/performance ratios and ease incremental growth. HPC clusters are designed using cost-effective components; however, large clusters—especially those comprising hundreds of servers—require substantial investments. As a result, most HPC clusters are shared among several departments within an organization. Once a cluster is up and running, the challenge for system administrators is to ensure fair access and maximize utilization ratios in order to enhance return on investment. An efficient workload manager is critical for achieving these objectives. Workload managers vary in complexity and features. Some are feature rich but difficult to configure, while others are simple but buggy.

Figure 1 shows the typical components from which a Platform Rocks[1] cluster is built—including everything from the OS to benchmarking tools such as Linpack. A key enabling component of a Platform Rocks cluster is the free Platform Lava workload management software system. Platform Lava is based on Platform Load Sharing Facility (LSF) HPC, which has become the de facto workload manager for many production clusters. Platform Lava enables Platform Rocks users to submit their work easily to the cluster, while providing administrators with enough flexibility and control to help ensure that high-priority jobs complete on time.

[1] Platform Rocks includes software developed by the Rocks Cluster Group at the San Diego Supercomputer Center and its contributors.

A critical component of any workload manager is the job scheduler. When users submit jobs, the scheduler combines all requests and decides when and where to execute the jobs. Typically, workload managers have internal job schedulers. However, system administrators can substitute an external scheduler for the internal scheduler to enhance functionality. Platform Lava has a simple but stable first-come, first-served scheduler that is useful in most situations. Platform Lava's interface is compatible with Platform LSF HPC, providing a clear migration path. Platform LSF HPC can also use external schedulers such as Maui for enhanced features; for example, Maui provides advanced scheduling algorithms that are specifically designed to maximize utilization in policy-driven, heterogeneous HPC environments.

Features of Platform Lava

Platform Lava's features and capabilities can provide benefits for many cluster environments. Primary among these benefits are ease of installation and use, scalability, reliability, and flexibility and control.

Ease of installation and use. Platform Lava is easy to install on a Platform Rocks cluster. The Platform Lava Roll is installed first on the front-end node, and the necessary configuration files and services are copied, installed, and configured on each compute node as it joins the cluster. For single-user clusters, the default configuration is sufficient. Administrators who need to control access to the cluster or manage multiple groups with different applications on the cluster can easily edit the text-based configuration files using the Platform Lava simple command set. Once the changes are made, the administrator simply notifies the system to reload its configuration files.

Scalability. Platform Lava scales from 1 to 500 nodes with a single Platform Rocks front end. The workload manager also scales with the number of users, jobs queued, and jobs running in the system. Platform Lava is designed to queue up and run jobs from hundreds of users.

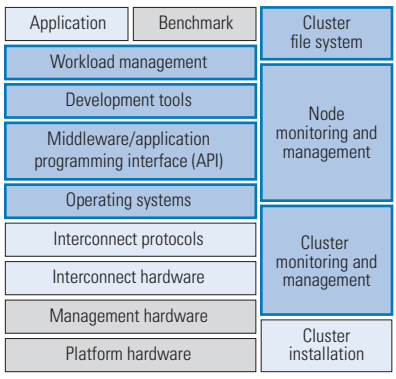


Figure 1. Components of a Platform Rocks cluster

Reliability. Platform Lava keeps track of work submitted and maintains accurate job records so that jobs will continue to run as long as the compute nodes running the application do not fail. If the front-end node fails or Platform Lava components fail, Platform Lava is designed to rebuild the state of the system and all jobs running on the system when those failed components are restarted. Even if an application completes or fails when Platform Lava is down, the software is designed to rebuild the state and notify the user that the job has finished or failed once Platform Lava is restarted.

Platform Lava keeps track of work submitted and maintains accurate job records so that jobs will continue to run as long as the compute nodes running the application do not fail.

Flexibility and control. For shared cluster resources, manual rules need to be established between users to enforce controlling and modifying the cluster for each user's needs and applications. In practice, this works when there is a small set of users or when all of the users know each other and effectively police each other. But as the number of users on the cluster grows, more elaborate control features are needed. This is where flexibility and control help the Platform Lava administrator. Platform Lava can be configured to change a cluster's configuration during different times of the day and to control access to the cluster by user, host, queue, and general Linux® limits. By modifying the text configuration files, the Platform Lava administrator controls work on the cluster.

Benefits of migrating to advanced workload managers

The advantages of migrating from Platform Lava to Platform LSF HPC include enhanced scalability, parallel job control, advanced scheduling features and algorithms, and the ability to link multiple clusters into a single compute resource. Many production environments exceed the capabilities for basic workload managers. Typically, when an environment grows, the complexity of user needs, the capacity and capability of computing resources, and even the ability to connect to other independent clusters become important factors in managing a compute infrastructure. Best practices recommend that complex compute infrastructures be handled with advanced workload managers.

In general, as a compute infrastructure grows in complexity, the emphasis is placed either on high throughput or on high performance. Another way to look at this division is to consider high-throughput computing as *capacity* computing and high-performance computing as *capability* computing. The challenge in capacity computing is maximizing the number of serial jobs

vmware®

VIRTUAL addition

The New Math: Less = More

Addition
Rapidly Add New Server Resources

Subtraction
Subtract Server Maintenance Downtime

Multiplication
Multiply Your Business Continuity

Division
Divide and Conquer

Virtual Addition = The New Math

Dell™ PowerEdge™ Servers
\+ VMware Virtual Machines
\+ Dell/EMC Storage

= unparalleled flexibility, fast time to productivity and economic benefits

Get the Power of VMware Virtualization

GET MORE OUT OF NOW

Get The Power of VMware Virtualization Today
Visit www.vmware.com/POV

that can be pushed through the compute infrastructure. Key concerns are job-priority management and job scheduling without compromising resource utilization. In contrast, high-performance computing maximizes fine-grained parallel processing. Among the many issues that arise in such a capability computing environment, key considerations include the management and monitoring of parallel subtasks and the optimization of scheduling to leave as few gaps in compute utilization as possible. Platform LSF HPC can be an excellent choice for industrial-strength capacity and capability computing.

Inevitably, as organizations change over time, so do their HPC requirements. Platform Lava offers support for organizational growth. However, when administrators must manage demanding workloads across a large cluster, Platform LSF HPC may be required. Because they share the same code base, Platform Lava easily integrates with Platform LSF HPC. The configuration files and functionality of Platform Lava are a subset of those of Platform LSF HPC. Dell and Platform Computing support a clear upgrade path from Platform Lava to Platform LSF HPC.

Advanced scheduling features in Platform LSF HPC can address the issues posed by HPC environments by employing advanced algorithms such as hierarchical fair share and multiple configurable policies. Effective configuration of scheduling policies can enhance resource utilization. Once business policies have been defined to identify critical projects, users, or groups, those policies can be included in the scheduler configuration. This approach enables the scheduler to make decisions dynamically regarding the priority of individual jobs—allowing for equitable and efficient division of resources based on the assigned policies.

Because they share the same code base, Platform Lava easily integrates with Platform LSF HPC. The configuration files and functionality of Platform Lava are a subset of those of Platform LSF HPC.

Platform LSF HPC helps support scalability into the range of 5,000 nodes, with 100,000 active and queued jobs in the cluster. Multicluster features built into Platform LSF HPC help extend scalability by integrating geographically dispersed clusters. These tools are designed to enable clusters not only to import and export compute jobs, but also to share resources between clusters automatically and transparently. In addition, jobs can be forwarded and resources leased to remote clusters.

The steps required to migrate from Platform Lava to Platform LSF HPC are basic. To install on Platform Rocks 3.3.0, administrators perform a "front-end upgrade" and select the Platform LSF HPC Roll instead of the Platform Lava Roll. On a Platform Rocks 4.0.0 cluster, the ability to add and remove Rolls is built into the infrastructure. Platform Lava can be removed as a Roll, and then Platform LSF HPC can be added. At this point, Platform LSF HPC can be configured to take advantage of the specialized features that are desired. To make use of the parallel management capabilities, administrators simply use prebuilt parallel job launchers that are part of the Platform LSF HPC distribution.

Upgrading does not require code changes in the applications running under the workload manager. Under Platform, multicluster jobs can be passed back and forth between clusters based on send and receive policies. An additional advantage of upgrading from Platform Lava is that Platform LSF HPC includes a GUI.

Key features of advanced workload managers

In production environments, job scheduling can be complex because multiple factors need to be considered. The quality of a schedule depends on the sophistication of the scheduling algorithms employed. Advanced scheduling algorithms use techniques such as advanced reservation and backfill.

Advanced reservation of resources is a technique used by Platform LSF HPC. Jobs that require a large portion of the available resources or have very long anticipated runtimes can use advanced reservation to have a predictable starting time. Advanced reservation of high-priority jobs helps deliver the required quality of service (QoS). Backfill is a common technique that can enhance the quality of the schedule generated. Given a schedule with advanced-reserved, high-priority jobs and low-priority jobs, a backfill algorithm tries to fit the small jobs into scheduling gaps. This allocation does not alter the sequence of jobs previously scheduled but helps improve system utilization by running low-priority jobs between high-priority jobs.

Platform LSF HPC provides a dynamic scheduling decision mechanism, which uses the processing load to make scheduling decisions. Based on these decisions, jobs can be migrated among compute nodes or rescheduled. Loads can also be balanced among compute nodes in heterogeneous environments. In addition, Platform LSF HPC can dynamically migrate jobs among compute nodes. Furthermore, Platform LSF HPC can apply multiple scheduling algorithms to different queues simultaneously.

Platform LSF HPC is targeted toward parallel HPC applications. It offers parallel job control and HPC-specific job scheduling features. In HPC clusters, the scheduling of parallel jobs requires special attention because parallel jobs comprise several subtasks. Each subtask is assigned to a unique compute node during execution, and nodes constantly communicate among themselves during execution. The manner in which the subtasks are assigned to processors is called mapping. Because mapping

affects execution time, the scheduler must map subtasks carefully. To achieve high resource utilization for parallel jobs, both job efficiency and advanced scheduling are required. Platform LSF HPC can make intelligent scheduling decisions based on the features of advanced interconnect networks, thus enhancing process mapping for parallel applications.

Platform LSF HPC can make intelligent scheduling decisions based on the features of advanced interconnect networks, thus enhancing process mapping for parallel applications.

Platform LSF HPC can interface with external schedulers like Maui, which may have complementary features. For example, Maui can further enhance the utility of compute resources by considering jobs submitted from all queues concurrently. Internally, Maui is based on a single unified queue. This scheduler offers a variety of QoS settings to system administrators and access levels to users and jobs. Maui can stop a job during execution—a task called preemption—under several conditions, including submission of a higher-priority job.

Workload management for production HPC environments

Platform Lava is a full-featured workload manager that is available at no charge—providing an excellent opportunity for organizations deploying HPC clusters to evaluate it. Dell™ HPC clusters offer a supported and dependable upgrade path to industry-leading workload managers. The principal benefits of upgrading include enhanced scalability; advanced scheduling that can lead to optimal resource utilization; tight job control; multilevel fault tolerance; and topology-aware scheduling. Furthermore, the upgrade does not require any code changes in applications.

Saeed Iqbal, Ph.D., is a systems engineer and advisor in the Scalable Systems Group at Dell. His current work involves evaluation of resource managers and job schedulers used for standards-based clusters. He is also involved in performance analysis and system design of clusters. Saeed has a B.S. in Electrical Engineering and an M.S. in Computer Engineering from the University of Engineering and Technology in Lahore, Pakistan. He has a Ph.D. in Computer Engineering from The University of Texas at Austin.

Yung-Chin Fang is a senior consultant in the Scalable Systems Group at Dell. He specializes in HPC systems, advanced HPC architecture, and cyberinfrastructure management. Yung-Chin has published more than 30 conference papers and articles on these topics. He also participates in HPC cluster–related open source groups as a Dell representative.

Keith Priddy is an integration architect at Platform Computing. His responsibilities include integrating clustering products with the Platform resource management framework. Keith has a B.S. in Computer Science from Texas A&M University.

Bill Bryce is the senior product manager for Platform Rocks at Platform Computing, where he has worked for the past 10 years. His interests include distributed computing, parallel programming, communications protocols, and operating systems. Bill has a B.S.C. in Computer Science from the University of Waterloo.

FOR MORE INFORMATION

Dell HPC clusters:
www.dell.com/hpcc

Platform LSF HPC:
www.platform.com

Maui Cluster Scheduler:
www.clusterresources.com

Dell OpenManage Tools

for High-Performance Computing Cluster Management

High-performance computing (HPC) clusters use industry-standard computing, storage, and interconnect components to aggregate cost-effective supercomputing power. As the number of nodes in a typical HPC cluster continues to escalate, efficient remote cluster management is becoming a necessity. Components of the Dell™ OpenManage™ software suite can be used to enhance HPC cluster management.

BY YUNG-CHIN FANG; ARUN RAJAN; MONICA KASHYAP; SAEED IQBAL, PH.D.; AND TONG LIU

Related Categories:

Cluster management

Clustering

Dell OpenManage

Dell PowerEdge servers

High-performance computing (HPC)

Systems management

Visit www.dell.com/powersolutions for the complete category index.

The Dell OpenManage software suite enables administrators to monitor and manage Dell PowerEdge™ servers remotely, centralizing IT resources and streamlining systems management tasks to help reduce total cost of ownership. Particular tools within the Dell OpenManage suite are designed to provide management and monitoring capabilities for PowerEdge servers composing a high-performance computing (HPC) cluster.

Components of the Dell OpenManage suite

Dell OpenManage software consists of four CDs: the Installation and Server Management CD, the Systems Management Consoles CD, the Service and Diagnostics Utilities CD, and the Documentation CD. Dell also provides Web-downloadable Dell OpenManage update packages.[1]

Dell OpenManage Installation and Server Management CD

The Dell OpenManage Installation and Server Management CD contains three tools: Dell OpenManage Server Administrator (OMSA), Dell OpenManage Array Manager, and Dell OpenManage Server Assistant.

Dell OpenManage Server Administrator. OMSA is a secured management agent that provides flexible command-line, Web, and Simple Network Management Protocol (SNMP) interfaces; detailed fault and performance information is reported in the selected user interface. OMSA is designed to review and report the hardware, firmware, and management software configuration and status of a server and a host-based RAID subsystem. The online diagnostic tool included with OMSA can be used to determine the root cause of hardware problems while the system is still operational. OMSA also supports OS-level remote power management, such as graceful shutdown, as well as remote BIOS and system firmware updates. The OMSA storage service is designed to configure, monitor, and manage locally attached RAID and non-RAID disk storage.

Dell OpenManage Array Manager. This tool is designed to configure, monitor, and manage storage devices attached to the server, including controllers, array disks, enclosures, channels, and other physical components. Array Manager can obtain information about a storage system's logical components, such as virtual disks and volumes, and displays the connections among the logical

[1] For Dell OpenManage downloads, visit support.dell.com/support/downloads/devices.aspx?c=us&cs=555&l=en&s=biz&SystemID=PWE_1850&category=36&os=LE30&osl=en.

and physical storage components. In addition, Array Manager can create and delete virtual disks on the storage system and rebuild and reconstruct data on a system's locally attached RAID storage.

Dell OpenManage Server Assistant. This step-by-step system configuration utility supports the installation process for Microsoft® Windows®, Red Hat® Enterprise Linux®, and Novell® NetWare® operating systems as well as for the latest RAID controllers, network devices, and other optimized device drivers for Dell servers. In one node recovery scenario, Server Assistant can be used to recover and reinstall the OS on a failed node.

Dell OpenManage Systems Management Consoles CD

The Dell OpenManage Systems Management Consoles CD provides four console tools: Dell OpenManage IT Assistant (ITA), the remote access controller (RAC) console, the Array Manager console, and the baseboard management controller (BMC)[2] management utility (BMU).

Dell OpenManage IT Assistant. ITA is a centralized, one-to-many remote hardware monitoring and management console. Through a Web-based graphical user interface (GUI), ITA provides cluster hardware fault monitoring with notification mechanisms such as e-mail, paging, and console alerting to inform administrators of events. For example, events may include exceptional disk, memory, voltage, fan, and thermal conditions. Inventory and asset reporting are also supported. In addition, ITA allows utilities—such as Array Manager, OMSA, remote access services, the Dell PowerConnect™ switch manager, and remote console software for Dell digital KVM (keyboard, video, mouse) switches—to be launched in context.

RAC console. A RAC supports virtual media, which can be used for diskless HPC cluster configuration. The RAC console is designed to access RACs via a Web interface or command-line interface (CLI). An OS-independent remote console that supports graphics mode, such as the RAC console, is often used in HPC clusters. This console does not require OS-level console redirection settings. A dedicated RAC network interface card (NIC) is designed to help reduce the performance impact on communication-sensitive applications because RAC management traffic is routed through a dedicated fabric. The RAC family of hardware includes the Dell Remote Access Card III (DRAC III), DRAC III/XT, Dell Remote Access Controller 4 (DRAC 4), Embedded Remote Access (ERA), ERA Option (ERA/O), and ERA/Modular Chassis (ERA/MC).

Array Manager console. Array Manager is a storage configuration and management tool that allows cluster administrators to configure and manage local and remote storage attached to a system.[3] The Array Manager console communicates with the

[2] For more information, see "Remote Management with the Baseboard Management Controller in Eighth-Generation Dell PowerEdge Servers" by Haihong Zhuo; Jianwen Yin, Ph.D.; and Anil V. Rao in *Dell Power Solutions,* October 2004, www.dell.com/downloads/global/power/ps4q04-20040110-Zhuo.pdf.

[3] For more information, see the *Dell OpenManage Array Manager User's Guide* at support.dell.com/support/edocs/software/smarrman.

server management framework to provide storage management functions, including physical and logical views of storage; context-sensitive menus; dialog boxes; wizards; and property pages that display alerts, events, and so forth.

BMC management utility. The BMU is designed for Intelligent Platform Management Interface (IPMI)–compliant servers.[4] The BMU provides out-of-band management and BMC configuration features. Features provided in the BMU include the IPMI shell (ipmish), a CLI to correspond with the remote BMC, and a Remote Management and Control Protocol (RMCP)–based proxy that allows BIOS- and OS-level console redirection through the LAN—which is also known as Serial Over LAN (SOL). In the deployment phase, the BMU (via ipmish) is typically used to power up a new server remotely. In the operational phase, the BMU (via ipmish) can remotely power cycle a hung node, and it can be used (via SOL) as an out-of-band remote console.

Dell OpenManage Service and Diagnostics Utilities CD

The Dell OpenManage Service and Diagnostics Utilities CD provides utilities, updates, and diagnostic tools from Dell.

Utilities and updates. This CD provides the latest BIOS, firmware, services, and OS-based online diagnostics for supported systems. *Note:* Updates using the CD are supported only on systems running Microsoft Windows. To use the CD on Novell NetWare or Red Hat Enterprise Linux platforms, administrators must first use a Microsoft Windows–based system to extract the required drivers from the CD; then they can share those drivers with the systems running other operating systems.

Diagnostics. Hardware and firmware diagnostic utilities can be used locally and remotely on Dell PowerEdge servers. The diagnostic utilities are supported on certain Microsoft Windows NT® and Windows 2000 versions. Administrators can select the tests to be run—simultaneously or individually—on various components of a server. However, these diagnostics utilities are limited to addressing problems on individual servers and will not resolve or identify problems that arise at the network level. Examples of components that can be diagnosed include hard disk drives, various media drives, Peripheral Component Interconnect (PCI) buses, SCSI devices, serial and parallel ports, and NICs. Easy-to-use GUIs enable administrators to access functions such as the Test Queue Viewer (lists the currently selected tests, queued to be run sequentially); the Progress Viewer (shows the test's progress while it is running); the Diagnostic tree (lists the available tests based on the component); and the Component Selector (lists the diagnosable components).

In addition to the tools provided on the Service and Diagnostic Utilities CD, the latest BIOS, firmware, drivers, and Dell OpenManage applications can be obtained from the Dell support Web site at support.dell.com.

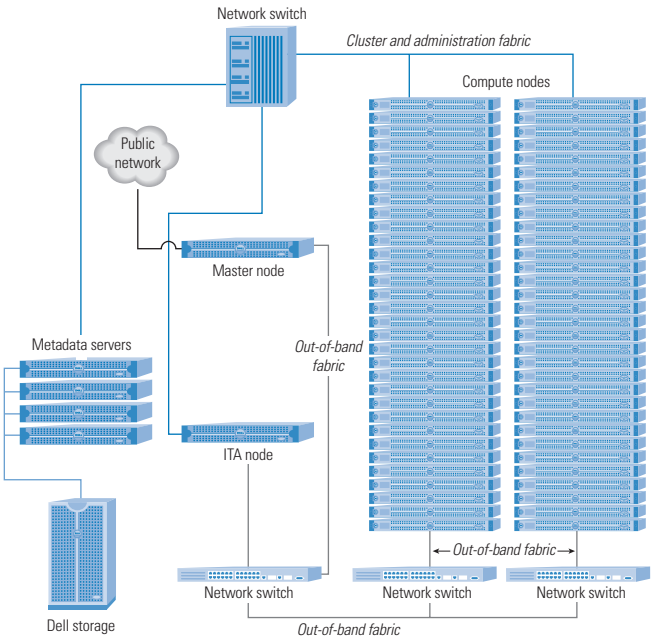


Figure 1. Architecture of a typical HPC cluster

Dell OpenManage Documentation CD

The Dell OpenManage Documentation CD provides manuals for Dell hardware and software.

Hardware manuals. The Documentation CD contains user's guides, installation guides, and troubleshooting guides for Dell PowerEdge systems. It also provides RAID controller–related documentation, such as user guides and driver installation guides, as well as adapter card and modem documentation.

Operating manuals for Dell OpenManage components. This CD also includes the *Software Quick Installation Guide* and guides for various Dell OpenManage components such as Array Manager, DRAC, ITA, the BMC, and the Server Update Utility (SUU).

Web-downloadable components

Various Dell OpenManage components are also available on the Web.

Server Update Utility. The SUU is a CD-based application that can be used to identify and apply the latest updates (BIOS, firmware, and drivers) to a Dell PowerEdge server. The application provides a comparison report differentiating component versions. It also allows administrators to update components using preconfigured System Update Sets. The SUU uses a database of firmware, drivers, and BIOS components for Dell PowerEdge servers called

[4] For more information, see "Efficient BMC Configuration on Dell PowerEdge Servers Using the Dell Deployment Toolkit" by Anusha Ragunathan, Alan Brumley, and Ruoting Huang in *Dell Power Solutions,* February 2005, www.dell.com/downloads/global/power/ps1q05-20040219-Brumley.pdf.

the repository. Administrators needs root (Linux) or administrative (Windows) permissions to apply updates using the SUU.

Dell Update Packages. These packages can be an easy-to-use, flexible method for updating system software on Dell PowerEdge servers. Dell Update Packages provide a CLI that can be used to perform multiple updates as a batch process and a GUI that can be used to apply individual updates to the system.

For the latest information about the SUU and Dell Update Packages, visit the Dell support Web site.

Building blocks of the HPC cluster architecture

The architectural figure shown in Figure 1 represents a typical HPC cluster. The diagram depicts the master node, compute nodes, external network that connects to the public network, private cluster and administration fabric, out-of-band management fabric, and attached storage. It also shows the dedicated, centralized one-to-many management console: the ITA node. The metadata servers and attached storage are designed for an optional parallel file system.

Master node. The master node typically runs a Linux OS and stores compute node configuration data in a database. Administrators can log on to the master node via the public network, develop applications on the master node, submit jobs from the master node, and use performance-monitoring tools to observe cluster performance. When an event occurs, the master node can either launch an OS-level remedy to a remote node or reboot a hung node via the BMU (using ipmish). In many configurations, the Network File System (NFS) running on the master node also serves as a shared file system for compute nodes. The NFS can reside on local storage or on an attached storage chassis for high capacity.

Compute nodes. Compute nodes typically run a Linux OS and interface with the master node through the internal cluster and administration fabric, which is also used for installation and OS-level administration. Compute nodes perform the number-crunching tasks. In a Dell HPC cluster, each node has an embedded BMC, which connects to the NIC for out-of-band management; OMSA runs on the OS level of the node for in-band management. When cluster administrators notice that a compute node is not reporting to ITA, they can remotely diagnose the issue via the out-of-band fabric to help bring the node back online.

ITA node. In a Dell HPC cluster, one node runs IT Assistant to monitor and manage the cluster hardware. The ITA node also can interface through the out-of-band fabric to the master node, compute nodes, and any metadata servers—allowing for remote monitoring and management of the HPC cluster.

Management fabrics. This includes both in-band and out-of-band fabrics. In-band fabric requires the presence of an OS, while out-of-band fabric can function in an OS-absent state. These fabrics include various mechanisms that use available resources to manage

and monitor cluster components. With IPMI 1.5, the host OS and out-of-band traffic share the same NIC (reducing the environment to one fabric) to ease cable management. In an IPMI-compliant system, the serial port also can participate in management to pass IPMI commands to the BMC via a serial port concentrator. The optional ERA card provides remote manageability on a dedicated out-of-band management fabric via a network switch—for example, using virtual media and a software-independent remote graphics console. The supported RAC CLI provides scripting ability for cluster administrators to tailor management features to meet specific needs.

Digital/analog KVM switch. This switch helps significantly reduce cable volume, secure digital access, and enable flexible remote server management. The Dell PowerEdge 2161DS Console Switch[5] consists of a rack-mountable KVM switch with 16 analog Avocent Rack Interface (ARI) ports, which can connect to devices over a standard LAN. Remote console switch software must be installed to allow administrators to view and control servers attached to the switch and help prevent unauthorized access via IP. The switch uses Ethernet networking infrastructure and TCP/IP to transmit KVM information. A KVM switch is typically used for the master node and the ITA node to share the monitor, keyboard, and mouse.

HPC cluster deployment using Dell OpenManage

The pre-OS deployment phase involves setting up the BMC, BIOS, and other components for all nodes. Administrators can use the Dell OpenManage Deployment Toolkit (DTK) to set up the BMC and BIOS remotely via the Preboot Execution Environment (PXE). The DTK is a set of utilities—which can be downloaded from the Dell support Web site—designed to help configure Dell PowerEdge servers in a pre-OS state. Using `biosconfig` and `bmcconfig` commands from the CLI, administrators can configure the BIOS (including the boot order) and the BMC (including IP, RAC, and storage settings), respectively. The Dell PowerEdge SC line of systems uses the Configuration Toolkit (CTK), which provides similar utilities for BMC and BIOS setup on those servers. For more detailed information, refer to the Dell OpenManage Documentation CD.

If a node is equipped with a RAC device, that RAC device also can be used to set up the BMC and BIOS remotely. For example, administrators can wrap the DTK and script into a bootable image, place the bootable image into virtual media, and boot from the virtual media. Using the dedicated RAC out of band requires that the RAC be set up accordingly. This setup may be automated by programming the out-of-band network switch to assign IP addresses one-by-one to the RACs by masking all but one port of the switch at each step.

Note: Deploying large-scale clusters in an unattended fashion requires extra caution because of potential problems such as power surges. Best practices recommend performing segmented parallel remote power-up for a cluster, incorporating a delay between each node power-up to avoid an aggregated power surge that may damage power facilities.

Installing OMSA

To install OMSA on all the compute nodes, administrators can create a server model–specific OMSA slice (refer to the *Installation and Security User's Guide* on the Documentation CD) that can be wrapped into the cluster OS image for the compute nodes. The other method for remotely deploying OMSA is to first create the server model–specific OMSA slice on NFS, and then use a parallel shell and the slice on NFS to remotely install OMSA in silent and unattended mode on all compute nodes. After the OS installation, the nodes can be rebooted and post-configuration tasks can be performed, including system diagnostics for staging and stabilizing the cluster hardware. OMSA also can be used to change the boot order from PXE to booting from the hard drive first.[6]

Installing the OS and software stack

Remote deployment of the OS and cluster computing software stack is often a difficult task when moving from a newly configured cluster to a fully deployed HPC cluster. An efficient and highly automated way to conduct this deployment is to perform a PXE boot from an image server.[7] This can be achieved by having a Bootstrap Protocol (BOOTP)/Dynamic Host Configuration Protocol (DHCP) server and an NFS server on the network. The kickstart file and OS image should be housed on the NFS server, while the networking information, boot kernel, RAM disk, and kickstart file are placed on the BOOTP/DHCP server. This approach enables unmanaged PXE-based OS installation across a newly deployed, large-scale HPC cluster.

Efficient remote cluster management

The Dell OpenManage suite comprises various tools, which are available on four CDs as well as on the Dell support Web site. These tools are designed to ease the deployment and operational phases of large-scale HPC clusters. Dell OpenManage tools can be used to

[5] For more information about the Dell 2161DS Console Switch, visit support.dell.com/support/edocs/systems/smarcon/en/2161DS/hardware/hardware.pdf.

[6] For information about using Platform Rocks to automate BMC and BIOS configuration during deployment, see "Configuring the BMC and BIOS on Dell Platforms in HPC Cluster Environments" by Garima Kochhar, Rizwan Ali, and Arun Rajan in *Dell Power Solutions,* November 2005, www.dell.com/downloads/global/power/ps4q05-20050222-Kochhar.pdf.

[7] For more information see "Installing Linux High-Performance Computing Clusters" by Christopher Stanton, Rizwan Ali, Yung-Chin Fang, and Munira A. Hussain in *Dell Power Solutions,* Issue 4, 2001, ftp.us.dell.com/app/4q01-Lin.pdf.

configure, monitor, and manage the various components within an HPC cluster environment, including the master node, compute nodes, management fabrics, switches, and other devices. With these tools, IT organizations can streamline the process of scaling out HPC clusters to support growing data centers.

Yung-Chin Fang is a senior consultant in the Scalable Systems Group at Dell. He specializes in HPC systems, advanced HPC architecture, and cyberinfrastructure management. Yung-Chin has published more than 30 conference papers and articles on these topics. He also participates in HPC cluster–related open source groups as a Dell representative.

Arun Rajan is a systems engineer in the Scalable Systems Group at Dell. He has a B.S. in Electronics and Communications Engineering from the National Institute of Technology, Tiruchirappalli, in India and an M.S. in Computer and Information Science from The Ohio State University.

Monica Kashyap is a senior systems engineer in the Scalable Systems Group at Dell. Her current interests and responsibilities include in-band and out-of-band cluster management, cluster computing packages, and product development. She has a B.S. in Applied Science and Computer Engineering from the University of North Carolina at Chapel Hill.

Saeed Iqbal, Ph.D., is a systems engineer and advisor in the Scalable Systems Group at Dell. His current work involves evaluation of resource managers and job schedulers used for standards-based clusters. He is also involved in performance analysis and system design of clusters. Saeed has a B.S. in Electrical Engineering and an M.S. in Computer Engineering from the University of Engineering and Technology in Lahore, Pakistan. He has a Ph.D. in Computer Engineering from The University of Texas at Austin.

Tong Liu is a systems engineer in the Scalable Systems Group at Dell. His current research interests are HPC cluster management, high-availability HPC clusters, and parallel file systems. Tong serves as a program committee member of several conferences and working groups on cluster computing. Before joining Dell, he was an architect and lead developer of High Availability Open Source Cluster Application Resources (HA-OSCAR). Tong has an M.S. in Computer Science from Louisiana Tech University.

FOR MORE INFORMATION

Dell OpenManage:
www.dell.com/openmanage

Dell HPC clusters:
www.dell.com/hpcc

The Dell Life-Cycle Approach to

Implementing HPC Cluster Services

Dell™ high-performance computing (HPC) clusters are designed to help solve complex business problems and perform resource-intensive calculations and data analysis. Dell PowerEdge™ servers can be essential building blocks for HPC clusters because they provide an excellent price/performance ratio. Dell Services, a consulting organization within Dell, helps organizations plan, deploy, and manage HPC cluster solutions, supporting every phase of the cluster life cycle.

BY SIMON JANDRESKI, J. LANCE MILLER, AND JIM SKIRVIN

Related Categories:

Clustering

Dell PowerEdge servers

High-performance computing (HPC)

Professional services

Services

Visit www.dell.com/powersolutions for the complete category index.

Built using industry-standard components for computing, storage, and network interconnects, Dell high-performance computing (HPC) clusters can be designed to address particular computational requirements or to solve specific, complex data-analysis problems. HPC clusters enable organizations to replace expensive proprietary supercomputers and specialty computers designed for parallel computing without sacrificing computational power or research time. Commonly used in various scientific and commercial fields, HPC clusters can provide cost-effective computational power for high-end, floating-point-intensive scientific and engineering problems and data-intensive commercial tasks. Common uses of HPC clusters include:

- Weather modeling
- Seismic analysis for oil exploration
- Computational fluid dynamics for aerodynamic simulation of automobiles and aircraft
- Bioinformatics and protein folding for molecular modeling in biomedical research
- Data mining and finance modeling for business analysis

Dell Services has implemented Dell platform–based HPC cluster projects for configurations ranging from 4 to as many as 4,096 nodes. In June 2005, 20 Dell platform–based clusters were named in the TOP500 Supercomputers Sites list[1]—a semiannual list of the 500 most powerful computer systems in the world. Dell manages HPC cluster projects using the life-cycle approach shown in Figure 1, which provides an overview of key Dell services: planning (assess and design), implementing (build and deploy), and maintaining (support and optimize).

[1] The TOP500 Supercomputers Sites list for June 2005 is available at www.top500.org/lists/2005/06.

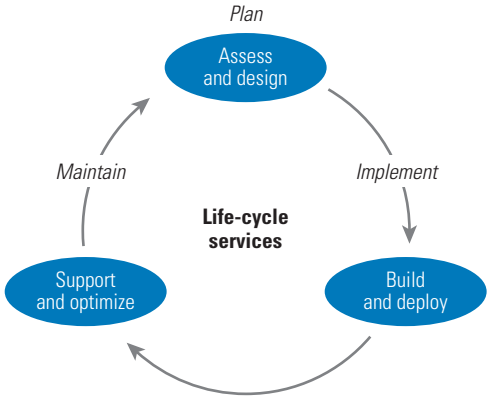


Figure 1. The Dell life-cycle approach to HPC cluster services

Assess and design the environment

The planning phase is the optimal entry point for Dell Services involvement in an HPC cluster implementation. Many organizations prefer to perform the planning step themselves and are fully capable of doing so. However, Dell's operational excellence and execution experience in planning a wide range of HPC cluster solutions can benefit organizations by helping to avoid costly mistakes and helping to ensure that the maximum processing power is obtained from the available budget.

Dell's operational excellence and execution experience in planning a wide range of HPC cluster solutions can benefit organizations by helping to avoid costly mistakes and helping to ensure that the maximum processing power is obtained from the available budget.

Dell Services solution architects begin by assessing an organization's applications and processing requirements, and then they assist in decision making to help select components that are most appropriate for the intended processing use. Using details uncovered during the assessment phase, solution architects can design the HPC cluster following best-practices methodologies that employ Dell's expertise in building enterprise solutions and leverage Dell's alliances with industry leaders to enhance performance and end-to-end interoperability. Solution architects work closely with both external and internal organizations to help select the necessary building blocks appropriate for their required use, including consideration of processor and interconnect needs, storage and storage area network (SAN) requirements, and systems management and job-scheduling necessities. Figure 2 shows a sample HPC cluster architecture.

To help validate the proposed solution, Dell Services can employ the support of the Dell Scalable Enterprise Computing (SEC) team. The SEC team weighs enterprise requirements and the type of processing the HPC cluster must perform against the proposed equipment stack and then makes a feasibility determination. This process enables Dell to validate that all components necessary for a successful deployment have been identified, thereby minimizing the possibility that critical objects have been overlooked and helping to avoid costly delays—all before the order is placed.

For example, Dell Services provided this type of oversight when a major aerospace laboratory designed a 512-node cluster. Hardware, software, peripherals, floor layout, network trays, and electrical concerns were all addressed during the planning stage of the engagement.

Build and deploy the cluster

The potential chaos incurred during the deployment of a large cluster can degrade and sometimes overwhelm the day-to-day operations of an IT facility. In a typical deployment, trucks arrive at a loading dock and drop off more than a thousand individually packed servers, racks, and interconnect components. Inherent in this approach is the need to divert valuable IT resources just to manage delivery logistics and dispose of the packaging.

The Dell Services proposition is simple: provide high quality and value in services while enhancing customer experience. To help reduce the chaos, the clutter, and the on-site setup time, which affects both service quality and value, the Austin Merge Center (AMC)—which is located near Dell's server manufacturing facilities—serves as a logistical assembly point for hardware

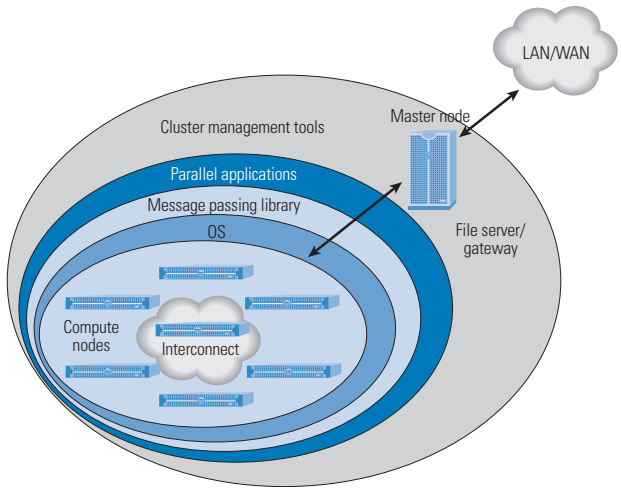


Figure 2. A sample HPC cluster architecture

components, including components that may require some degree of customization and varying build times from other manufacturers.

By leveraging the AMC facility as a staging and assembly point for HPC cluster engagements, Dell Services can manage much of a large-scale deployment away from the customer's site. Individual components of an HPC cluster solution can be received, unboxed, racked, cabled, and tested before they are shipped to the customer for deployment. At the final destination, deployment then consists of rolling up to the loading dock with a rack of components that are already configured and ready to be moved to their final location in the data center. Once all the prebuilt racks are in place, the preconfigured and labeled cable bundles can be used to connect the racks in a fraction of the time it would normally take to build the racks on-site.

Dell assigns a single point of contact for the cluster implementation from conception to support. Regardless of whether the base components are Dell branded, Dell will support the recommended HPC cluster configuration.

Recently, this technique was employed successfully at a large university in the western United States. Components of the university's HPC cluster were racked, cabled, and shipped directly from the AMC. The university received racks completely assembled and populated, virtually eliminating the need to manually install components or dispose of boxes and packing materials from the university's data center. Before the components were delivered, Dell Services conducted an on-site Data Center Environment Assessment (www.dell.com/dcea) to help ensure that the university's facilities had the power and cooling capacity for the cluster before it was installed. Key to the success of this implementation was the Dell Services project management team. Assigned to oversee actual installation at the university, this team helped ensure that deployment went smoothly, that issues were resolved expeditiously, that the university received regular status reports, and that the project stayed on schedule.

Support and optimize the data center

Dell assigns a single point of contact for the cluster implementation from conception to support. Regardless of whether the base components are Dell branded, Dell will support the recommended HPC cluster configuration. In the case of Dell-branded components, both the OS and hardware are supported through Gold and Platinum support contracts.[2] Additionally, other components may be covered under both Gold and Platinum support options.

Following installation and deployment of an HPC cluster, organizations may contact Dell Services and request that the HPC cluster be further tuned to help extract optimal processing power and speed from their systems. On many occasions, Dell has helped optimize OS and networking options to help ensure peak performance. For example, Dell Services worked with a major technical university in the southeastern United States after its lead researcher noticed performance issues with the code running on the newly installed 256-node HPC cluster. Through persistence, expertise, and the support of Intel developers, the Dell team was able to assist the university in achieving the requisite performance level.

Achieve high performance and meet IT needs

By leveraging in-depth knowledge of Dell hardware and interconnect components as well as Dell's relationships and alliances with key partners, Dell Services can help organizations assess and design, build and deploy, and support and optimize HPC clusters—and through operational excellence and reliable execution, deliver high quality and value in services. The Dell team of solution architects, delivery consultants, project managers, and other resources has efficiently implemented numerous HPC cluster projects, helping to identify the proper cluster components to achieve the requisite price/performance for the intended use. The Dell Services approach is designed to provide best-practices services throughout the planning, implementation, and maintenance phases of the HPC cluster life cycle, assigning a single point of accountability for each HPC cluster project. The result is designed to be a tested and validated cluster environment that can meet the specific computational requirements or solve the complex data-analysis problems that a particular organization needs to address.

Simon Jandreski is a network engineer who works for Dell Services, where he helps standardize Dell service offerings and delivery methodology.

J. Lance Miller is a program manager responsible for the development of service offerings for Dell Services. His current focus is on HPC clusters, data center facilities, and Oracle® implementations.

Jim Skirvin is a solutions architect with Dell Services, where he serves as the services offering lead and a subject-matter expert for HPC clusters.

[2] For more information about Dell support options, visit www1.us.dell.com/content/topics/global.aspx/services/en/pesstiers?c=us&cs=555&l=en&s=biz.

Building Fast, Scalable I/O Infrastructures for High-Performance Computing Clusters

High-performance computing clusters require extremely scalable disk-based storage infrastructures that can support large-scale shared file systems to sustain high throughput and data integrity even under severe failure conditions.

BY BRAD WINETT

Related Categories:

Cluster management

Clustering

DataDirect Networks

High-performance computing (HPC)

Interconnects

Performance

Storage

Storage architecture

Visit www.dell.com/powersolutions for the complete category index.

By the late 1990s, clusters had become an accepted paradigm for high-performance computing (HPC) applications. At the same time, shared file systems that could scale in performance and capacity, also known as parallel or clustered file systems, began to mature and high-throughput, low-latency interconnects emerged—providing two of the three critical architectural components for a storage I/O hierarchy that could support the cutting-edge performance requirements of clustered computing. Recognizing the need for a storage system specifically designed to enable shared file systems for HPC clusters, DataDirect Networks developed the Silicon Storage Appliance™ (S2A) system to provide the third critical element for HPC cluster data access.

DataDirect Networks and Dell have collaborated on several successful cluster deployments. These sites use Dell-supplied DataDirect S2A storage controllers and the IBRIX Fusion file system from IBRIX, Inc., or the Lustre file system from Cluster File Systems, Inc., to create production-quality I/O environments that are designed to meet the performance requirements of highly demanding sites and to remain available during storage system, file server, or disk failures.

Enhancing shared file systems for HPC clusters

Shared file systems for HPC clusters require high sustained throughput to individual volumes; any disruption in storage subsystem performance or access can affect the entire cluster. Additionally, many scientific simulation applications require intense collaboration among compute nodes, and the loss of a compute node can terminate an entire job. To reduce the risk of having to restart an entire job, cluster administrators can configure checkpoints that take a full snapshot of the involved compute nodes at frequent points in time. These checkpoints are essentially full memory dumps (writes) to the shared-file-system storage to facilitate the restoring and restarting of a job from a "known good" point in time. Thus, the shared-file-system storage must have exceptional write performance to expedite the checkpoint process and maximize computing efficiency.

DataDirect S2A storage controllers are specifically designed to address the shared-file-system needs of HPC cluster environments: high throughput for reads and writes; enhanced scalability; sustained performance during disk drive failures; transparent, real-time I/O node failover; and management and statistical analysis tools.

High throughput for reads and writes. The DataDirect S2A9500 supports high-throughput performance by enabling superfast reads and writes. In addition, the S2A system is designed to reduce the complexity of cluster file system configurations through its PowerLUN feature, which enables high-performance volume creation without file-system-level striping. As a result, the S2A system can ease storage deployment, optimization, and system maintenance—helping to simplify the overall HPC cluster environment.

Enhanced scalability. The variety of HPC cluster environments can stretch the limits of any storage system, from simulation architectures with high-performance, low-capacity needs to imaging applications with immense storage requirements and moderate performance needs.

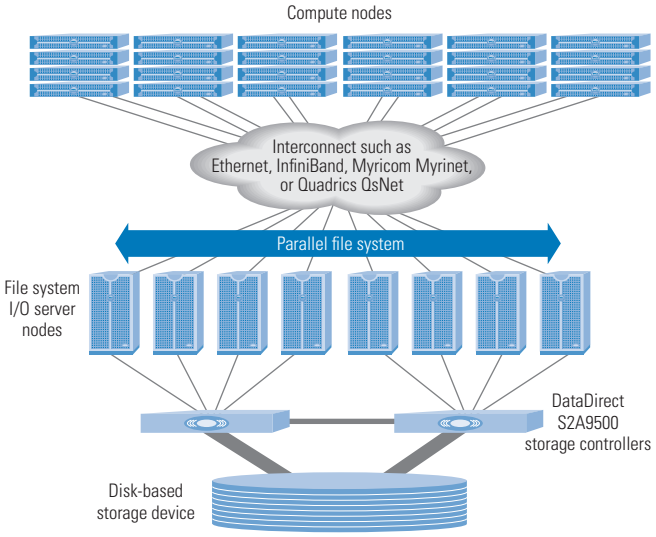


Figure 1. Typical HPC cluster scalable storage unit using DataDirect S2A storage controllers

A DataDirect S2A system's ability to scale from a few terabytes to more than 400 TB of capacity in a high-density form factor can help protect IT investments by growing flexibly as business requirements evolve.

Sustained performance during disk drive failures. Large shared file systems logically stripe across many disk drives to create single, high-performance volumes. The failure of a disk drive should not cause an entire file system to experience performance loss. The DataDirect S2A architecture provides a quality of service (QoS) capability that is designed to ensure that the performance of large file systems will not be affected by individual drive failures. The DataDirect S2A9500 system's dual-parity capability is designed to prevent performance impact or loss of data access even when two drive failures occur in the same RAID group.

Transparent, real-time I/O node failover. Typical storage systems use slow, intrusive daemons or scripting to migrate storage targets, known as logical units (LUNs), when a server node fails. In large clusters, this disruption in I/O service can cause performance problems or even job failures that can affect the computing activities of the entire cluster. The DataDirect S2A architecture is designed to provide system resiliency through its unique host-port parallelism to each I/O server node, which inherently provides simultaneous access to LUNs, enabling fast failover and trivial failback capability. This parallelism helps streamline administration and enhance the performance of the entire cluster.

Management and statistical analysis tools. Given the complexity and scale of a typical HPC cluster, the management capabilities of the underlying storage environment can be critical to maintaining peak cluster efficiency. The ability to apply statistical methods to optimize the data flow and examine I/O patterns can provide invaluable feedback for sustaining and improving the performance of the overall cluster. The DataDirect S2A storage controllers are designed to provide comprehensive reporting of system operating status; disk and host command efficiencies; and I/O pattern recognition that can be used to optimize the storage infrastructure, the shared file system, and even the compute applications themselves.

Creating robust building blocks for cluster architectures

The I/O storage system design for an HPC cluster depends on the scale of the compute architecture itself. Typical clusters require approximately 1 GB/sec of bandwidth and roughly 10 TB of storage capacity per trillion floating-point operations per second (TFLOPS) of compute power. One approach to cluster design is to divide clusters into "scalable units," which enables a manageable and repeatable building-block approach to cluster architecture. Figure 1 shows a typical HPC cluster network scenario with a scalable storage unit designed to deliver more than 2 GB/sec of sustained read or write bandwidth to the compute nodes; a 10 TFLOPS cluster could use five of these scalable units, for example. In this scenario, a parallel file system—enabled by the underlying block-level storage architecture of the DataDirect S2A system—could segment or stripe across the scalable units. In this way, a large file system could be created for the overall cluster to use.

A real-world example of such a cluster deployment took place at a university in the southeastern United States, where Dell, DataDirect Networks, and IBRIX configured an entry-level cluster consisting of approximately 400 compute nodes and a DataDirect S2A3000 storage controller designed to provide approximately 20 TB of Serial ATA (SATA) disk storage capacity and designed to enable a sustained bandwidth of roughly 500 MB/sec. In another real-world deployment, Dell helped a major supercomputing center build a 1,300-node, Intel® Xeon™ processor–based cluster designed to provide approximately 15 TFLOPS of compute power. This cluster had 150 TB of Fibre Channel disk storage that was designed to enable a sustained bandwidth of approximately 12 GB/sec, and it used DataDirect S2A8500 storage controllers to power a Lustre-based parallel file system. In these example deployments, Dell and DataDirect Networks hardware were configured to enable resilient HPC cluster storage at extremely attractive price/performance levels.

Optimizing HPC cluster efficiency and performance

Deploying storage systems that are expressly designed to enable the shared file systems required for HPC clusters can help streamline cluster administration and enhance the performance and efficiency of a cluster's I/O architecture. The read and write performance, scalability, quality of service, fault tolerance, and analytical tools of a storage system are important factors to consider when designing an HPC cluster. HPC clusters require highly resilient storage systems such as DataDirect Silicon Storage Appliance systems, which are designed to run optimally even under severe failure conditions.

Brad Winett is vice president of business development and marketing at DataDirect Networks (www.datadirectnet.com).

Reap the Benefits of SQL Server 2005

Originally published by *SQL Server Magazine* as part of its "SQL Server 2005 Upgrade Handbook," this article explores key SQL Server 2005 features, including enterprise data management, application development, and business intelligence.

BY DOUGLAS McDOWELL

Related Categories:

Application development

Business continuity

Business intelligence

Clustering

Database

High availability (HA)

Microsoft SQL Server

Microsoft SQL Server 2005

Microsoft Windows

Online analytical processing (OLAP)

Security

Visit www.dell.com/powersolutions for the complete category index.

If your organization depends on SQL Server today, prepare for Microsoft® SQL Server™ 2005—an upgrade that is designed to deliver immediate results for existing applications and continue to deliver results as you enhance existing applications and develop new applications to fully exploit the expanded capabilities. Today's business environment demands a comprehensive data-management platform that delivers business results with efficiency. SQL Server 2005 provides a comprehensive data-management platform (see Figure 1), integrating the development, management, and operations of relational data as well as extraction, transformation, and loading (ETL); online analytical processing (OLAP); and reporting with security, performance, and availability designed to meet the needs of the most demanding enterprise applications.

You can intelligently share data across databases, devices, and applications from multiple vendors by adopting SQL Server 2005's standardized data platform. This strategy can deliver value by letting you make IT investments based on open standards and widely available developer and administration skill sets and tools.

SQL Server 2005 also helps you control costs without sacrificing enterprise-level performance, reliability, or scalability. Through a comprehensive enterprise data-management feature set, commodity hardware support, and highly productive integrated development and management environments for developers and database administrators (DBAs), SQL Server 2005 is designed to enable low upfront implementation and ongoing maintenance costs. The family of SQL Server 2005 editions—Express, Workgroup, Standard, and Enterprise—includes necessary components in one product, without incremental fees for comprehensive, out-of-the-box data integration, management, analysis, and reporting functionality. Let's look at how the SQL Server 2005 release delivers value in the key areas of enterprise data management, developer productivity, and business intelligence.

Enhanced enterprise data management

SQL Server 2005 is ready for the enterprise, offering exceptional data availability and manageability, hardened security, and the ability to scale from handheld mobile devices to the most demanding online transaction processing (OLTP) systems and multi-terabyte data warehouses. This release enhances DBA productivity by automating routine tasks and letting administrators focus on high-value activities. SQL Server 2005 also provides an extensible framework for proactive health and performance monitoring. Here's a look at SQL Server 2005's enterprise data-management features.

Enterprise performance. SQL Server 2005 boasts impressive component-level performance, from the relational engine to business intelligence. The relational engine handles demanding OLTP workloads and multi-terabyte data warehouses. SQL Server 2005 Integration Services (SSIS) is designed to move millions of rows per second while performing in-memory transformations with delivery to multiple destinations. Analysis Services offers unified access to information, and is designed to provide sub-second query times, advanced caching, and data mining

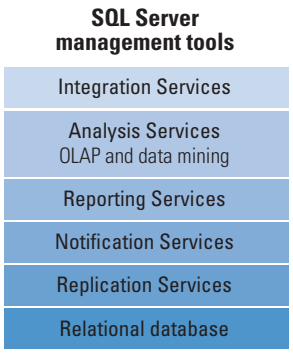


Figure 1. SQL Server 2005 tools enabling integrated data management for Microsoft Windows Server 2003, Microsoft Office, Microsoft Office SharePoint Portal Server 2003, Microsoft Visual Studio.net, and third-party applications

with predictions for very large data sets. Notification Services enables support for hundreds of thousands of subscription-based users with numerous business rules for event polling. And Reporting Services has deployment models to scale up or scale out, coupled with advanced caching and snapshot strategies to support high user concurrency. Bulk Load has provided a fast way to insert data into SQL Server, and in SQL Server 2005, you'll find ultrafast performance for bulk-loading in and out of the SQL Server process using Bulk Copy Program (bcp.exe) and BULK INSERT (Transact-SQL [T-SQL]). SQL Server 2005 also features enhanced performance and additional options for the OPENROWSET() function, which you can also use as a scalable method for loading XML documents.

Business continuance. For mission-critical applications, SQL Server 2005 enables 24/7 data availability, providing end users with consistent access to information. Failover clustering and database mirroring technologies let you deliver highly reliable, available applications to your employees, customers, and partners with minimal downtime (see Figure 2). Features such as online indexing, piecemeal backup and restore, partitioning, dynamic configuration, and support for hot memory swapping minimize and help eliminate downtime, enabling uninterrupted user access to enterprise data—even during disaster recovery operations. SSIS offers distributed deployment and restartability, promoting resilient (ETL) processes. The integration of Notification Services and Reporting Services with Microsoft Internet Information Services (IIS) enables network load balancing (NLB) to help maintain uptime. And SQL Server 2005 provides affordable, enhanced disaster recovery options, with peer-to-peer replication, database mirroring, log shipping, and the Analysis Services Server-Sync functionality allowing multiple servers to support primary servers. And to help ensure direct access to SQL Server for vital database recovery operations, SQL Server 2005 adds a dedicated administrator connection.

Manageability. SQL Server 2005 gives you a single, unified management tool—SQL Server Management Studio—that lets you manage the SQL Server platform from one interface (see Figure 3). This integration boosts DBA productivity across the SQL Server implementations in your enterprise. The release also lets you easily identify, troubleshoot, and resolve performance problems. SQL Server Profiler benefits from enhanced trace capabilities that encompass core products in SQL Server, including SQL Server Database Services, Analysis Services, and Integration Services. With SQL Server's extensible XML-based definition, you can capture and effectively analyze more details, use aggregated views, and perform correlations with Windows Event Logs and a recently expanded set of performance counters. Microsoft has added graphical deadlock and Showplan enhancements along with comprehensive access to crucial metadata through catalog views for database objects and Dynamic Management Views (DMVs) for ongoing server activity, dynamically changing state, and diagnostic information. You can also automate all repetitive or common SQL Server administrative tasks—such as programmatically retrieving configuration settings, creating new databases, applying T-SQL scripts, creating SQL Server Agent jobs, and scheduling backups—by using SQL Management Objects (SMO) and the Profiler API (application programming interface). SMO also enables independent software vendors and partners to build on top of the management framework and provides enhanced scalability and performance compared to SQL Distributed Management Objects (SQL-DMO). SQL Server 2005 continues to support applications written in SQL-DMO with DMO9.You can also use SQL Server Agent to build a proactive performance monitoring solution and speed problem isolation and identification. And the Database Tuning Advisor (DTA), which replaces the Index Tuning Wizard (ITW), can help you resolve performance issues. Furthermore, additional security contexts let you grant users who do not have system administrator access the ability to create database traces and use DTA for database tuning.

Security. Microsoft set a standard for security with its Trustworthy Computing initiative, which is designed to ensure a safe and reliable computing experience. SQL Server 2005 helps deliver on this from the design of the product through its final deployment: secure by design, secure by default, and secure in deployment. By default, SQL Server 2005 helps maximize security with a minimal surface area. A dedicated security-configuration interface called the Surface Area Configuration (SAC) consolidates access to configurable services and settings and gives you brief configuration descriptions to help you make informed decisions. SQL Server 2005 introduces fine-grain administration rights, which let you grant levels of development and administrative rights decoupled from the levels of data access rights in each platform subcomponent. SQL Server 2005 also gives you enhanced control over grantable permissions, password policy enforcement, and high levels of data encryption for storage and transmission. A secure development environment also lets developers easily sign, verify, and manage code—including Common Language Runtime (CLR) assemblies that run in the database—and managed code uses Code Access Security (CAS) to prevent assemblies from performing certain operations, protecting the OS or database server from compromise.

Developer productivity

SQL Server 2005's many development features and a comprehensive enterprise tool set empowers developers—whether on small or large

project teams—to rapidly deliver robust enterprise database applications. Here's how SQL Server 2005 helps your development team reduce time to market and collaborate to produce enhanced solutions.

Time to market. Deep integration between Visual Studio 2005 and SQL Server 2005 fosters rapid development and testing cycles. The developer interface enhances efficiency with wizards, accelerators, and step-by-step documentation letting you develop often complex processes in a fraction of the time required using previous versions of Visual Studio and SQL Server. With SQL Server 2005, developers can use one development framework for relational, XML, and OLAP applications integrated with Visual Studio 2005 for fast, efficient development and debugging. Developers can take advantage of the architectural switch from batch-level recompilations to statement-level recompilations, which requires less coding to prevent undesired recompilations. The introduction of large value types—varchar(max), varbinary(max), and nvarchar(max)—helps reduce the complexity of database programming, decreasing the special requirements for blob data types. SQL Server 2005 also adds Query Notifications, which enable an application to request a notification from SQL Server when the results of a query change. This functionality lets developers design efficient applications that aren't unnecessarily polling the underlying database for changes and using caching or disconnected record-sets when the database hasn't been updated. You can also gain enhanced development efficiencies by using data-access-layer features, including the support for Multiple Active Result Sets (MARS), which allow applications to have more than one active default result set per connection.

Team collaboration. Larger groups of SQL Server developers can now work together interactively on complex or small projects by using the mature Visual Studio 2005 deployment model. SQL Server 2005's integration with Visual Studio Team Services and other source-control platforms promotes developer consistency and accountability and enables enhanced configuration-management processes. Support for mature configuration-management practices and procedures can help reduce rework and miscommunication.

Interoperability. Through support for Web services and the .NET Framework, SQL Server 2005 supports interoperability with multiple platforms, applications, and programming languages. Support for existing and emerging open standards such as HTTP, XML, Simple Object Access Protocol (SOAP), XQuery, and XML Schema Definition (XSD) facilitates communications across your extended enterprise systems. Native XML support in SQL Server 2005 runs deep: You will find XML storage in a dedicated XML data type that has its own index type, strong query capabilities via XQuery and XPath, and the ability to create XML code from relational data by using the XSD language. FOR XML PATH, a relational query output format, allows nested FOR XML queries, which is designed to simplify queries for which you might currently be using the FOR XML EXPLICT option. Microsoft has even added an XML output format as an alternative to the conventional Showplan for query-plan evaluation.

Developers can build secure, reliable, and scalable applications using the SQL Server 2005 Service Broker technology. Service Broker provides queuing and reliable messaging between SQL Server instances, supporting scalable applications that benefit from a robust asynchronous programming model. Your applications can also embed enhanced reporting functionality when you use the Visual Studio 2005 Report Controls for Reporting Services. These controls let you deploy reports whether or not you have a Reporting Services report server available for report processing and rendering. With native support for the .NET Framework CLR, SQL Server 2005 and Visual Studio 2005 converge to let developers write stored procedures, triggers, user-defined functions (UDFs), user-defined types (UDTs), and user-defined aggregates in managed code. And because managed code compiles to native code before execution, you can gain significant performance enhancements in some scenarios.

Availability feature	Database mirroring	Failover clustering
Automatic failover	Yes	Yes
Transparent client redirection	Yes, auto-redirect	Yes, reconnect to same IP address
Impact on overall throughput	No	No
Protection against work loss	Yes	Yes
Certified hardware required	No	Yes
Redundant data provided	Yes	No

Figure 2. SQL Server 2005 database mirroring and failover clustering features

Advanced business intelligence

SQL Server 2005 provides components and tools to help implement a comprehensive, end-to-end business intelligence solution—whatever your analysis and reporting needs are. And implementing the release's entire integrated business intelligence feature set can magnify the benefits enabled by a SQL Server 2005 upgrade. How can the enhanced business intelligence features help you deliver solutions that give your users and business decision makers access to the information they need?

Single version of the truth. SQL Server 2005 is designed to provide a holistic view of your business for reporting and analysis, integrating reporting with OLAP. This paradigm, called the Unified Dimensional Model (UDM), handles complex data modeling scenarios, combining benefits of relational and traditional OLAP to enable a balance between data latency and query performance. UDM also helps you with localization issues and other unique reporting and analysis circumstances that historically require multiple products. The Business Intelligence Development Studio, integrated with

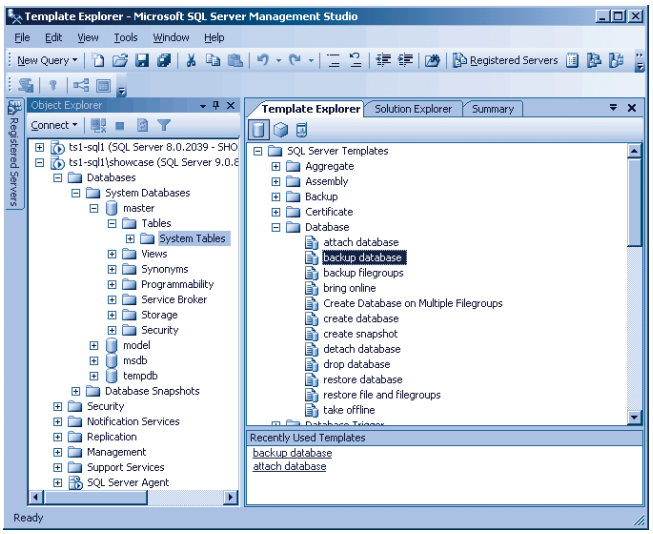


Figure 3. SQL Server 2005 provides integrated management features through the new SQL Server Management Studio

Visual Studio, encompasses business intelligence development and extends developer productivity benefits into virtually every aspect of SQL Server development. In addition, SSIS's enterprise ETL features, including high-performance data-movement and advanced data-cleansing capabilities, enable easy creation and maintenance of a centralized data store to represent your entire business.

Timely business insights. By letting your organization quickly and easily integrate and analyze business data from multiple heterogeneous data sources, SQL Server's business intelligence capabilities can help your business gain crucial insight into its markets and compete effectively. In SQL Server 2005, minimized data latency enabled by SSIS, SQL Server Query Notifications and real-time OLAP capabilities within the UDM, timely data-mining predictions, and compelling presentation options (including Reporting Services, Microsoft Office, and Microsoft SharePoint® Server) give developers multiple ways to build solutions that help users attain rapid and critical business insights.

Advanced analytics and reporting. Through rich reporting, advanced analytics, data mining, and familiar tools such as Microsoft Office, SQL Server 2005 lets you give users the power to build their own views of business information. With the introduction of Report Builder, the Reporting Services ad hoc reporting tool, end users can intuitively navigate data sources to build and share reports that let them drill down from summary data into details. Report Builder provides a data-modeling tool that you as developer can use in a one-time process to create an intuitive semantic model for your users. With this semantic model, users can then use a lightweight browser-deployed .NET Smart Client interface for ongoing ad hoc report generation. SQL Server 2005 also provides rich analytics that you can use to build applications that integrate data-mining models into daily business operations. SQL Server 2005 adds six data-mining algorithms.

Competitive features

A SQL Server 2005 upgrade also presents a great opportunity to migrate applications to SQL Server. Features such as snapshot-related connection contexts, Read Committed Snapshot Isolation, and Snapshot Isolation let many Oracle-hosted applications behave on SQL Server 2005 just as they do on Oracle platforms. Impressive enhancements to the T-SQL query language accelerate SQL Server's competitive lead. T-SQL now includes constructs such as EXCEPT and INTERSECT, PIVOT and UNPIVOT, RANK, and TOP N Sort in addition to common table expressions (CTEs), which enable advanced recursive queries. And the release adds error-handling capabilities with TRY…CATCH statements. You can also use Data Definition Language (DDL) triggers, a special kind of trigger that fires in response to DDL statements, to perform administrative tasks in the database, such as auditing and regulating database operations.

Compared to previous SQL Server releases, SQL Server 2005 features a simplified licensing model (per processor or per server with client access licenses) for every feature a specific software edition offers. This enables SQL Server 2005 to play a central role in an organization where you can extend the value of an initial licensing investment by using additional features of the comprehensive platform at no incremental cost. For example, you can expand a SQL Server 2005 data-storage platform upgrade to also offer management, replication, analysis, and reporting functionality.

In addition, SQL Server 2005 can be an excellent solution for small and medium businesses. Microsoft has responded to the needs of small businesses with the introduction of two entry-level licensing options: the freely distributable SQL Server 2005 Express and the cost-effective SQL Server 2005 Workgroup Edition. Designed to scale from the large enterprises down to small business, SQL Server 2005 enables the same level of performance, security, reliability, and business value to all customers.

Prepare to realize the benefits of upgrading to SQL Server 2005, which provides an enterprise data-management platform with advanced business intelligence functionality and impressive developer and administrator productivity features. Start digging into SQL Server 2005's advantages today and begin your upgrade planning so that your DBAs, developers, and entire organization can reap the benefits that SQL Server 2005 brings.

Douglas McDowell is director of operations for business intelligence at Solid Quality Learning. He is a mentor, solution architect, project manager, and a founder of Atlanta.mdf, an Atlanta SQL Server users' group. He is a Microsoft Certified Systems Engineer (MCSE), a Microsoft Certified Database Administrator (MCDBA), a Microsoft Certified Trainer (MCT), and winner of Microsoft's Worldwide Business Intelligence Solution of the Year.

Oracle 10g Real Application Clusters:

Building and Scaling Out a Database Grid

on Dell PowerEdge Servers and Dell/EMC Storage

Database grids allocate services across multiple, standards-based nodes to provide high performance and high availability—if a node fails or the workload fluctuates, the grid can automatically adapt in response. Oracle® 10g Real Application Clusters (RAC) software running on standards-based Dell™ PowerEdge™ servers and Dell/EMC storage can provide a flexible, reliable platform for a database grid. In particular, Oracle 10g RAC databases on Dell hardware can easily be scaled out to provide the redundancy or additional capacity that a grid environment requires.

BY ZAFAR MAHMOOD AND ANTHONY FERNANDEZ

Related Categories:

Cluster management

Clustering

Database

Dell PowerEdge servers

Dell/EMC storage

High availability (HA)

Oracle

Scalable enterprise

Visit www.dell.com/powersolutions for the complete category index.

In the past, IT professionals typically added computing resources to database servers by scaling up vertically—that is, when database application response time degraded because of additional users, more-complex queries, or database growth, administrators would add hardware or replace existing hardware with larger, faster components. Scaling up often involves proprietary hardware and large symmetric multiprocessing (SMP) systems, which must be replaced once they reach capacity limits. This approach can ultimately lead to significant losses—not only in terms of hardware resources, but also in terms of performance and availability. After reaching maximum capacity, further scaling up can become prohibitively expensive or even impossible until vendors release the next generation of hardware.

Scaling up can be problematic for software as well because database servers have internal mechanisms that handle locking and other multiuser issues. For example, the 3 GB limit on the maximum user-space memory available to a process on a 32-bit OS can limit the amount of data that a relational database management system (RDBMS) can cache on a single server. Software limitations become the primary impediment to continued scaling up. For this reason, SMP performance usually fails to demonstrate linear upward scalability as more and more processor power or memory is added. Typically, if this performance were graphed, after the curve started to flatten, the SMP system would require expensive hardware upgrades to gain miniscule performance improvements.

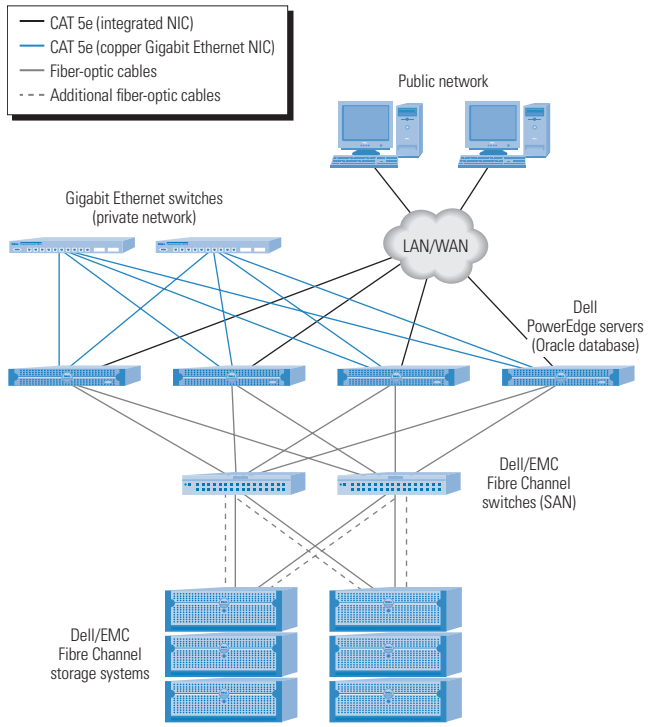


Figure 1. Typical four-node Oracle 10*g* RAC database cluster on Dell hardware

A scale-out database environment

A cost-efficient growth strategy is to implement a database system that can horizontally scale out. Using standards-based hardware, the scale-out environment can provide redundant servers, storage systems, and components—resulting in a high-performance, highly available architecture with low total cost of ownership (TCO). The major benefit of scaling out is the ability to incrementally expand database applications as business needs grow simply by adding standards-based hardware components to the existing infrastructure.

Implementing the scalable Dell and Oracle database environment

Oracle 10*g* Real Application Clusters (RAC) software on Dell PowerEdge servers and Dell/EMC Fibre Channel storage can be used to build a scalable database environment. A shared-storage database cluster built with these components is designed to provide cluster nodes with equal access to the same set of database objects. Each node in the cluster has its own set of Oracle background processes and memory structures, which are synchronized over a high-speed interconnect. Oracle 10*g* RAC uses Cache Fusion technology for global cache synchronization across nodes. Figure 1 depicts a typical four-node Oracle 10*g* RAC database cluster on high-availability Dell hardware. This architecture is designed to provide redundancy down to the component level.

Scale-out tasks for the Dell and Oracle 10*g* RAC platform

An Oracle 10*g* RAC database is designed with the inherent capability to scale out. RAC software includes built-in tools and utilities to scale out the database from the clusterware level to the database instance. For these tools to work, however, some basic prerequisites must be fulfilled before the RAC application can be scaled out:

- Additional nodes must be installed in the exact manner as the existing nodes were installed—with the same OS mount points and configuration parameters.
- Additional nodes must be made part of the cluster at the network layer.
- Additional nodes must be made part of the storage area network (SAN).
- Additional nodes must be prepared for either Oracle Cluster File System (OCFS) or Oracle Automatic Storage Management (ASM) database storage before they are made part of the Oracle 10*g* RAC database.

After these prerequisites are met, Oracle tools such as Oracle Universal Installer and Oracle Database Configuration Assistant can help ease the integration of an additional node into the existing cluster. The following steps can be taken to make an additional node part of the Oracle 10*g* RAC environment at the clusterware and database layers:

1. **Add the node at the Oracle clusterware layer:** In this step, the node is added to the existing cluster using Oracle Universal Installer. This process modifies the existing cluster layer and makes existing nodes aware of the newly added node. The required Oracle clusterware files and binaries are automatically copied to the added node from the existing cluster nodes, and the Oracle Notification Services are set up for cluster-wide events among all the nodes, including the newly added node.
2. **Add the node at the Oracle database layer:** This process extends the Oracle database home from the existing nodes to the newly added node. Through the use of Oracle Universal Installer, the required Oracle binaries are copied from the existing nodes to the newly added node, and the virtual IP address is set up for the newly added node. This process also defines the private interconnect to be used for cluster-wide Oracle Cache Fusion traffic. Oracle Cluster Repository information is updated to reflect the inclusion of another node in the cluster.
3. **Add the node at the Oracle database instance layer:** The final step is to extend the existing Oracle database instances to create and add the new instance of the database that will run on the newly added node. For this step, Oracle Database Configuration Assistant is used. After this process has

	Hardware	Software
Cluster nodes	Dell PowerEdge 2850 servers, each with: • Two Intel® Pentium® 4 processors at 3.6 GHz • 4 GB of RAM • 800 MHz frontside bus • 1 Gbps* Intel NIC for the public LAN • Two 1 Gbps LAN on Motherboards (LOMs) teamed for the private interconnect • Two QLogic QLA2340 HBAs • Dell Remote Access Controller	• Red Hat Enterprise Linux AS 4 QU1 • EMC® PowerPath® 4.4 • EMC Navisphere® agent • Oracle 10*g* R1 10.1.0.4 • Oracle ASM 10.1.0.4 • Oracle CRS 10.1.0.4 • Linux bonding driver for the private interconnect • Dell OpenManage
Storage	• Dell/EMC CX700 storage array • Dell/EMC Disk Array Enclosure (DAE) with 30 disks (73 GB 15,000 rpm) • Two 16-port Brocade SilkWorm 3800 Fibre Channel switches • Eight paths configured to each logical volume	• EMC FLARE™ Code Release 16
Network	• 24-port Dell PowerConnect™ 5224 Gigabit Ethernet switch for the private interconnect • 24-port Dell PowerConnect 5224 Gigabit Ethernet switch for the public LAN	• Linux binding driver used to team dual on-board NICs for the private interconnect

**This term does not connote an actual operating speed of 1 Gbps. For high-speed transmission, connection to a Gigabit Ethernet server and network infrastructure is required.*

Figure 2. Hardware and software configuration for the Oracle 10*g* RAC cluster

completed, database users can connect to the newly added node and perform database operations.

The final step in scaling out an existing Oracle 10*g* RAC database cluster accomplishes the following tasks:

- An ASM instance is created and started on each newly added node if the existing instances use ASM.
- A new database instance is created on each newly added node.
- High-availability components are created and configured.
- Node applications for the Global Services Daemon (GSD), Oracle Net Services listener, and Enterprise Manager agent are configured and started.
- The Oracle Net configuration is created.
- The new instance is started.
- Services on the new instance are created and started, if required.

For detailed information about these processes, refer to the Oracle 10*g* information available at www.dell.com/oracle.

Automating the Dell and Oracle 10*g* RAC scale-out tasks

Although Oracle 10*g* RAC provides tools and utilities to scale out the database, it cannot perform bare-metal deployments. Consequently, configuration tools are needed to help meet prerequisites before the Oracle scale-out tasks can be fully automated. Several third-party applications and tools are available to address this need. For example, Altiris® Deployment Solution™ for Dell Servers, which also integrates with Dell OpenManage™ software, can help facilitate bare-metal deployments and enable newly added nodes to be quickly integrated into the Oracle cluster. After the OS installation, the newly added node can be integrated seamlessly into an existing cluster using the Oracle tools discussed earlier in the "Scale-out tasks for the Dell and Oracle 10*g* RAC platform" section in this article.

Testing Oracle 10*g* RAC scalability

In June 2005, a team of Dell engineers tested the scalability of Oracle 10*g* RAC software on Dell hardware. Figure 2 shows the hardware and software configuration used in the test environment. The test started with one node and scaled up to four nodes. A typical online transaction processing (OLTP) workload was run on one-node, two-node, three-node, and four-node RAC configurations while all other variables—such as the size of the System Global Area (SGA), DB_BLOCK_SIZE, OS tuning parameters, storage configuration and disk sizes, and the number of database connections on each RAC node—remained constant. The only variable that changed was the number of cluster nodes. For each RAC configuration, the number of transactions per second was captured for comparison. Each node simulated 200 concurrent users performing database transactions. As Figure 3 shows, near-linear scalability was achieved from one to four nodes in terms of the number of transactions per second on each node.

Oracle best practices recommend that all nodes in a cluster have the same hardware and software configuration so that the scalability may be as linear as possible.

Oracle best practices recommend that all nodes in a cluster have the same hardware and software configuration so that the scalability may be as linear as possible. To compare test results when the hardware configuration is not identical, the test team ran a second test with the same scenario as the first test except that the fourth node was added to the cluster with only one host bus

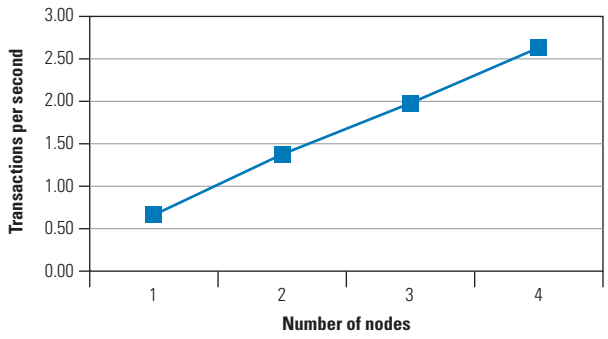


Figure 3. Transactions per second as identically configured nodes are added

adapter (HBA) and no teaming on the private interconnect network interface cards (NICs)—the other three nodes each had two HBAs and used NIC teaming. With this modified configuration, the fourth node did not have the I/O load-balancing capability that is possible through multipathing; and the bandwidth on the private interconnect for intra-node, memory-to-memory block transfers was limited. Because the fourth node was not configured like the other three cluster nodes, this node's number of transactions were almost half that of the other nodes (see Figure 4).

Note: The results of these tests are not actual numbers and have been normalized. These results represent the relative scalability of a RAC database with minimal database and OS tuning. Actual transactions per second will vary based on hardware configuration as well as OS and Oracle tuning parameters.

Oracle 10*g* RAC services framework

Scale-out capability is the basic component of grid computing. In a grid environment where large numbers of nodes are part of an Oracle 10*g* RAC database, applications connect to the RAC grid using RAC services. RAC services tightly integrate with Oracle Cluster Ready Services (CRS). Together, these services provide the following features and capabilities:

- Applications connect to services, which are defined within the database grid.
- Services can span multiple instances, with additional instances being made available in response to failures or workload demands.
- Services are available continuously, and the load is shared across one or more instances.
- When a RAC grid is configured, the nodes that will host the services can be one of two types: preferred instances, which are the first to start the service and are the primary nodes; and available instances, which are used to replace failed preferred instances or to meet increasing workload demands.
- Services are available somewhere in the database grid as long as one surviving node exists in the grid.

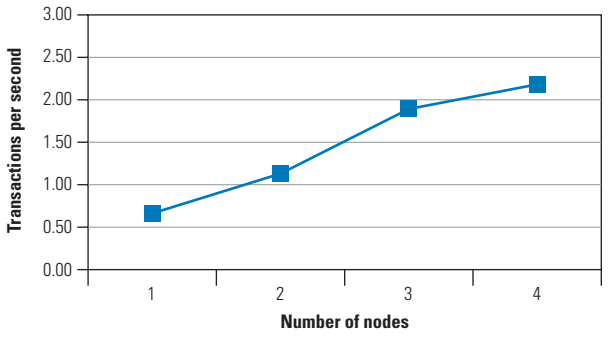


Figure 4. Transactions per second when fourth node has different configuration

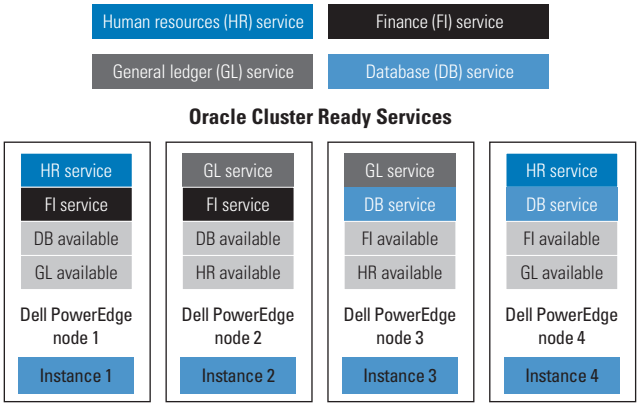


Figure 5. RAC database grid comprising four nodes and four services

Figure 5 depicts a four-node RAC grid with four services configured. Each service is configured with two primary instance nodes; the other two nodes serve as the available instances. These RAC services can span over to additional nodes in response to increased workloads on the preferred nodes or failures in the RAC grid.

Testing failover and load balancing in Oracle 10*g* RAC services

In the June 2005 study discussed earlier in this article, Dell engineers also tested the failover and load-balancing capabilities of an Oracle 10*g* RAC database. Using the same hardware and software setup, the team configured a two-node RAC database with a service called orasrv. The two nodes—oradb1 and oradb2—both served as preferred instances. An order-entry client application performed various Data Definition Language (DDL) and Data Manipulation Language (DML) operations on the RAC database using the orasrv service.

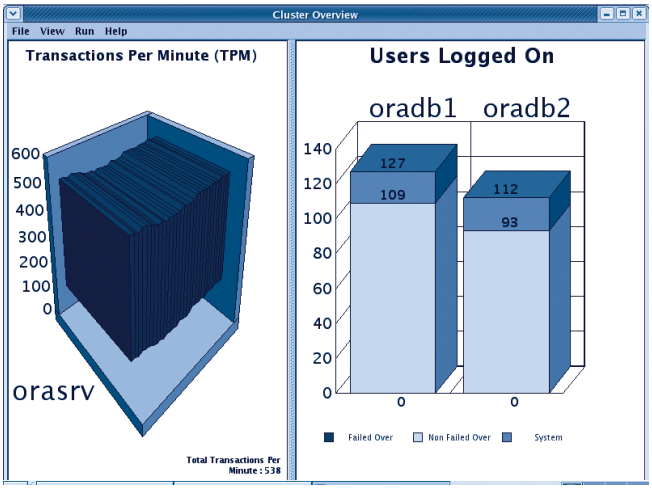


Figure 6. Transactions per minute and user connections when both database nodes are available

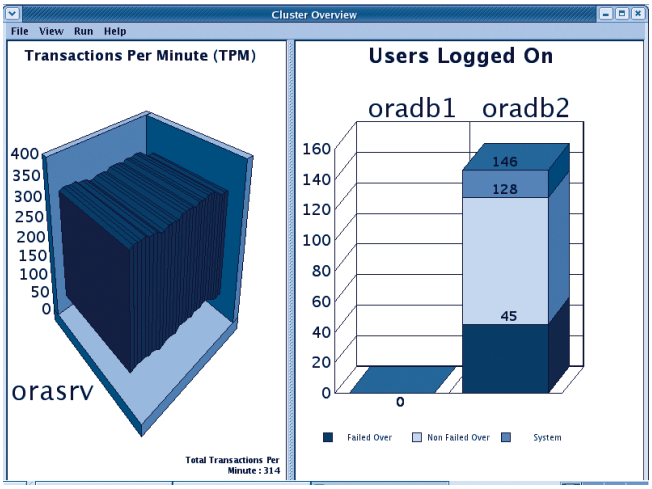


Figure 7. Transactions per minute and user connections when one node becomes unavailable

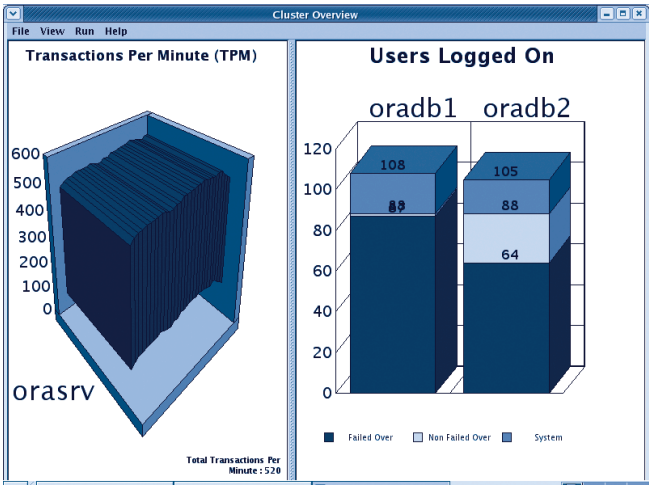


Figure 8. Transactions per minute and user connections when the failed node is back online

The client application provides a graphical user interface (GUI) that reports performance. Figure 6 shows the performance results, via this GUI, in terms of throughput (transactions per minute) and user connections when the client application was running on both nodes. When oradb1 was powered down, the user connections failed over to the remaining node (oradb2) using the Oracle Transparent Application Failover feature (see Figure 7). This feature enabled all the new connections to automatically go to oradb2 and allowed the database to continue executing transactions.

As soon as the failed node was brought back online, Oracle CRS automatically restarted the database instance and the orasrv service on oradb1. It also notified oradb2 of the other node's availability to handle connections and to execute database transactions. Over time, the client connections automatically load balanced across both nodes using the server-side load-balancing feature of Oracle 10*g* RAC (see Figure 8).

A scalable, highly available database grid

Today's data center demands an infrastructure that is highly available and scalable. Oracle 10*g* Real Application Clusters database software on Dell hardware is designed to provide high levels of redundancy and scalability. This database platform is designed to overcome the limitations that monolithic, single-node environments face in terms of CPUs, memory, I/O bandwidth, and interconnects. With a combination of tools provided by Oracle and third-party OS deployment vendors such as Altiris, administrators can easily deploy and scale out this database environment. Furthermore, Oracle 10*g* RAC is helping define grid computing technology through its robust services framework, which is designed to fail over, relocate, and expand services transparently in response to failures and workload demands.

Zafar Mahmood is a senior consultant in the Dell Database and Applications Team of Enterprise Solutions Engineering, Dell Product Group. He has been involved in database performance optimization, database systems, and database clustering solutions for more than eight years. Zafar has a B.S. and an M.S. in Electrical Engineering, with specialization in Computer Communications, from the City University of New York.

Anthony Fernandez is a senior analyst with the Dell Database and Applications Team of Enterprise Solutions Engineering, Dell Product Group. His focus is on database optimization and performance. Anthony has a bachelor's degree in Computer Science from Florida International University.

FOR MORE INFORMATION

Dell and Altiris deployment solutions:
www.dell.com/altiris

Dell and Oracle 10*g* database solutions:
www.dell.com/10g

Mahmood, Zafar, Joel Borellis, Mahmoud Ahmadian, and Paul Rad. "Enabling a Highly Scalable and Available Storage Environment Using Oracle Automatic Storage Management." *Dell Power Solutions,* June 2004. www.dell.com/downloads/global/power/ps2q04-008.pdf.

Oracle Database 10*g* documentation:
www.oracle.com/technology/documentation/database10g.html

Oracle Grid Computing:
www.oracle.com/technologies/grid/index.html

Oracle Real Application Clusters:
www.oracle.com/database/rac_home.html

Exploring Dell-Supported Configurations for

Oracle Database 10g Standard Edition with RAC

For small and growing enterprises, Dell and Oracle offer cost-effective Oracle® Real Application Clusters (RAC)–based cluster configurations on Microsoft® Windows Server™ 2003, Standard Edition, with Service Pack 1. These highly available clusters are offered as two-node direct attach and Fibre Channel configurations using low-cost storage based on Dell/EMC AX100 and CX300 arrays.

BY CHETHAN KUMAR

Related Categories:

Clustering

Dell PowerEdge servers

Dell/EMC storage

Microsoft Windows Server 2003

Oracle

Storage architecture

Visit www.dell.com/powersolutions for the complete category index.

Driven by business requirements for a low-cost, highly available database system, enterprises may now consider validated cluster configurations based on Oracle® Database 10*g* Standard Edition with Real Application Clusters (RAC). By bundling Oracle Database 10*g* Standard Edition with industry-standard Dell™ PowerEdge™ servers and low-cost storage such as Dell/EMC AX100 and CX300 arrays, small and growing enterprises can implement an efficient, highly available cluster database system that also offers a smooth upgrade path to Oracle Database 10*g* Enterprise Edition for enhanced scalability.

Architecture and components of the supported cluster configurations

Dell offers two configurations for Oracle Database 10*g* Standard Edition with RAC on Microsoft® Windows Server™ 2003, Standard Edition, with Service Pack 1 (SP1): direct attach cluster and Fibre Channel cluster.[1] The main difference exists in the use of Fibre Channel switches, which doubles the number of paths available from the nodes to the shared storage enclosures. The nodes are interconnected through a Gigabit Ethernet switch, forming a private network for cluster intercommunication.

In the direct attach cluster architecture, the hosts are connected directly to the storage, as shown in Figure 1. Each cluster node has two host bus adapters (HBAs), each of which connects to a different storage processor. These connections offer redundancy in case of failure of components such as an HBA, an optical cable connecting the host HBA to the storage, or a storage processor.

In the Fibre Channel cluster architecture, the cluster nodes are connected to storage through high-speed Fibre Channel switches, as shown in Figure 2. Each node has two HBAs, and each HBA is connected to a separate Fibre Channel switch. Each Fibre Channel switch in turn is connected to both storage processors in the external storage. As a result, each cluster node has four paths available for communicating with the shared storage. This switch fabric architecture is designed to offer high redundancy for an Oracle Database 10*g* Standard Edition with RAC cluster.

Nodes: Dell PowerEdge 2850 servers

In Dell-supported cluster configurations, Dell PowerEdge 2850 servers are the nodes. The PowerEdge 2850 is an affordable, full-featured, rack-mountable 2U server that

[1] A Fibre Channel cluster is also referred to as a storage area network configuration.

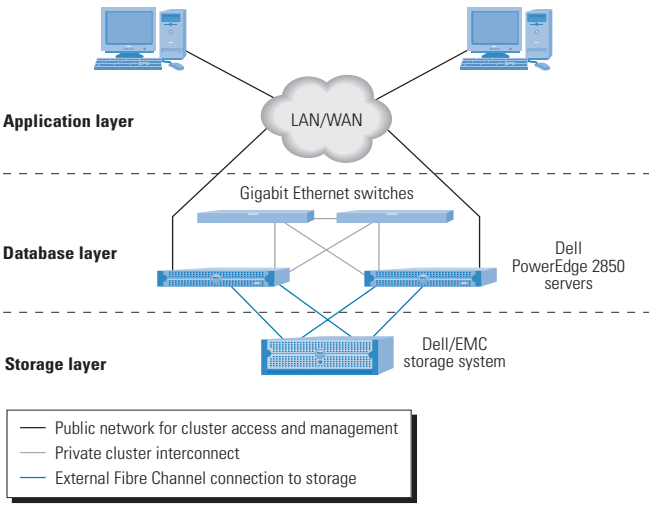


Figure 1. Dell/Oracle direct attach cluster architecture

is well suited for multiserver rack environments. It incorporates advanced technology such as dual Intel® Xeon™ processors, Peripheral Component Interconnect Express (PCI Express) I/O, the Intel E7520 chipset with support for dual-channel memory architecture, and up to 16 GB of double data rate 2 (DDR2) at 400 MHz SDRAM memory. The PowerEdge 2850 also includes dual Intel PRO/1000 MT on-board network interface cards (NICs). Supported HBAs include the QLogic QLA200,[2] QLA2340, QLA2342, and QLE2360 adapters and the Emulex LP10000 and LP1050e adapters. The nodes run Microsoft Windows Server 2003, Standard Edition, with SP1.

Storage: Dell/EMC AX100 and CX300 arrays

For the cluster configurations, Dell offers a powerful yet economical storage enclosure in the form of the Dell/EMC AX100 and an entry-level product for a storage area network (SAN) configuration in the form of the Dell/EMC CX300. The Dell/EMC AX100 features a compact, rack-mountable 2U storage enclosure that includes 3 to 12 Serial ATA (SATA) disk drives. A single Dell/EMC AX100 array is designed to provide storage capacity ranging from 480 GB to 3 TB.

The Dell/EMC CX300, an entry-level RAID storage system with Fibre Channel disk drives, is capable of operating as direct attach storage. Alternatively, it can be configured in a SAN. The Dell/EMC CX300 incorporates updated storage processors while remaining compatible with existing CX series components such as disks, Disk Array Enclosures (DAEs), 40U racks, and software. This compatibility facilitates data-in-place upgrades for existing Dell/EMC CX200, CX400, and CX600 arrays. The Dell/EMC CX300 consists of a 2 Gbps Disk Processor Enclosure (DPE2) and an optional 2 Gbps DAE (DAE2). In a single 3U chassis, the DPE2 contains two storage processor boards; up to 15 one-inch Fibre Channel disk drives; and redundant, hot-swappable fans and power supplies.

I/O path management: EMC PowerPath software

The cluster nodes run EMC® PowerPath® software to enable multipath management and path failover. PowerPath includes integrated volume management capabilities that help keep costs low by maximizing the efficiency of storage allocation. Key features of PowerPath include:

- Multipath management for high availability and performance
- Dynamic multipath load balancing
- Proactive I/O path testing and automatic path recovery
- Automatic path failover

Database: Oracle Database 10*g* Standard Edition

Oracle Database 10*g* Standard Edition is designed to be a highly efficient, reliable, secure database that is well suited for small and growing data center environments. It supports a clustered environment with a maximum of four processors. Because each node in the two-node Dell cluster has dual processors, the configurations offer the maximum processing capacity supported by Oracle licensing for Oracle Database 10*g* Standard Edition.

Oracle Database 10*g* Standard Edition includes Oracle RAC capabilities to protect against hardware failures and the flexibility to scale up hardware resources. The database is easy to install and configure and includes clusterware and storage management capabilities. In addition, Oracle Database 10*g* Standard Edition is built from the same code base as Oracle Database 10*g* Enterprise Edition, so it can scale easily to Enterprise Edition. Oracle Database 10*g* Standard Edition also offers inherent built-in support for clusters, including Automatic Storage Management (ASM), Cluster Ready Services (CRS), and Cluster Synchronization Services (CSS).

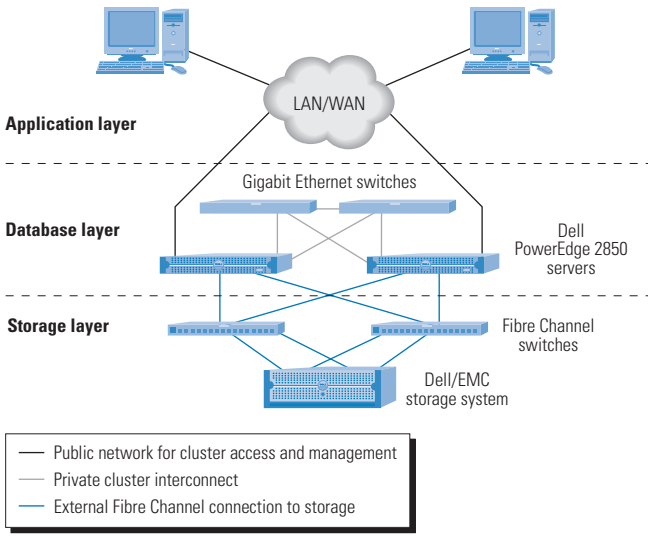


Figure 2. Dell/Oracle Fibre Channel cluster architecture

[2] The QLogic QLA200 is supported only with the Dell/EMC AX100 in a Fibre Channel cluster configuration.

Automatic Storage Management. ASM provides a simple storage management interface that is consistent across storage and server platforms. It virtualizes the database into disk groups and automates the file management within those disk groups. Data is spread evenly across available storage disks to minimize hot-spot formation, thus helping to eliminate manual I/O performance tuning. ASM also enables database administrators to change the storage configuration without taking the database offline. ASM is designed to automatically rebalance the files across disks after the disks are added or dropped.

Cluster Ready Services. CRS offers a high-availability framework for uninterrupted operation of the database. When a failure occurs, the services are designed to continue functioning on the nodes and the instances unaffected by the failure. CRS also provides interfaces to relocate, disable, and enable services.

Cluster Synchronization Services. CSS provides cluster management and node management. It monitors ASM and ASM shared components such as disks and disk groups. ASM registers itself and disk groups it has mounted with CSS on startup, which allows CSS to keep the disk group metadata in sync across cluster nodes. Any new disk groups that are created are dynamically registered and communicated to the other nodes.

Features and benefits of the cluster configurations

The Dell-supported configurations for Oracle Database 10*g* Standard Edition with RAC on Microsoft Windows Server 2003, Standard Edition, with SP1 are well suited for small and growing enterprises—offering highly available, scalable, easy-to-deploy database clusters.

High availability through redundancy

The Dell-supported configurations are redundant at both the hardware and the software level. Each of the two nodes in the cluster has dual Intel Xeon processors. Both nodes share common storage connected through multiple optical paths.

The storage array has dual processors, and each node is connected to each storage processor through separate HBAs. Each storage processor forwards the I/O requests to the disks it controls. In the event of a failure, the request is rerouted through the other functional storage processor to the relevant disk.

Each node runs a separate instance of ASM and database services. Moreover, each node is capable of servicing client requests if any of the hardware components in the cluster—HBA, NIC, or storage processor—fails. The redundancy of each configuration's hardware and software enables the cluster to be highly available.

Scalability for future growth

The Dell/EMC CX300 offers a good entry point for growing organizations because it provides the capability to build a SAN with other Dell/EMC CX series storage arrays. Organizations can also increase the number of nodes in a cluster by adding more Dell PowerEdge servers as they migrate to an enterprise database. This hardware scalability and the upward compatibility of Oracle Database 10*g* Standard Edition with Oracle Database 10*g* Enterprise Edition can enable organizations to create a cost-effective configuration that can be scaled incrementally in response to business requirements.

Ease of deployment

Deploying a database server from bare metal can be a daunting task for many administrators. It can be even more difficult to install a cluster of servers and keep them consistent. Finding the optimal combination of validated hardware and software is crucial for any database infrastructure. There is no guarantee that each hardware and software component, which may have been tested individually, will work as expected when put together in a clustered environment. To help reduce the burden on IT departments for finding a working set of components, Dell has tested and validated the Dell-supported hardware configuration discussed in this article.[3]

The Dell Deployment CD helps simplify the deployment of software components that are required for the proper functioning of the cluster. Administrators can follow well-defined steps by using the Dell Deployment CD to install the OS, drivers, and Oracle software packages—often enabling clusters to be deployed within a few hours. All required steps are documented in the *Oracle Database 10*g *Standard Edition Real Application Clusters for Microsoft Windows Deployment Guide*.

Affordable clusters for high availability

Both Dell-supported cluster configurations for Oracle Database 10*g* Standard Edition with RAC on Microsoft Windows Server 2003, Standard Edition, with SP1 are designed to provide high availability that is normally reserved for enterprise systems. Hardware redundancy is available at each level—node, HBAs, NICs, storage connections, and storage—contributing to the cluster's high availability. Moreover, key components in the hardware and software stack are cost-effective—including the Dell PowerEdge 2850 servers, QLogic and Emulex HBAs, Dell/EMC AX100 and CX300 storage, and Oracle Database 10*g* Standard Edition. The cluster configurations also enable upward compatibility with Oracle Database 10*g* Enterprise Edition, thereby helping to reduce total cost of ownership as enterprise requirements grow.

Chethan Kumar is a systems engineer and advisor in the Database and Applications Group at Dell. He has an M.S. in Computer Science and Engineering from The University of Texas at Arlington.

[3] For more information, refer to the Solution Deliverable List, which is accessible by visiting www.dell.com/10g, clicking "Oracle Database 10*g* Standard Edition with Real Application Clusters" under "Microsoft Windows 2003 SP1" and then clicking "Dell PowerEdge and Oracle Database 10*g* Standard Edition with Real Application Clusters on Windows 2003 Standard Edition SP1 (updated 7/01/2005)".

High-Availability Blade Server Clustering with the Dell PowerEdge Cluster FE555W

Dell high-availability clusters such as the Dell™ PowerEdge™ Cluster FE555W are based on cost-effective, modular Dell PowerEdge servers. The flexible, rack-mountable PowerEdge 1855 blade server supports 10 removable server blades, fabric switches, and Ethernet switches—all efficiently stored within the 7U Dell Modular Server Enclosure. By incorporating the dense hardware configuration and redundant components of the PowerEdge 1855 blade server, the PowerEdge Cluster FE555W is designed to save data center space and enable high availability.

BY GREG BENSON, BRYANT VO, AND FARRUKH NOMAN

Related Categories:

Blade servers

Clustering

Dell PowerEdge blade servers

Dell PowerEdge servers

Dell/EMC storage

Fibre Channel

High availability (HA)

Microsoft Windows

Visit www.dell.com/powersolutions for the complete category index.

Dell high-availability (HA) clusters are designed to provide maximum uptime for business-critical environments with continued efficiency and effectiveness. The Dell PowerEdge Cluster FE555W incorporates the flexible, industry-standard Dell PowerEdge 1855 blade server—a rack-mountable server system that supports 10 removable server blades, fabric switches, and Ethernet switches within the 7U Dell Modular Server Enclosure. The PowerEdge Cluster FE555W supports two server modules running the Microsoft® Windows® 2000 Advanced Server OS and two to eight server modules running Microsoft Windows Server™ 2003, Enterprise Edition or Enterprise x64 Edition. For storage, the cluster uses Fibre Channel–based Dell/EMC CX300, CX500, or CX700 arrays.

Understanding the building blocks of the Dell PowerEdge Cluster FE555W

By incorporating redundant hardware features, Dell HA clusters are designed to ensure that resources remain available to network clients and the applications can recover from failure. The PowerEdge Cluster FE555W comprises cluster nodes, Fibre Channel modules, Ethernet modules, and storage systems.

Cluster nodes. The individual server blades of the Dell PowerEdge 1855 blade server, which use dual Intel® Xeon™ processors, function as the cluster nodes. Multiple clusters can be configured depending upon the number of available blades. Supported configurations may use a maximum of two nodes per cluster running Windows 2000 Advanced

Module type	Cluster configuration	Private network configuration
Dell PowerConnect 5316M Gigabit Ethernet switch	Two or more cluster nodes reside in one PowerEdge 1855 blade server	The private network is established using internal connections in the Dell Modular Server Enclosure.
	Two or more cluster nodes reside in two PowerEdge 1855 blade servers	The private network switch module on one PowerEdge 1855 blade server is connected to the private network switch module on the other PowerEdge 1855 blade server using a standard Ethernet cable.
Ethernet pass-through module	Two nodes reside in one or two PowerEdge 1855 blade server(s)	A standard Ethernet cable is connected directly to the corresponding cluster node ports on the Ethernet pass-through module. *or* Standard Ethernet cables are connected from the pass-through module ports (corresponding to the cluster nodes) to an external switch.
	Three or more cluster nodes reside in one or two PowerEdge 1855 blade server(s)	The private network is established by connecting standard Ethernet cables from the pass-through module ports (corresponding to the cluster nodes) to an external switch.

Figure 1. Private network configurations for different Ethernet modules

Server and a maximum of eight nodes per cluster running Windows Server 2003, Enterprise Edition or Enterprise x64 Edition.

Fibre Channel modules. Each server blade on the PowerEdge 1855 blade server includes an embedded dual-port Dell 2342M Fibre Channel host bus adapter (HBA). Each HBA Fibre Channel port is connected through the midplane to either an embedded Fibre Channel switch or a Fibre Channel pass-through module. The Fibre Channel switch connecting to the midplane is a 14-port, 2 Gbps Brocade SW3014 or McDATA 4314 switch with 10 downlink ports (for internal connection) and 4 uplink ports (for external connection to storage). The pass-through module provides one-to-one connection to external storage area network (SAN) devices and supports both 2 Gbps and 4 Gbps Brocade and McDATA Fibre Channel switches.

Ethernet modules. Each server blade is configured with two integrated Gigabit Ethernet[1] network interface cards (NICs) to provide cluster heartbeat communication and public client access on two separate networks. These NICs are internally connected to a corresponding port on either an Ethernet pass-through module or a switch module. The Ethernet pass-through module provides a non-switched, one-to-one connection between the server blade and the external Gigabit Ethernet device or another server blade. These ports are preset to communicate only at 1 Gbps and will not auto-negotiate to a slower speed. In contrast, the Gigabit Ethernet switch module—the Dell PowerConnect™ 5316M switch—provides a switched connection with 6 uplink ports and 10 downlink ports. The option of using any one of these modules for the private and public networks provides flexibility to meet different user requirements. In a configuration with a large number of cluster nodes or with multiple clusters, best practices recommend employing switch modules. Figure 1 shows the private network cabling for pass-through and switch modules in various cluster configurations.

Storage systems. Server nodes within a cluster can share one or more external storage systems connected through Fibre Channel switch fabric modules or pass-through modules. The PowerEdge Cluster FE555W solution supports Dell/EMC CX300, CX500, and CX700 storage arrays, which can be configured through a management station using EMC® Navisphere® Manager software.

Choosing the appropriate configuration

Direct attach storage (DAS) and SAN-based configurations are widely used for HA clusters. Choosing an appropriate configuration depends on the load of the cluster nodes and the application servers that need to participate in the cluster. A DAS-based cluster, suitable for light loads, requires the use of Fibre Channel pass-through modules to communicate directly with the storage system; whereas a SAN-based cluster uses an embedded fabric switch to open the scope for multiple nodes in the cluster to share heavy loads. Using the PowerEdge Cluster FE555W in a DAS environment (where the cluster is typically limited to two nodes) would not be an efficient use of the large number of resources, unless multiple clusters were employed to make use of all the server blades.

The current release of the Dell PowerEdge Cluster FE555W supports a 2 Gbps end-to-end cluster environment in a direct attach configuration and both 2 Gbps and 4 Gbps cluster environments in a SAN-based configuration. In the future, Dell plans to provide support for 4 Gbps end-to-end cluster environments in both direct attach and SAN configurations. Based on the enterprise structure, specific storage requirements, and the number of clusters, IT organization can implement various SAN configurations.

The current release of the Dell PowerEdge Cluster FE555W supports a 2 Gbps end-to-end cluster environment in a direct attach configuration and both 2 Gbps and 4 Gbps cluster environments in a SAN-based configuration.

Single-server, single–storage system configuration. In this configuration, a maximum of four ports are available from the fabric switches to provide connectivity to a dual–storage processor (SP) storage system with four ports. Based on the available ports, any Dell/EMC CX storage array can be connected to the

[1] This term does not connote an actual operating speed of 1 Gbps. For high-speed transmission, connection to a Gigabit Ethernet server and network infrastructure is required.

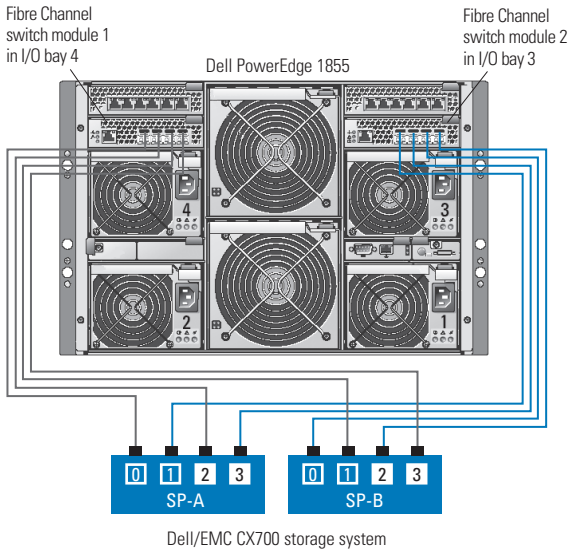


Figure 2. Connections in a single-server, single–storage system cluster configuration

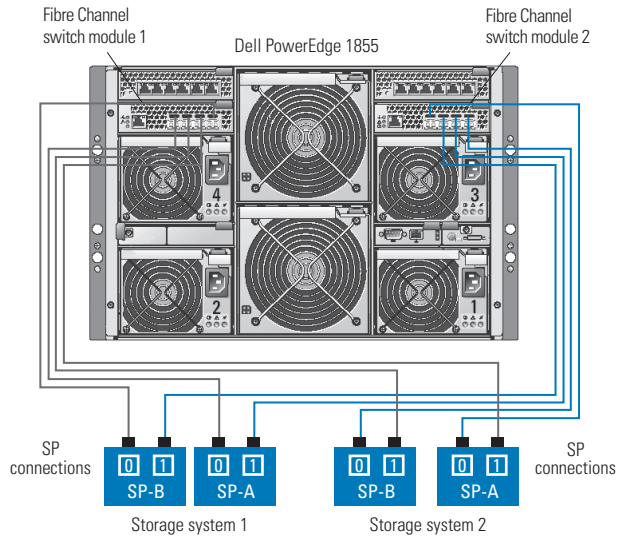


Figure 3. Connections in a single-server, dual–storage system cluster configuration

PowerEdge 1855 blade server. Figure 2 shows the redundant connections to a Dell/EMC CX700 array.

Single-server, dual–storage system configuration. Cluster storage capacity can be increased by connecting to two storage systems. Only two dual-SP storage systems with four ports (two from each SP) can be connected to the fabric ports on the PowerEdge 1855 blade server, unless an external fabric is linked to the ports via Inter-Switch Links (ISLs). Figure 3 depicts the cluster configuration with two storage systems.

Dual-server, single–storage system configuration. Multiple nodes or clusters can share a single storage system. In this scenario, if multiple PowerEdge 1855 blade servers are employed, then each server requires a direct path to the storage system to help ensure high availability for applications. At least two ISLs are required to provide connectivity between each Fibre Channel switch of the PowerEdge 1855 blade servers. A link to connect the private networks is also required if a cluster is using server blades from both PowerEdge blade servers. Figure 4 shows the connections of two PowerEdge 1855 blade servers to a single storage system.

Dual-server, dual–storage system configuration. A second storage system can be added to a dual-server cluster to increase the storage capacity and help ensure high availability by providing direct paths from Fibre Channel modules of both PowerEdge 1855 blade servers. Figure 5 shows the connections between two PowerEdge 1855 blade servers and two storage systems.

Setting up the SAN

Setting up the SAN involves configuring the Fibre Channel connections from the HBAs to the storage targets through the fabric. Logistically, this setup requires administrators to configure the Fibre Channel daughtercard (HBA) on each blade, set up the storage access, and create fabric zones to secure host-to-storage connectivity. A direct attach configuration employs the same set of rules except for fabric zoning.

Fibre Channel HBA setup. The Fibre Channel daughtercard on each server blade in a PowerEdge 1855 blade server must be configured for the SAN topology. Administrators can verify or change these settings by using the QLogic SANsurfer utility or the Fast!UTIL option during the boot-up process. Through either tool, the connection and topology settings can be configured for a fabric or direct attach environment.

Storage setup. The storage system should be updated with the core software, and EMC Access Logix™ software should be enabled

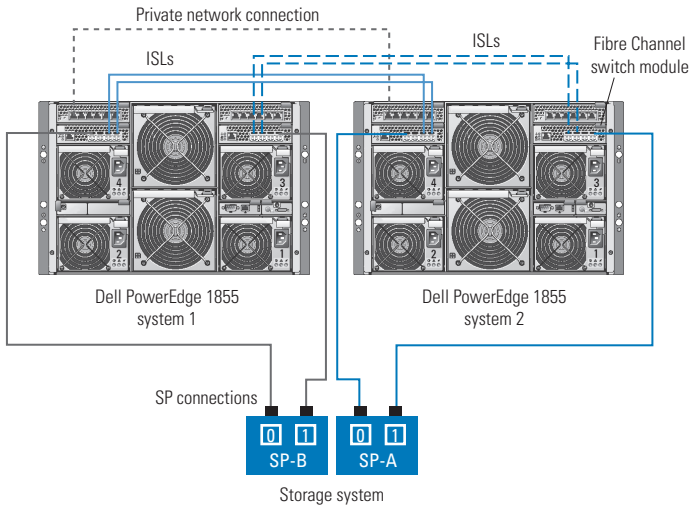


Figure 4. Connections in a dual-server, single–storage system cluster configuration

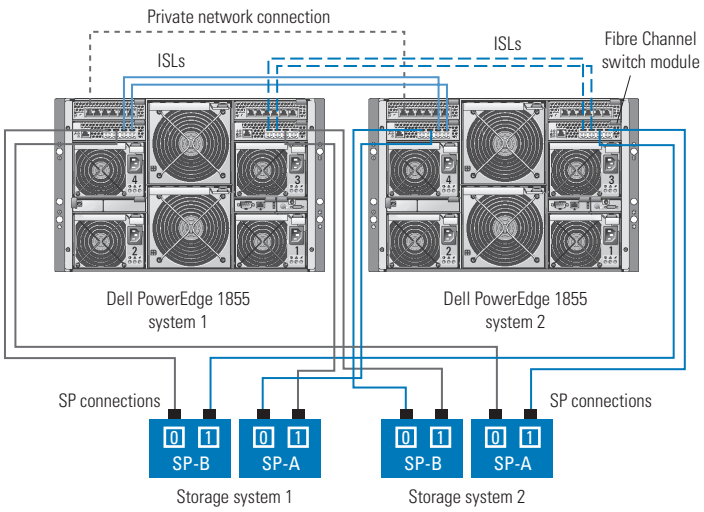


Figure 5. Connections in a dual-server, dual–storage system cluster configuration

to help prevent data corruption. Access Logix software is highly recommended on a Dell/EMC CX storage system if one of the following conditions is true:

- Two or more stand-alone servers or clusters are configured to access the same storage system.
- EMC MirrorView™, SnapView™, and SAN Copy™ software are installed on the storage system and being used by the cluster.

Access Logix restricts access to logical units (LUNs) by organizing them into storage groups for particular hosts. Administrators can activate these access restrictions by enabling Access Control (a feature of Access Logix) in the storage system properties and placing the hosts and their assigned LUNs in a storage group. A set of hosts residing in one storage group will not be allowed to join any other storage group.

Fabric zones setup. Zoning is implemented on Fibre Channel switches to isolate servers and storage systems from each other. In an environment that uses zoning and Access Logix software, multiple PowerEdge clusters can share Dell/EMC storage systems in a switched fabric. Dell supports only single-initiator zoning, in which each Fibre Channel daughtercard port resides in a separate zone with the storage ports—to help ensure that Fibre Channel communication between the cards and their target storage systems do not affect one another. Each Fibre Channel port can be connected to a maximum of two storage systems.

Configuring the cluster

Once all the cluster components are configured properly and the EMC Navisphere agents are active on the cluster nodes, the Fibre Channel HBAs should be able to make logical connections to the storage. The Navisphere agent allows each host (cluster node) to register with the storage system. EMC PowerPath® software also should be installed on the hosts to provide automatic rerouting features for the Fibre Channel I/O traffic, to help offset primary-path failures and address load-balancing needs.

The participating cluster nodes are configured in a storage group with shared LUNs. These LUNs, which can now be accessed from the disk management feature of each node, are prepared with unique disk signatures to actively participate in a cluster. The public and private networks are also verified to be uniquely identifiable in separate domains. Microsoft Cluster Service can be configured with the available resources on one node at a time. If the cluster is formed properly, resources can fail over between different nodes.

Gaining high availability and modularity

Business-critical applications that require maximum uptime are excellent candidates for clustering. The Dell PowerEdge Cluster FE555W can help deliver continuous infrastructure availability for critical applications by leveraging the flexibility and scalability of a SAN. This approach provides enterprise-class features while enabling the PowerEdge Cluster FE555W to help organizations keep acquisition costs low, maximize rack density, and enhance power efficiency.

Greg Benson is a project manager in the Dell Scalable Systems Group. He has a B.A. in Business Administration with a minor in Economics from the University of Florida.

Bryant Vo is a systems engineer in the High-Availability Cluster Development Group at Dell. His current interests include business continuity, clustering, system architecture, and storage technologies. He has a B.S. in Computer Science from the University of Houston.

Farrukh Noman is a systems engineer in the High-Availability Cluster Development Group at Dell. His current interests include development of solutions for Fibre Channel–based DAS and SAN clusters, as well as Internet SCSI (iSCSI) networks. Farrukh has an M.S. in Computer Engineering from the University of Minnesota, Twin Cities.

FOR MORE INFORMATION

Dell high-availability clusters:
www.dell.com/ha

Realizing Multi-Core Performance Advances in Dell PowerEdge Servers

With the introduction of multi-core Intel® Xeon™ processors, IT organizations have the opportunity to migrate to multi-core platforms without changing the underlying system architecture. Benchmark tests on Dell™ PowerEdge™ servers indicate that multi-core processors can help increase performance by more than 50 percent on existing Intel Pentium® or Intel Xeon architectures.

BY JOHN FRUEHE

Related Categories:

Dual-core technology

Intel

Multi-core technology

Multiprocessor (MP)

Performance

Processors

Scalable enterprise

System architecture

Visit www.dell.com/powersolutions for the complete category index.

Multi-core technology—which integrates two or more processor logic cores into a single physical processor—has been available for several years. However, it has traditionally been used in niche processors or in high-end proprietary systems. Now, Intel Corporation is building multi-core processors based on its existing Intel Pentium and Intel Xeon architectures. This approach enables multi-core processors to be backward compatible with single-core platforms, thereby easing the deployment of this emerging technology in critical environments. Although multi-core technology can deliver significant performance gains, IT organizations must plan migrations carefully to help ensure that multi-core technology is introduced in a way that helps to reduce long-term costs, not increase them.

Multi-core technology in Dell PowerEdge servers

Dell is taking a leadership role in advancing enterprise data center technology by delivering a consistent architecture that enables organizations to move from single-core processors to multi-core processors on the same underlying system architecture. In comparison, many vendors either provide multiple multi-core architectures to address different performance objectives or force an architecture change in the migration to multi-core performance. Adding multiple architectures to accommodate multi-core processors can increase complexity and total cost of ownership for IT organizations. Changing architectures to take advantage of multi-core processors can lead to higher initial investments and fewer long-term cost benefits than evolving to multi-core architecture on a system platform that is consistent with current single-core processor architecture. The optimal solution is to migrate to multi-core on existing architecture that is already deployed. This approach is designed to reduce data center complexity as well as costs.

Dell not only adheres to common architectures such as Intel Pentium and Intel Xeon processors, but also makes migrating to multi-core processors easy. Dell PowerEdge 1850, PowerEdge 2800, and PowerEdge 2850 servers share a common system image that enables backward compatibility with existing models. Whether single-core or dual-core processors are used, the system image is designed to work with other products in the PowerEdge 1850, PowerEdge 2800, and PowerEdge 2850 server family. In addition, the PowerEdge 1855 blade server supports multi-core processors and has a backward-compatible system image to support both single-core and dual-core processors in PowerEdge 1855 systems.

Single-socket, multi-core PowerEdge servers

Based on Intel Pentium 4 technology, the Intel Pentium D processor was the first server processor from Intel to integrate multi-core technology. This processor was designed for single-socket servers, which are predominantly used in small businesses, Web server farms, and development environments.

Single-socket servers can achieve significant benefits by integrating multi-core processor technology. For example, many small organizations cannot afford to use one server for each application, as large enterprises often do. Cost constraints and rapid growth typically require these organizations to run several applications together on a single server—which is an optimal environment for multi-core processor technology. In Web environments, complex applications such as Java and Secure Sockets Layer (SSL) usually run on front-end servers, which typically require a large amount of processing performance. Multi-core processors are designed to deliver high performance while helping to preserve the price/performance that is required for an effective single-socket server.

Comparing single-core and dual-core performance

In October 2005, Dell engineers, using several industry-standard benchmarks, tested the performance of the dual-core Dell PowerEdge 850 server and compared it to the performance of its predecessor, the single-core PowerEdge 750 server (see Figure 1). In these tests, the PowerEdge 850 was equipped with a dual-core Intel Pentium D processor at 3.0 GHz, and the PowerEdge 750 was equipped with a single-core Intel Pentium 4 processor at 3.0 GHz.

Benchmark	Single-core PowerEdge 750	Dual-core PowerEdge 850	Performance increase
SPECint_rate2000	16.8	30.5	81.55%
SPECjbb2000	46,632	67,371	44.48%
WebBench 6.0	2,608	3,303	26.65%

Figure 1. Benchmark performance results for single-socket Dell PowerEdge servers[1]

For processor-intensive, single-socket applications, the SPECint_rate2000 benchmark from the Standard Performance Evaluation Corporation (SPEC) is a good indicator of relative processor performance gains. On this benchmark, the PowerEdge 850 provided an 81.55 percent increase in performance over the PowerEdge 750. Technical applications and compute-intensive applications typically benefit most from such performance gains.

In addition, the Dell performance team used SPECjbb2000 to test performance for Java-based applications, which rely heavily on processor performance and memory latency. On this benchmark, the PowerEdge 850 achieved a 44.48 percent performance increase over the PowerEdge 750. Java is a commonly deployed application on rack-dense, single-socket 1U servers. Because Java tests stress both processor and memory, performance gains are expected to be less than those on the processor-centric SPEC benchmarks.

Another key area for single-socket server deployment is in large Web farms that run edge-of-network applications or Web hosting. While these applications do rely on processor and memory resources, they typically rely more on the I/O subsystem than the applications represented by SPECint_rate2000 and SPECjbb2000. The WebBench 6.0 benchmark is a Common Gateway Interface (CGI)–based test suite that measures the performance output for a Web-serving application. Equipped with a Peripheral Component Interconnect (PCI) Express bus, the PowerEdge 850 delivered a 26.65 percent performance increase over the PowerEdge 750, which was equipped with a PCI Extended (PCI-X) bus. This performance gain can provide a significant advantage for front-end Web servers.

In each benchmark studied for this article, the PowerEdge 850 server outperformed the PowerEdge 750 server. This performance gain was the result of two factors: the generational change from the PowerEdge 750 to the PowerEdge 850, which provided increased memory speed, and the advancement from a single-core to a dual-core processor architecture.

Dual-socket, multi-core PowerEdge servers

Dual-socket servers represent the majority of x86-based deployments worldwide. These servers support varying workloads—from simple file-and-print sharing to complex high-performance computing (HPC), technical, Java, and database applications.

[1] Performance results are based on benchmark tests performed by Dell labs in October 2005. Dell PowerEdge 850 configuration included one dual-core Intel Pentium D processor at 3.0 GHz with 1 MB level 2 (L2) cache per core, 2 GB of double data rate 2 (DDR2) RAM at 667 MHz, an 800 MHz frontside bus (FSB), and Microsoft Windows Server™ 2003 with Service Pack 1 (SP1). Comparable PowerEdge 750 configuration included one single-core Intel Pentium 4 processor at 3.0 GHz with 1 MB L2 cache, 2 GB of DDR RAM at 400 MHz, an 800 MHz FSB, and Windows Server 2003 with SP1. Actual performance will vary based on configuration, usage, and manufacturing variability.

Benchmark	Single-core architecture	Dual-core architecture	Performance increase
SPECint_rate2000			
PowerEdge 1850	38.1	58.5	53.54%
PowerEdge 1855	38.3	58.9	53.79%
PowerEdge 2800	38.2	58.5	53.14%
PowerEdge 2850	38.1	58.7	54.07%
SPECjbb2000			
PowerEdge 1850	104,172	149,801	43.80%
PowerEdge 1855	104,778	151,061	44.17%
PowerEdge 2800	104,443	150,700	44.29%
PowerEdge 2850	104,139	150,151	44.18%
TPC-C			
PowerEdge 2800 (performance)	28,122	38,622	37.57%
PowerEdge 2800 (price/performance)	US$1.40/tpmC	US$0.99/tpmC	29.29%
MMB3			
PowerEdge 1850	8,000	9,500	18.75%

Figure 2. Benchmark performance results for dual-socket Dell PowerEdge servers[2]

Dual-socket servers also typically provide a large proportion of enterprise services, so performance increases for these servers can enable significant productivity benefits for the entire enterprise, not just for a few niche applications.

Dual-socket Dell PowerEdge servers—such as the Dell PowerEdge 1850, PowerEdge 1855, PowerEdge 2800, and PowerEdge 2850 servers—use the Intel Xeon processor, which is a common platform for myriad x86 applications. The prevalence of dual-socket servers and Intel Xeon processors indicates that performance gains for dual-socket, Intel Xeon processor–based platforms are likely to have a significant impact on the price/performance of enterprise applications.

The Hyper-Threading Technology of the Intel Xeon processor enables a dual-socket server with two dual-core Intel Xeon processors to execute up to eight simultaneous threads in unison. This enhancement is designed to significantly improve the efficiency of applications, which allows dual-core Intel Xeon processors to help increase the performance benefits of threading in a dual-socket environment. Eventually, this capability may influence software developers to begin focusing on improvements in application threading rather than relying on faster clock speeds to enhance application performance.

Dual-socket servers are commonly deployed in database environments, where they typically run Microsoft® SQL Server™ or Oracle® database applications. For 64-bit database software such as Oracle 10*g* and Microsoft SQL Server 2005, organizations can combine the multithreading capability of the Intel Xeon processor with 64-bit memory addressing to build a powerful platform for database applications.

Comparing single-core and dual-core performance

Dell engineers in August and September 2005 tested the performance of dual-socket, dual-core PowerEdge servers and compared them against dual-socket, single-core PowerEdge servers tested in February 2005. Except where noted, the single-core Intel Xeon processors used in these benchmark tests ran at 3.6 GHz and the dual-core Intel Xeon processors ran at 2.8 GHz. Figure 2 shows the results of these benchmark tests.

On the SPECint_rate2000 benchmark, the dual-core Intel Xeon processors provided performance gains of up to 54.07 percent. This performance increase indicates that critical compute-intensive applications like HPC, technical application loads, and other integer-heavy applications may see an appreciable performance increase when compared with the fastest single-core Intel processors available (3.6 GHz at the time of testing).

On the SPECjbb2000 benchmark, up to 44.29 percent performance gain was achieved on dual-socket, dual-core PowerEdge 2800 servers. Java applications are often deployed in conjunction with databases, using a two-tier architecture. Many Web applications that run on Java are deployed on dual-socket servers because dual-socket servers are designed to provide greater reliability and scalability compared to single-socket servers. The Java performance gains demonstrated by the benchmark tests discussed in this article indicate that organizations deploying memory-intensive Java applications on Dell PowerEdge 1850, PowerEdge 1855, PowerEdge 2800, and PowerEdge 2850 servers with dual-core Intel Xeon processors are likely to obtain higher performance, better threading, and improved overall throughput compared to dual-socket, single-core servers.

The Dell team also tested the dual-socket servers on the TPC-C benchmark from the Transaction Processing Performance Council (TPC). The test results showed that the PowerEdge 2800 server with one dual-core Intel Xeon processor at 2.8 GHz achieved a 37.57 percent performance increase over the PowerEdge 2800 server with one single-core Intel Xeon processor at 3.4 GHz as well as a price/performance ratio of US$0.99/transactions per minute (tpmC), breaking the dramatic barrier of US$1/tpmC. In

[2] SPEC performance results for PowerEdge 2850 servers are based on benchmark tests performed by Dell labs on comparable configurations. Dual-core processor tests were performed in August 2005 on a PowerEdge 2850 server with two dual-core Intel Xeon processors at 2.8 GHz and 2 MB L2 cache per processor core. Single-core processor tests were performed in February 2005 on a PowerEdge 2850 with two single-core Intel Xeon processors at 3.6 GHz and 2 MB L2 cache per processor. Both systems were configured with 8 GB of error-correcting code (ECC) DDR2 SDRAM, an 800 MHz FSB, a PCI Express bus, one 36 GB SCSI drive, and Windows Server 2003, Standard Edition, with SP1. Actual performance will vary based on configuration, usage, and manufacturing variability.

★★★★

PERFORMANCE. HASSLE-FREE MULTI-CORE SERVERS.

THE DELL™ POWEREDGE™ 1850, 2800, 2850, AND THE 1855 BLADE SERVERS FEATURE DUAL-CORE INTEL® XEON™ PROCESSORS FOR OUTSTANDING PERFORMANCE.

DELL'S EASY TO DEPLOY MULTI-CORE TECHNOLOGY.

Get up to a 53% gain in performance* with Dual-Core Intel® Xeon™ Processors in Dell™ PowerEdge™ Servers. Working with your existing architecture greatly reduces the number of system images for easier deployment and management. It's the right technology at the right time.

Click www.dell.com/power25
Call (toll free) 1.877.486.DELL

***Based on the SPECint_rate2000 benchmark test performed by Dell Labs in February and July 2005 comparing a Dell PowerEdge 2850 configured with two 3.60GHz w/2MB single-core Intel Xeon Processors, 8GB DDR-2 memory, 1x36GB SCSI HDD, Windows Server 2003 Standard with the same system configured with two 2.80GHz w/2MB dual-core Intel Xeon Processors. Actual performance will vary based on configuration, usage and manufacturing variability. Results can be found at http://www.spec.org.**

Benchmark	Value system: single-core, quad processor (3.66 GHz, 2 MB cache)	Performance system: single-core, quad processor (3.33 GHz, 8 MB cache)	Dual-core, quad processor (3 GHz, 2 MB cache)	Dual-core performance increase over value system	Dual-core performance increase over performance system
SPECint_rate2000	56.4	72.2	111	97%	54%
SPECjbb2000	126,967	156,064	256,966	102%	65%

Figure 3. Benchmark performance results for quad-socket Dell PowerEdge 6850 servers[3]

fact, the Dell PowerEdge 2800 was the first system to shatter the US$1/tpmC mark in the TPC-C benchmark test.[4]

Messaging applications can also benefit from a performance increase with dual-core processors. In the MAPI (Messaging Application Programming Interface) Messaging Benchmark 3 (MMB3), which tests Microsoft Exchange performance, the dual-core, dual Intel Xeon processor–based PowerEdge 1850 server achieved an 18.75 percent performance gain over the single-core, dual Intel Xeon processor–based PowerEdge 1850 server. Although the Exchange application is limited to 4 GB of physical memory and relies heavily on disk subsystem performance, the dual-core PowerEdge 1850 server's significant performance increase can be attributed to increased threading and enhanced efficiency of dual-core processor architecture compared to single-core processor architecture.

Quad-socket, multi-core PowerEdge servers

Dell is currently developing the next generation of PowerEdge 6800 and PowerEdge 6850 servers and before year-end plans to deliver dual-core processor options and faster frontside buses (FSBs)—800 MHz compared to the 667 MHz of single-core PowerEdge quad-socket systems—in these new quad-socket servers. Currently, Dell ships two different single-core, four-processor systems: value systems with Intel Xeon processors at 3.66 GHz and 2 MB processor cache, and performance systems with Intel Xeon processors at 3.33 GHz and 8 MB processor cache. The dual-core, quad-socket system with Intel Xeon MP processors at 3 GHz is designed to outperform these single-core systems by a significant margin. In both SPECint_rate2000 and SPECjbb2000 benchmarks, the enhanced threading capabilities and 800 MHz FSB enabled the dual-core architecture to deliver outstanding performance. Applications that are generally run on such systems are highly scalable in order to take advantage of the four processors, so the introduction of dual-core technology and an 800 MHz FSB enable significant increases in overall performance, particularly for applications such as back-end databases, enterprise resource planning, and customer relationship management.

The system image for the new dual-core, four-processor servers is designed to be horizontally compatible across both the PowerEdge 6800 and PowerEdge 6850 servers as well as vertically compatible with existing single-core and dual-core processor architectures. Figure 3 compares benchmark performance of PowerEdge 6850 dual-core platforms with single-core platforms.

The next step in processor technology

Multi-core processing technology can provide significant performance increases for various types of applications across a variety of platforms. Whether employing a simple single-socket server running a Web server application, a dual-socket server running complex technical applications, or a quad-socket server supporting a large database, multi-core processor–based systems are designed to help boost performance and improve productivity. By enhancing compute efficiencies, new-generation servers with multiple processing cores and increased threading capabilities can help organizations build upon the IT framework that is already in place to keep cost of ownership low while advancing the goals of the scalable enterprise: simplified operations; improved resource utilization; and cost-effective, incremental data center growth that enables fast, flexible business response.

John Fruehe is a marketing strategist for the Dell Enterprise Product Group. He has worked at Dell for nine years. Prior to that, John was at Compaq and Zenith Data Systems. John has a B.S. in Economics from Illinois State University and has been in the technology field for 14 years.

FOR MORE INFORMATION

Dell multi-core technology:
www.dell.com/multicore

[3] Performance results are based on benchmark tests performed by Dell labs on comparable configurations. Dual-core processor tests were performed in October 2005 on a quad-socket PowerEdge 6850 with dual-core Intel Xeon MP processors at 3 GHz and 2 MB L2 cache per processor, 16 GB of ECC DDR2 SDRAM, and an 800 MHz FSB. Single-core processor tests were performed in March 2005 on a PowerEdge 6850 with four single-core Intel Xeon processors at 3.66 GHz and 2 MB L2 cache per processor and on a PowerEdge 6850 with four single-core Intel Xeon processors at 3.33 GHz and 8 MB L2 cache per processor; both single-core processor–based systems had 16 GB of ECC DDR2 SDRAM and a 667 MHz FSB. All three systems were configured with a PCI Express bus; one 36 GB SCSI drive; and Windows Server 2003, Standard Edition. Actual performance will vary based on configuration, usage, and manufacturing variability.

[4] TPC-C performance results are based on benchmark tests performed by Dell labs on comparable configurations. Dual-core processor tests were performed in September 2005 on a PowerEdge 2800 server with one dual-core Intel Xeon processor at 2.8 GHz and 2 MB L2 cache per processor core, 8 GB of ECC DDR2 SDRAM, an 800 MHz FSB, a PCI Express bus, one 36 GB SCSI drive, Windows Server 2003, and Microsoft SQL Server 2005—resulting in TPC-C performance of 38,622 tmpC, price/tpmC of US$0.99, and a system availability date of November 8, 2005. Single-core processor tests were performed in February 2005 on a PowerEdge 2800 server with one single-core Intel Xeon processor at 3.4 GHz and 2 MB L2 cache, 2.5 GB of ECC DDR2 SDRAM, an 800 MHz FSB, a PCI Express bus, one 36 GB SCSI drive, Windows Server 2003, and Microsoft SQL Server 2000—resulting in TPC-C performance of 28,122 tmpC, price/tpmC of US$1.40, and a system availability date of April 30, 2005. Actual performance will vary based on configuration, usage, and manufacturing variability. The top TPC-C price/performance results are available at www.tpc.org/tpcc/results/tpcc_price_perf_results.asp.

Creating Flexible, Highly Available SAP Solutions

Leveraging Oracle9*i* and Linux on Dell Servers and Dell/EMC Storage

Flexible, scalable, and highly available servers and storage systems are required for SAP® software implementations. IT administrators now have an alternative to the expensive, proprietary UNIX® platforms typically used for SAP environments: cost-effective, standards-based Dell™ servers and Dell/EMC storage leveraging Oracle9*i*™ Real Application Clusters and the Red Hat® Enterprise Linux® OS. The key to success is proper configuration of the hardware as well as the Linux, Oracle9*i*, and SAP software stack. Best practices for infrastructure planning, setup, and installation can help ensure that systems function as expected and enable optimal performance, easy scalability, and quick implementation.

BY DAVID DETWEILER, ACHIM LERNHARD, FLORENZ KLEY, THORSTEN STAERK, AND WOLFGANG TRENKLE

Related Categories:

Clustering

Database

Dell PowerEdge servers

Dell/EMC storage

High availability (HA)

Linux

Oracle

Red Hat Enterprise Linux

SAP

Scalable enterprise

Visit www.dell.com/powersolutions for the complete category index.

Whether upgrading existing SAP systems, migrating to the next SAP software release, or making new investments in mySAP™ technology and applications or the SAP NetWeaver® platform, IT organizations invariably face the paramount consideration of how to enhance value with limited resources. Enabling technologies and service-delivery capabilities in both software and hardware infrastructure can help organizations adapt the IT framework to the expanding boundaries set by evolving enterprise priorities.

Enterprise solutions such as the mySAP Business Suite need high availability and easy scalability—regardless of the size of the organization. Users of SAP solutions demand uninterrupted database access even if hardware or software failures occur. As businesses and organizations grow, their IT requirements can change rapidly. Adding capacity or changing workloads for SAP systems may require reconfiguring or replacing the database server to adapt to the new situation. Deploying a database cluster instead of a single, dedicated server can help maximize overall system availability in such situations. In addition, some SAP components such as the SAP central instance—which includes the Message Service and the Enqueue Replication Service—can be made redundant to enhance the robustness of an SAP environment.

Working closely with SAP, Oracle, Intel, and Red Hat, Dell has leveraged Oracle9*i* Real Application Clusters (RAC) technology to demonstrate the technical viability of running mySAP solutions in a robust, highly available, and scalable standards-based environment. Those wanting to implement

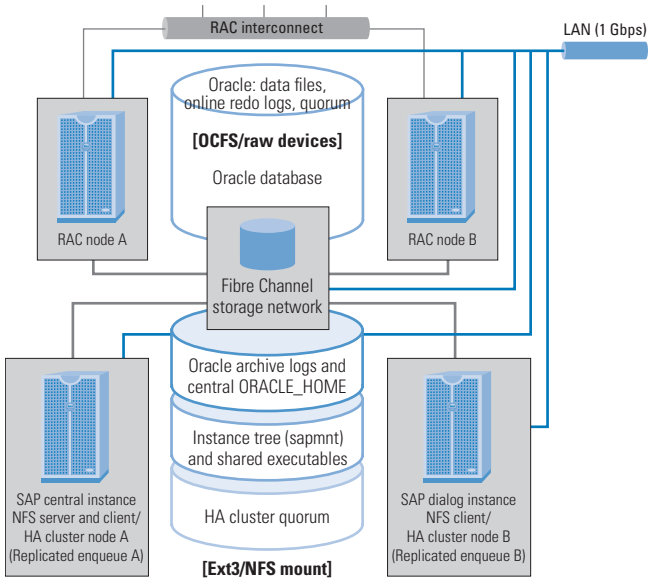


Figure 1. Basic architecture of an SAP central instance using Oracle9*i* RAC databases

an SAP solution on a Linux platform using Oracle9*i* RAC technology have various options for configuring the overall SAP infrastructure to best meet specific business objectives and budgetary requirements. Important aspects of determining the architecture include:

- Whether the SAP system will be used in a production or test/development environment
- The priority and time requirements or constraints of database availability in the overall infrastructure
- The expected growth of the user base or SAP functionality scope
- The importance of continuous SAP central instance availability in the overall SAP system landscape

Figure 1 shows the basic architecture of a fully redundant SAP system that is set up with two Oracle9*i* RAC database nodes and the highly available SAP central instance using the SAP Enqueue Replication Service. This setup can be adapted or scaled to meet various performance and deployment needs, as shown in the example configuration in Figure 2.

IT organizations have various options for possible SAP system architectures—and Dell best practices can be used to help guide installation of these architectures. However, some installation procedures discussed in this article are based on SAP and Oracle requirements for SAP systems, and may deviate from Dell-standard procedures for Oracle9*i* RAC and Red Hat Enterprise Linux installations. *Note:* All server names and IP addresses used in this article exemplify an implementation and should be changed to match an organization's specific requirements.

Setting up Dell servers and storage

When an IT organization decides to implement an Oracle9*i* RAC system on a Dell/Oracle-certified configuration, hardware setup and installation can be performed by Dell Services. This can include physically installing the servers in a rack at the organization's site, connecting all necessary cables, installing the OS (Red Hat Enterprise Linux AS 3 is used in the example scenario in this article), and initializing the shared Dell/EMC storage system.

LUN planning

Optimally, logical unit (LUN) planning should be completed before Dell Services sets up the shared storage and the LUNs, enabling the service representative to immediately set up the LUNs according to the specifically planned SAP implementation. Administrators should determine the LUN requirements according to Oracle and SAP

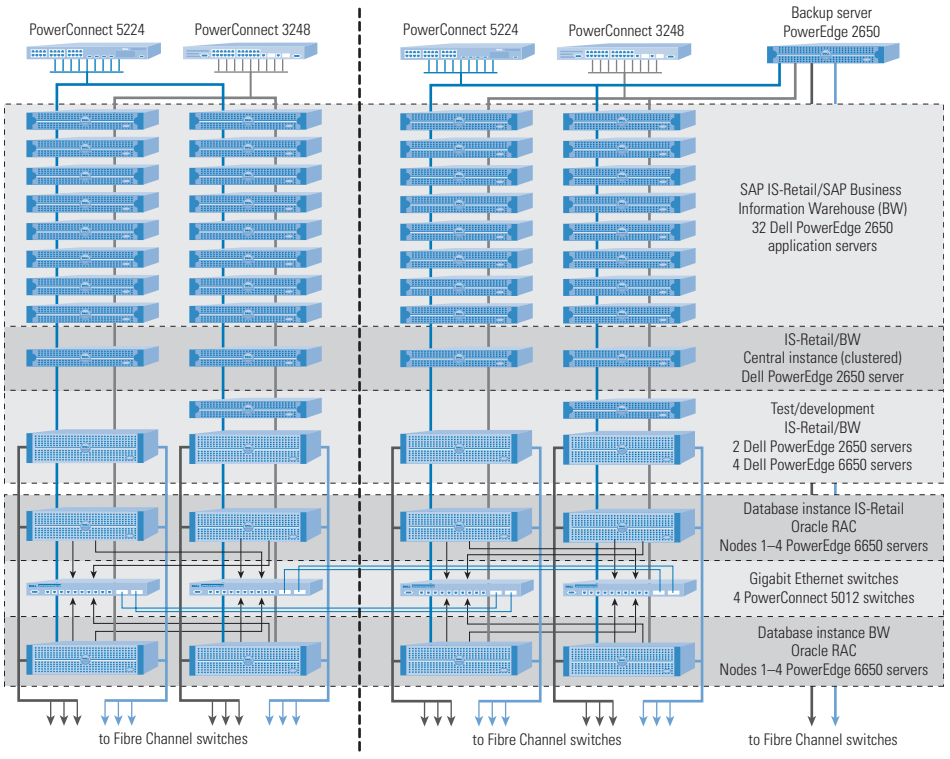


Figure 2. Example deployment of an SAP central instance using Oracle9*i* RAC databases

recommendations for a single-instance Oracle9*i* database for an SAP system. They should then add the LUNs required for RAC communication and the second database node, as well as the LUNs for the Highly Available Network File System (HA NFS) service.

The following LUNs are required:

- A set of LUNs for a single-node Oracle9*i* database for SAP
- A set of LUNs for the redo thread of each additional database instance (online redo logs and the undo/rollback tablespace)
- Two (small) LUNs for the quorum of the Red Hat Enterprise Linux cluster
- A LUN for the shared executables ($ORACLE_HOME and SAP executables)

The LUNs should be organized into two separate storage groups governing access. One storage group should contain all database-relevant LUNs; the other storage group should contain all LUNs belonging to the Red Hat Enterprise Linux cluster.

If installing the database on raw devices and not on OCFS, administrators should be aware that what can be a file in an OCFS implementation must be configured on a separate raw device, and thus must be on a separate LUN or partition:

- A small LUN for the RAC cluster quorum (to ensure exclusive access on the SCSI device level, this must be a separate LUN and must not be a partition on a LUN shared with other partitions)
- A small LUN for the shared configuration file (this can be one of several partitions on a LUN)
- Three or more small LUNs for the database control files (these can be distributed among several partitions on a LUN)

Visual check of the hardware

To verify that systems are wired correctly, administrators should start by visually checking the overall system. The Dell/EMC storage system, the two Dell PowerEdge™ servers for the database cluster, and the two PowerEdge servers for the SAP central instance cluster should each be connected to the Fibre Channel switch. Additional SAP application servers may not need storage area network (SAN) connectivity.

The primary network card in each server and the network interface of the Dell/EMC storage system should be connected to the data center LAN, over which these servers will be accessed from the outside. The network should satisfy SAP requirements for the connection of SAP database servers and SAP application servers.

The secondary interface of the database and the SAP central instance cluster servers should be connected to a private LAN, over which these systems connect only to the other nodes in the respective cluster. In the case of a two-node Oracle9*i* RAC configuration with a two-node SAP central instance cluster, each secondary interface can be connected with a cross-wired Ethernet cable (also referred to as a cross-connect or X-connect) to its clustered counterpart. Alternatively, administrators can connect the clustered servers over a dedicated Gigabit Ethernet switch (a requirement in Oracle9*i* RAC configurations with more than two database nodes). The only fabric currently supported for the database interconnect is Gigabit Ethernet.[1]

ONLINE EXTRA: LOGIN AND MANAGEMENT TOOLS FOR SAP ON ORACLE/LINUX PLATFORMS

To learn more about remote login capabilities, the Linux desktop, and versioning tools for SAP on Oracle/Linux platforms, visit *Dell Power Solutions* online at www.dell.com/powersolutions. This special supplement to the print article also provides example LUN configurations for Oracle9*i* RAC database nodes and high-availability NFS clusters in an SAP environment.

Installing the Linux, Oracle9*i* RAC, and SAP software

When installing Red Hat Enterprise Linux, Oracle9*i*, and SAP software, administrators should heed the following recommendations:

- Install a valid hardware/software combination that is certified by Dell, Red Hat, Oracle, and SAP. Dell-certified hardware for SAP and Linux can be found in SAP Note 300900 or at www.dell.com/sap. Because SAP specifically requires running only SAP-certified Linux kernel and glibc library versions, administrators should make sure to use only these versions and to exclude the Linux kernel and glibc library from software that automatically installs other versions of these packages. The latest information on certified Linux and glibc versions is available at www.sap.com/linux. The latest information on platform availability for Oracle9*i* RAC can be found in SAP Note 527843 (a username and password for the SAP Service Marketplace is required).
- Unless otherwise noted, use the original installation media provided by Red Hat, Oracle, and SAP to help ensure uniformity. *Note:* The Dell Deployment CD for Oracle9*i* RAC can be used to install Red Hat Enterprise Linux and Oracle on Dell PowerEdge servers. However, this CD currently does not support specific SAP requirements on RAC setups for SAP and therefore should not be used for RAC installations in SAP environments.
- Plan and configure storage systems carefully, optimally with the help of a Dell Services representative. Storage plays a

[1] This term does not connote an actual operating speed of 1 Gbps. For high-speed transmission, connection to a Gigabit Ethernet server and network infrastructure is required.

vital role in the success of any SAP project—particularly for implementations that include Oracle9*i* RAC technology.
- Use standard SAP installation procedures and utilities for Oracle9*i*. A single-instance database and the SAP software should be installed and then additional nodes can be added, activating the RAC feature of the database.
- Pay careful attention to the installation and the high-availability aspects of the implementation when configuring Oracle9*i* RAC for SAP software.

Red Hat Enterprise Linux 3

Before installing Red Hat Enterprise Linux 3, administrators should upgrade the Fibre Channel adapter firmware to the most recent version listed in the EMC compatibility matrix (www.emc.com/interoperability/index.jsp). During the installation, administrators should disconnect the Fibre Channel cables from the host bus adapters (HBAs) and set up access to the shared storage with the OS in place. Administrators should install the Red Hat Enterprise Linux OS on each Dell PowerEdge server according to the documentation provided with the server and the Red Hat installation guide. Best practices recommend using English as the default language.

Additional recommendations include the following:

- Partition the hard drives manually to properly reserve space for the mySAP application system data.
- Select "No firewall" when asked in the dialog box to configure the firewall. If a firewall is required, configure it manually after the system installation. In this case, ensure that the required ports on the mySAP servers are available from the other servers. The used ports on the mySAP servers can be ascertained after the installation in /etc/services.
- Select the time zone to be Coordinated Universal Time (UTC) and check the "System clock uses UTC" box. Ensure that the system's BIOS clock is set to UTC.
- Select "Customize the set of packages to be installed" in the Package Default dialog box. In the Package Group Selection dialog box, select the following as a minimum (additional items can be chosen): X Window system, Editors, Development Tools, Legacy Software Development, Administration Tools, and Printing Support.

After the first system boot, a configuration tool should appear. If a Network Time Protocol (NTP) Time Server is available at the site, Dell best practices recommend using this NTP server as a common time source for all servers in the SAP environment. Administrators should select "Enable Network Time Protocol," and under "Date and Time" they should enter the address of the time server.

The system should be registered with Red Hat Network so that update packages are automatically made available. To exclude the Linux kernel and important libraries from this auto-update, administrators should issue the `up2date --configure` command and configure the "Package Exceptions" to exclude the following: kernel*, glibc*, nscd, and nptl-devel.

Post-installation tasks for SAP environments

Once the initial installation and configuration on each server has been completed, administrators must conduct various tasks before proceeding with the Oracle and SAP installation. First, the Linux kernel parameters must be changed to meet the requirements for SAP software. The following lines should be added to the /etc/sysctl.conf file (or those already there should be edited to match the following):

```
#SAP settings
kernel.shmmax=2313682943
kernel.msgmni=1024
kernel.sem=1250 256000 100 1024
fs.file-max=8162
```

Once the preceding lines have been added, administrators should activate these settings with the `sysctl -p` command.

A temporary file system (tmpfs) needs to be configured on each SAP system. Space for the tmpfs is allocated in the system's virtual memory, which is composed of the system's physical memory (RAM) and the configured swap space. Tmpfs holds the SAP system's shared memory. The default setting for tmpfs is generally half of the total size of system memory. This setting can be changed in the /etc/fstab:

```
tmpfs   /dev/shm   tmpfs   size=size 0 0
```

To determine the proper size of tmpfs, administrators must consider the amount of virtual memory available and the memory requirements of the SAP system itself. If the system has 4 GB of RAM and 8 GB of configured swap space, the tmpfs can be limited to 6 GB. Administrators would then edit the line in /etc/fstab so that *size* is "6G" or "6144M":

```
tmpfs   /dev/shm   tmpfs   size=6G 0 0
```

However, administrators should note the following:

- Tmpfs must always be smaller than the amount of virtual memory available.
- Tmpfs should not be larger than twice the system's physical memory.

To avoid a situation in which swap partitions cannot be found, administrators should check that the order of the HBAs in each

system is correct. After the initial installation, they should check in the /etc/modules.conf file to help ensure that the SCSI RAID controller for the internal hard drives comes before the Fibre Channel HBA. For example, on a Dell PowerEdge 2650 server, the /etc/modules.conf file should look similar to the following:

```
alias scsi_hostadapter aacraid
alias scsi_hostadapter1 qla2300
```

If this is not the case, administrators should change the file and run `mkinitrd` to help ensure the order is kept. If using the kernel 2.4.21-15.ELsmp, administrators should issue the following command:

```
mkinitrd initrd-2.4.21-15.ELsmp.img
   2.4.21-15.ELsmp -f
```

Alternatively, administrators can use the device file system (devfs) or devlabel.[2]

Checking the network connections

Administrators must ensure that all servers are properly connected and integrated into the network. The servers each require a public IP address, and name resolving must function among the servers through the use of either host files on all participating systems or Domain Name System (DNS). Because reliance exclusively on host files is waning, best practices recommend using DNS. However, to prevent DNS name resolving from being a point of failure or a source of instability in the cluster, administrators should set up host files on all servers in the database cluster and the SAP central instance cluster, as well as on additional SAP application servers, to resolve names. Also, administrators should have procedures in place to keep those hosts files current—for example, after changing the DNS records.

Additionally, all database nodes must have an internal IP address for the database interconnect. This internal IP address should be reachable and resolvable from the database nodes exclusively, and does not have to be connected to the enterprise's network (see the "Visual check of the hardware" section in this article for information about cable connectivity).

Next, administrators should check the host names of the servers and test whether they can be reached over the network and whether name resolving works. Administrators should issue the `hostname` command, which should return the node name only—not the fully-qualified domain name (FQDN). Although the definition of a host name can vary—for example, whether it should include the domain name—for Oracle and SAP applications, the servers must be set up according to SAP note 722273 ("Red Hat Enterprise Linux 3 and 4: Installation and upgrade"):

```
# hosts for acme.com
127.0.0.1     localhost.localdomain  localhost
172.16.42.31  snoopy.acme.com        snoopy
```

Administrators should check and, if necessary, adapt the /etc/hosts file on each server. The entry for "localhost" must not contain an alias for the host name. Administrators should decide which name-resolving strategy will be used and test the name-resolving capability. They should then perform a similar test for connectivity with the `ping` command. If DNS is used, administrators should test both DNS and file resolution with the appropriate tools, such as `dig` and `getent`.

Checking the Dell/EMC shared storage

For Dell/EMC Fibre Channel storage, administrators should verify that the EMC® Navisphere® software agent is installed on each node. This installation is usually performed by Dell Services during the storage setup. Using the Red Hat Package Manager (RPM™), administrators can verify whether the Navisphere agent is already installed:

```
[root@rac1 root]# rpm -q naviagent
```

The expected response should look like the following:

```
naviagent-6.6.0.3.8-1
```

Version numbers may differ because software is upgraded over time.

Administrators should check whether LUNs have been created according to Oracle and SAP requirements (as discussed in the "LUN planning" section in this article). They also should check that each node is assigned to the correct storage group in the EMC Navisphere software; more information about this can be found in the documentation that accompanies the Dell/EMC Fibre Channel storage system. If not using EMC PowerPath software, administrators should use devlabel to create persistent device names.

Next, administrators should check in /proc/partitions for the shared storage LUNs. A list of the LUNs or logical disks that are detected by the node is displayed, as well as a list of the partitions that have been created on the external devices. Using this information, administrators can create partitions for OCFS volumes or raw

[2] For more information about devlabel, visit linux.dell.com/devlabel or see "Resolving Device Renaming Issues in Linux" by Gary Lerhaupt, *Dell Power Solutions*, February 2003, www1.us.dell.com/content/topics/global.aspx/power/en/ps1q03_lerhaupt.

devices. Administrators should ensure that the same shared disks are visible from all nodes.[3]

The listed devices vary depending on how the storage is configured. The primary SCSI drive or RAID array on each node will be listed as sda and will be partitioned. If any other SCSI disks or RAID arrays are on the node, they will be listed as sdb, sdc, and so on. The LUNs on the Fibre Channel storage system or SCSI enclosure should also be visible as SCSI devices. For example, if the node has one RAID array and the storage system has three logical disks, the node should identify the RAID array or internal disk as sda and the logical disks as sdb, sdc, and sdd. If three LUNs are on the Fibre Channel storage, the node should recognize the RAID array as sda and the Fibre Channel LUNs as emcpowera, emcpowerb, and emcpowerc (assuming EMC PowerPath® software is being used).

If a node does not detect the external storage devices, administrators should perform the following steps:

1. Reload the HBA driver on all nodes to synchronize the kernel's partition tables on the nodes by entering (for QLogic HBAs):

```
rmmod qla2300
modprobe qla2300
```

2. Confirm that all nodes detect the external storage devices by entering:

```
cat /proc/partitions
```

Partitions on shared storage

Even if creating partitions and file systems on shared storage from any of the attached servers is possible, best practices recommend limiting these administrative tasks to only one of the nodes, usually the first node. Because no database or shared file system is running at this stage of the installation, the nodes do not yet have access to the shared storage and consequently do not need to be disconnected from it. However, once a RAC database, an HA NFS server, or a similar service accesses the shared storage from multiple nodes, all other nodes must stop I/O operations and cede control of the storage device on a software level, so that the management node can alter the shared device configuration. Usually, ceding control is accomplished by unloading the Fibre Channel driver module, which is possible only when the database, cluster manager, or a similar service is shut down. Taking into account these organizational measures, administrators should continue creating partitions on the management node as they would on locally attached disks.

Partitions on local disks in Oracle9*i* RAC database nodes

Administrators should create partitions in the local drives on each of the RAC database nodes for the Oracle database software. On the first server, node A, administrators should perform the following steps:

1. Create one partition of 10,000 MB and a partition for the remainder of the disk space using the following command:

```
fdisk /dev/sda
```

2. Reboot the server.
3. After the reboot, format the partitions:

```
mke2fs -j /dev/sda7
mke2fs -j /dev/sda8
```

4. Verify that the entries in the file /etc/fstab are as follows:

```
/dev/sda7   /oracle   ext3   defaults   1 2
/dev/sda8   /sapcd    ext3   defaults   1 2
```

5. Create the /oracle directory:

```
mkdir /oracle
```

6. Create the /sapcd directory:

```
mkdir /sapcd
```

7. Mount the directories according to /etc/fstab:

```
mount -a
```

On the second server, node B, administrators should perform the following steps:

1. Create one partition of 10,000 MB and a partition for the remainder of the disk space using the following command:

```
fdisk /dev/sda
```

2. Reboot the server.
3. After the reboot, format the partition:

```
mke2fs -j /dev/sda6
```

4. Verify that the entries in the file /etc/fstab are as follows:

```
/dev/sda6   /oracle   ext3   defaults   1 2
/dev/sda8   /sapcd    ext3   defaults   1 2
```

[3] For more information about how to ensure that all machines see the same devices under the same name, see "Resolving Device Renaming Issues in Linux" by Gary Lerhaupt, *Dell Power Solutions,* February 2003, www1.us.dell.com/content/topics/global.aspx/power/en/ps1q03_lerhaupt.

IMPORT PEERS.*;
IMPORT EXPERTS.*;
IMPORT YOU.*;

// EXPORT DEADLOCK
// GO TO SDN.SAP.COM

You're stuck. You need answers. Maybe you have a solution to share with other SAP developers or a question for an SAP insider. Get the experts, partners and your colleagues to weigh in. Now there's a single collaborative destination where you can all converge: SAP® Developer Network. Nowhere else can you download sample code, join a spirited discussion forum, take an e-learning course, test-drive new SAP technology and, in general, keep us on our toes.

// JOIN IN AT SDN.SAP.COM

5. Create the /oracle directory:

```
mkdir /oracle
```

6. Mount the directory according to /etc/fstab:

```
mount -a
```

Installing OCFS on each RAC database node

Administrators should place Oracle tablespaces and log files on OCFS. To install OCFS on Dell/EMC storage systems, the following packages are required:

- coreutils-debuginfo-4.5.3-35.i386.rpm
- coreutils-4.5.3-35.i386.rpm
- ocfs-2.4.21-EL-smp-1.0.12-1.i686.rpm
- ocfs-support-1.0.10-1.i386.rpm
- ocfs-tools-1.0.10-1.i386.rpm
- ocfs-tools-debuginfo-1.0.10-1.i386.rpm
- tar-debuginfo-1.13.25-16.i386.rpm
- tar-1.13.25-16.i386.rpm

The following packages should be installed with the `-ivh` option:

- coreutils-debuginfo-4.5.3-35.i386.rpm
- ocfs-2.4.21-EL-smp-1.0.12-1.i686.rpm
- ocfs-support-1.0.10-1.i386.rpm
- ocfs-tools-1.0.10-1.i386.rpm
- ocfs-tools-debuginfo-1.0.10-1.i386.rpm
- tar-debuginfo-1.13.25-16.i386.rpm

Packages that should be installed using the `-Uvh` option include coreutils-4.5.3-35.i386.rpm and tar-1.13.25-16.i386.rpm. These two packages contain tools that are OCFS aware; the standard coreutils and tar packages should not be used on OCFS file systems. Once the packages have been installed on both node A and node B, the ocfs.conf file must be edited. On node A, the file should look similar to the following:

```
node_name=nodeA
node_number=0
ip_address=1.0.102.65
ip_port=7000
```

Then, administrators should issue the command `ocfs_uid_gen -c`. This generates a unique node identifier from the information in /etc/ocfs.conf and appends that information to the file. On node B, the ocfs.conf file should look similar to the following:

```
node_name=nodeB
node_number=1
ip_address=1.0.102.66
ip_port=7000
```

As on the first node, administrators should issue the command `ocfs_uid_gen -c` on node B. Next, administrators should enter the `load_ocfs` command on both nodes. At this point, administrators can verify on the system log whether OCFS loaded successfully. Once that is completed, the OCFS partitions must be formatted. Figure 3 shows example formatting commands (SID is the SAP system ID).

```
mkfs.ocfs -b 128 -L saparch -m /oracle/SID/saparch -u 1046 -g 502 /dev/sdc1
mkfs.ocfs -b 128 -L sapdata1 -m /oracle/SID/sapdata1 -u 1046 -g 502 /dev/sdd1
mkfs.ocfs -b 128 -L sapdata2 -m /oracle/SID/sapdata2 -u 1046 -g 502 /dev/sde1
mkfs.ocfs -b 128 -L sapdata3 -m /oracle/SID/sapdata3 -u 1046 -g 502 /dev/sdf1
mkfs.ocfs -b 128 -L sapdata4 -m /oracle/SID/sapdata4 -u 1046 -g 502 /dev/sdg1
mkfs.ocfs -b 128 -L sapdata5 -m /oracle/SID/sapdata5 -u 1046 -g 502 /dev/sdh1
mkfs.ocfs -b 128 -L sapdata6 -m /oracle/SID/sapdata6 -u 1046 -g 502 /dev/sdi1
mkfs.ocfs -b 128 -L origlogA -m /oracle/SID/origlogA -u 1046 -g 502 /dev/sdj1
mkfs.ocfs -b 128 -L origlogB -m /oracle/SID/origlogB -u 1046 -g 502 /dev/sdk1
mkfs.ocfs -b 128 -L mirrlogA -m /oracle/SID/mirrlogA -u 1046 -g 502 /dev/sdl1
mkfs.ocfs -b 128 -L mirrlogB -m /oracle/SID/mirrlogB -u 1046 -g 502 /dev/sdm1
mkfs.ocfs -b 128 -L clusterquorum -m /clusterquorum -u 1046 -g 502 /dev/sdn1
mkfs.ocfs -b 128 -L racquorum -m /racquorum -u 1046 -g 502 /dev/sdo1
```

Figure 3. Commands for formatting OCFS partitions

The next step is to create on all nodes the matching directories as mount points for OCFS:

- /oracle/SID/sapdata1
- /oracle/SID/sapdata2
- /oracle/SID/sapdata3
- /oracle/SID/sapdata4
- /oracle/SID/sapdata5
- /oracle/SID/sapdata6
- /oracle/SID/saparch
- /oracle/SID/origlogA
- /oracle/SID/origlogB
- /oracle/SID/mirrlogA
- /oracle/SID/mirrlogB
- /racquorum

```
LABEL=sapdata1        /oracle/SID/sapdata1   ocfs   _netdev   0 0
LABEL=sapdata2        /oracle/SID/sapdata2   ocfs   _netdev   0 0
LABEL=sapdata3        /oracle/SID/sapdata3   ocfs   _netdev   0 0
LABEL=sapdata4        /oracle/SID/sapdata4   ocfs   _netdev   0 0
LABEL=sapdata5        /oracle/SID/sapdata5   ocfs   _netdev   0 0
LABEL=sapdata6        /oracle/SID/sapdata6   ocfs   _netdev   0 0
LABEL=saparch         /oracle/SID/saparch    ocfs   _netdev   0 0
LABEL=origlogA        /oracle/SID/origlogA   ocfs   _netdev   0 0
LABEL=origlogB        /oracle/SID/origlogB   ocfs   _netdev   0 0
LABEL=mirrlogA        /oracle/SID/mirrlogA   ocfs   _netdev   0 0
LABEL=mirrlogB        /oracle/SID/mirrlogB   ocfs   _netdev   0 0
LABEL=racquorum       /racquorum             ocfs   _netdev   0 0
```

Figure 4. Storage devices in the /etc/fstab directory

Figure 4 shows the entries in the file /etc/fstab. The option `_netdev` is designed to ensure that, during system startup, the OCFS volumes are mounted after the network becomes active, enabling OCFS cluster communication. These devices should be activated with the `mount -a` command. When all the devices on node A have been mounted, administrators should edit the /etc/fstab directory as shown in Figure 4 and issue the `mount -a` command. The ownership for these directories must be changed to the database and user group:

```
chown -R ora<sid>:dba /oracle
```

By completing the steps described in this article, administrators can help ensure that the shared storage for both Oracle9*i* RAC nodes is ready to function as expected.

Planning a highly available, flexible environment for SAP software

SAP software implementations require a robust IT infrastructure to support business-critical processes. Cost-effective, standards-based Dell servers and Dell/EMC storage leveraging Oracle9*i* RAC databases and the Red Hat Enterprise Linux OS can provide a highly available, flexible platform for SAP environments. However, proper planning and configuration—based on best-practices recommendations from Dell, Oracle, Red Hat, and SAP—are key to helping ensure that all systems function optimally.

David Detweiler is the Dell SAP Alliance Manager in Europe, the Middle East, and Africa (EMEA) and a member of the Dell SAP Competence Center in Walldorf, Germany. The Dell SAP Competence Centers help ensure that current and future Dell technologies work together with SAP solutions and provide customers with the architecture, functionality, reliability, and support expected of mission-critical applications.

Achim Lernhard has worked at the Dell SAP Competence Center in Walldorf, Germany, for three years as part of the SAP LinuxLab. He assisted the Oracle9*i* RAC on Linux pilot customer from installation to productivity and worked on the hardware certifications.

Florenz Kley is a consultant for SAP Technology Infrastructure. He has worked for five years at the Dell SAP Competence Center in Walldorf, Germany, as part of the SAP LinuxLab. He conducted performance benchmarks to help prove the scalability and performance of Oracle9*i* RAC for SAP on Linux and helped build the architecture for Dell's Oracle9*i* RAC on Linux pilot customer.

Thorsten Staerk is a consultant at the Dell SAP Competence Center in Walldorf, Germany, as part of the SAP LinuxLab. He has worked extensively on Oracle9*i* RAC technologies for SAP, researches new SAP technologies and functionality, and certifies Dell platforms for SAP on Linux.

Wolfgang Trenkle is a senior consultant at the Dell SAP Competence Center in Walldorf, Germany, and is also a member of the Dell EMEA Enterprise Solutions Center team in Limerick, Ireland. In addition to serving as a consultant and supporting proof of concepts, Wolfgang provides training materials and tools to Dell's global SAP community.

FOR MORE INFORMATION

Red Hat Enterprise Linux 3:
www.redhat.com/docs/manuals/enterprise
www.redhat.com/docs/errata

SAP LinuxLab:
www.sap.com/linux

Dell and Linux:
linux.dell.com/projects.shtml
linux.dell.com/projects.shtml#devlabel

Optimizing SQL Server 2005 Environments for Resiliency

Using VERITAS Storage Foundation HA *for Windows* from Symantec

Improving systems availability, scalability, management, and performance are key goals for every size enterprise. The VERITAS® Storage Foundation™ HA *for Windows* suite from Symantec helps meet these goals by providing common installation, configuration, and management tools to help optimize Microsoft® SQL Server™ database environments. The Symantec® suite is designed to enhance storage and application configurations while enabling online disaster recovery validation and testing.

BY KEVIN KNIGHT

Related Categories:

Database

High availability (HA)

Microsoft SQL Server

Microsoft SQL Server 2005

Server consolidation

Storage consolidation

Storage software

Symantec

Visit www.dell.com/powersolutions for the complete category index.

Enterprises must contend with ever-increasing business challenges, regulations, and scrutiny. At the same time, customers expect information to be available at all times, and even a few minutes of downtime can be costly. Meeting stringent service-level agreements can be difficult without careful planning and wise IT investments.

Traditionally, IT administrators have protected servers and applications by backing up data to tape. Although regular tape backups are essential, recovery-time objectives and recovery-point objectives cannot always be met by tape-based solutions alone. Today, Microsoft SQL Server administrators require a flexible, integrated backup system that helps reduce the time it takes to recover from data corruption or data center failures from hours or days to minutes. While unplanned outages are often unavoidable, administrators can enhance preparations for planned outages such as maintenance, patching, and disaster recovery validation and testing.

This article explains how VERITAS Storage Foundation HA *for Windows* from Symantec configured on Dell™ PowerEdge™ servers can help IT administrators achieve high levels of data and application availability and enhance Microsoft SQL Server database performance. VERITAS Storage Foundation HA *for Windows* is designed to support the Standard Edition and the Enterprise Edition of SQL Server, all versions of the Microsoft Windows® 2000 Server and Windows Server™ 2003 operating systems including 64-bit versions, and virtual environments. This article also discusses how testing and validating a disaster recovery plan can help administrators ensure business continuity by using Storage Foundation HA *for Windows* to perform a "fire drill" that is designed to validate the production environment without affecting online applications.

Designing a Microsoft SQL Server 2005 solution

With multiple versions available to suit any size enterprise, Microsoft SQL Server is fast becoming a database of choice for organizations around the world. VERITAS Storage Foundation HA *for Windows* helps extend the core benefits of logical disk management that the Microsoft Windows OS provides. Additional features and capabilities

can be enabled in Storage Foundation HA *for Windows* simply by adding the appropriate license key.

Before installing SQL Server, administrators should consider how databases will be placed on the underlying storage and how that storage can be configured both for performance and for fault tolerance. Microsoft provides specific recommendations and prescriptive architectures for many types of deployments; more information can be found at www.microsoft.com/sql/2005.

In addition to database storage and performance issues, administrators should consider the importance of clustering and fault tolerance. Will more systems and storage paths be required? Primary considerations for determining SQL Server architecture include the following:

In addition to database storage and performance issues, administrators should consider the importance of clustering and fault tolerance.

- What are the uptime and performance requirements for this application?
- How much data will reside on the server and how long will it take to back up and restore the data?
- After a failure, can the application be restored to production using tape backup alone in the time required?
- Could a second data center also host this application?
- If a second data center is available, what type of connectivity exists between sites?

Optimizing the storage configuration for SQL Server 2005

The example scenario in this article explains how a SQL Server system can be attached to a storage area network (SAN). In such an environment, administrators can use the Dynamic MultiPathing option available in VERITAS Storage Foundation HA *for Windows*. This option is designed to provide fault tolerance for the Fibre Channel path between Dell servers and frequently used storage arrays, whether that path involves multiple Dell/EMC storage arrays or a heterogeneous storage environment. If a path to the array fails for any reason, this option is designed to help ensure that SQL Server will continue running without interruption.

Administrators typically need to create logical units (LUNs) on the disk array once they have decided how to configure the underlying storage. In some cases, this decision will be made by the storage administrator or the array vendor. Administrators must determine the required size and the quantity of LUNs as well the RAID configuration. They must also consider how to separate the system databases from the user database volumes and where to place the logs and tempdb database. New features in SQL Server 2005 can make heavy use of the tempdb database, which may create a bottleneck if SQL Server 2005 is not properly configured for the specific application. In addition, administrators should decide whether to create disk-based snapshots of the databases for quick recovery from accidental or malicious changes or from data corruption. Snapshot volumes can be created when configuring volumes for the application or at a later time when required and can be placed on a less-expensive storage tier than is required for high-performance applications. Figure 1 shows the typical disk layout for a simple SQL Server configuration based on the best practices just described.

Following best practices for designing the SQL Server deployment, administrators can then take advantage of dynamic disk features. Dynamic disks allow the grouping of disk resources based on how those resources will be used, clustered, captured (as a snapshot), backed up, and so forth. In the base OS, only one disk group can be created, which prevents the use of advanced dynamic disk features such as clustering or off-host processing and backup. In the typical disk layout described in Figure 1, administrators can easily add volumes as necessary if data or logs grow larger than originally configured. Storage Foundation HA *for Windows* enables administrators to set automatic or manual volume growth, both of which can take place while SQL Server remains online.

Additionally, if certain data volumes experience performance problems, they can be moved to other locations to help reduce contention while SQL Server remains online. In Dynamic Disk Group 2 of the Figure 1 layout, two additional LUNs are used to enable quick recovery—one for storing snapshots of the database and one for storing snapshots of the logs. Administrators can perform such recoveries by using the snapshots to restore the desired volume(s) and rolling the logs forward if required (as in a database restore from backup). This approach enables recovery from a snapshot in minutes or even seconds, rather than in the hours typically required for tape-based recovery. Also, because the VERITAS FlashSnap™ capability is designed to make a complete copy of all data, the original data can be recovered to the point of the most current copy. Recovery can be performed from any node in a cluster, and logs can be used to recover to a specific point in time. In addition, on a shared SAN such as in

Location	Disk group	Volume	Drive letter	Drive contents
Local storage	Basic Disk Group	Volume 1	C:	OS and SQL Server installation
SAN	Dynamic Disk Group 1 (minimum of 2 LUNs)	Volume 1	S:	System databases
		Volume 2	T:	Tempdb database
SAN	Dynamic Disk Group 2 (minimum of 4 LUNs)	Volume 1	U:	User database
		Volume 2	L:	User database logs
		Volume 3	N/A	Snapshot of database
		Volume 4	N/A	Snapshot of logs

Figure 1. Typical SQL Server disk layout

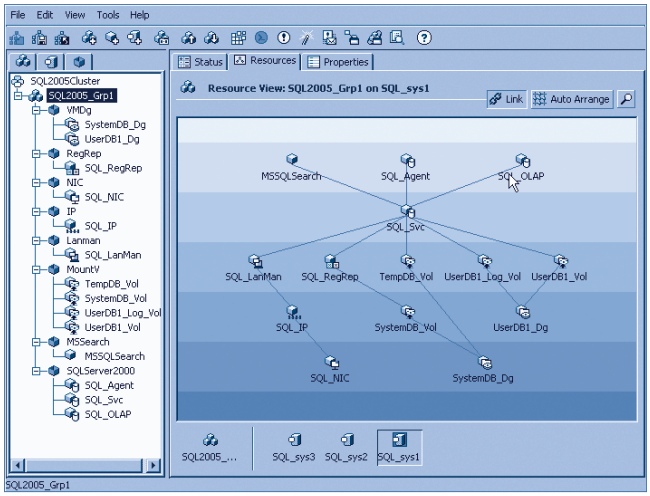


Figure 2. VERITAS Cluster Server console showing cluster resource dependency for SQL Server 2005

a cluster environment, dynamic disk groups are designed to protect disks from being accessed by other systems on the SAN.

Clustering with VERITAS Storage Foundation HA *for Windows*

Depending on the availability requirements of an application, administrators may also need to cluster SQL Server as well as any critical application that stores its data and depends on the SQL Server system. VERITAS Cluster Server (VCS) enables scalable, policy-based failover for up to 32 nodes in a single cluster and is designed to control startup and shutdown for multitier applications. Any-to-any failover is also supported.

Because SQL Server 2005 is designed to scale up to 50 instances per server (SQL Server 2000 scaled up to only 16 instances), policy-based failover and workload management can help ensure that critical applications get the appropriate server resources. In addition, this approach can enhance return on IT investments by helping to optimize server utilization. Figure 2 shows the VCS console in which cluster resource dependency for SQL Server 2005 is displayed.

Geographically dispersed clusters for remote-site disaster recovery

When multiple data centers are involved, administrators may need to plan for site failure and recovery in a remote data center. For up to four connected cluster sites, VCS is designed to seamlessly fail over SQL Server resources from one node to another—whether at the local site or across the globe. The VERITAS Volume Replicator provides synchronous or asynchronous block-level replication over IP to move data from one site to another when the distance is too far for standard Fibre Channel to reach. VCS can manage this replication as well as many types of hardware-based replication while providing simple site migration or recovery from a central display, with the click of a mouse. For optimal performance when replicating SQL Server, the tempdb database should not be replicated to the remote cluster.

Testing and validation

Testing and validation are key requirements for most organizations today. Many enterprises are now required to perform complete "lights-out" tests of their production data centers—in which a data center is actually shut down and recovery time is measured to help ensure that recovery procedures are sufficient—as frequently as once per quarter. VCS provides a "fire drill" feature, which is designed to clone the production cluster and disk configuration and bring the clone online for validation on a standby server—without affecting online production applications and servers. By performing such tests regularly, administrators can help ensure that downtime will be kept to a minimum when a real outage occurs.

Enhancing SQL Server environments with VERITAS Storage Foundation HA *for Windows*

Many factors, planned and unplanned, affect information availability. By implementing VERITAS Storage Foundation HA *for Windows*, organizations can benefit from a tool that is designed to build and optimize a resilient SQL Server system. Dell and Symantec are working together to enable a complete set of fully tested, integrated hardware and software solutions that are designed to provide high availability for Microsoft SQL Server 2005 environments. This scalable, easy-to-manage approach—which includes VERITAS Cluster Server, the key component of VERITAS Storage Foundation HA *for Windows*—is well suited for medium- to large-size enterprises implementing SQL Server in their data centers. On-site planning and assistance are also available through Symantec professional services.

Kevin Knight is a senior product manager at Symantec Corporation, where he focuses primarily on availability and disaster recovery solutions for enterprise applications on the Windows platform. Kevin frequently speaks at conferences as a subject-matter expert on Microsoft SQL Server and Microsoft Exchange Server.

FOR MORE INFORMATION

Symantec:
www.symantec.com

VERITAS Storage Foundation *for Windows* from Symantec:
www.veritas.com/Products/van?c=product&refId=31

Microsoft SQL Server 2005:
www.microsoft.com/sql/2005

Dell and Symantec:
www.dell.com/symantec

Maximizing SQL Server Performance

Using Symantec Indepth for SQL Server

Symantec® Indepth™ software for SQL Server is designed to keep Microsoft® SQL Server™ databases and related applications operating at peak performance levels. This software enhances the management of application performance by proactively monitoring, analyzing, and tuning SQL Server databases. In this way, Symantec Indepth for SQL Server helps database administrators optimize CPU utilization and correct potential performance degradation problems before they affect the response of business-critical applications.

BY RON GIDRON

Related Categories:

Application servers

Database

Microsoft SQL Server

Performance

Performance management

Symantec

Visit www.dell.com/powersolutions for the complete category index.

Large or small, organizations that use Microsoft SQL Server software all too often encounter performance problems. Such problems can be difficult to diagnose and costly to fix. Furthermore, requirements for increased productivity call upon many database administrators to manage a growing number of instances and databases—leaving them with little time to address performance problems that can affect business response.

In a resilient computing infrastructure, IT is aligned with business goals and responsive to changing business demands. Symantec Indepth for SQL Server enhances the efficiency of application performance management by providing a comprehensive view of application performance—capturing, measuring, and correlating performance metrics from critical system components. As a result, Indepth for SQL Server can help database administrators maximize the performance of SQL Server environments and optimize CPU resource utilization, enabling IT organizations to meet critical business needs. This article explains the features and benefits of Indepth for SQL Server and how this software can enhance the efficiency of the application management process.

Control multiple SQL Server instances from a single console

Symantec Indepth for SQL Server features a multiple-instance dashboard that is designed to provide administrators with a view of the performance and availability of all SQL Server instances and databases from a single screen. The dashboard helps administrators identify relevant performance information across a SQL Server network in a matter of seconds, including where heavy

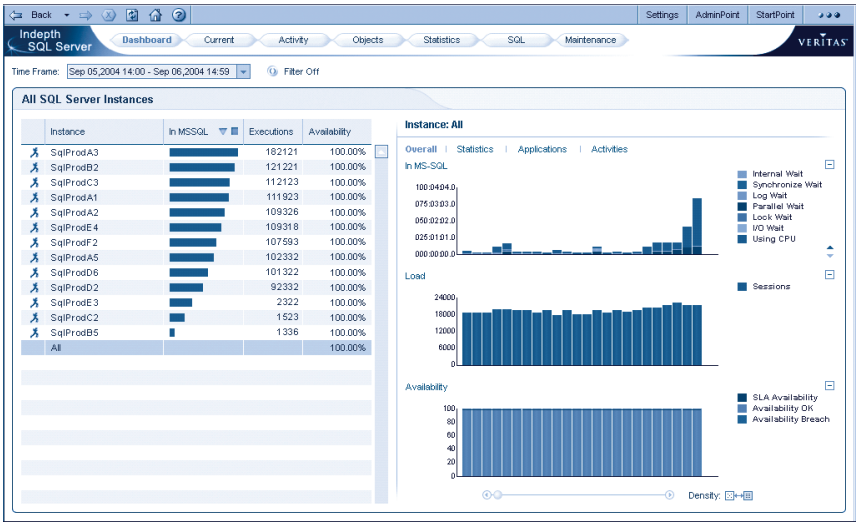


Figure 1. Indepth for SQL Server multiple-instance dashboard

processing loads may be bogging down the performance of databases, programs, statements, and batch jobs as well as which programs and statements are accessed by each user, program, remote system, and so forth.

When an application suffers from slow database response times, a tool such as Symantec Indepth for SQL Server can help administrators determine the root cause of the performance problem and identify the quickest way to correct the situation. Indepth for SQL Server is designed to offer the following capabilities:

- Correlate performance information from SQL Server sessions with user activity, program execution, batch jobs, database object activity, file and I/O activity, and so forth
- Correlate performance information with data about the database objects, tables, and indexes; storage layout and I/O activity on physical files; storage layout on central storage arrays such as Dell/EMC CX series arrays;[1] and data about enterprise resource planning (ERP) and customer relationship management (CRM) application servers, remote machines, and logged-on users
- Present the requisite information in a single, easy-to-understand screen display (see Figure 1), enabling database administrators to make an appropriate decision

Symantec Indepth for SQL Server also provides a built-in, intelligent index-recommendation mechanism called SmarTune. This mechanism features extensive capabilities for understanding and comparing execution plans as well as advanced tools to help resolve I/O, CPU, and locking problems. Consequently, Indepth for SQL Server is well suited for addressing performance problems and database tuning.

Increase productivity with continuous monitoring

Symantec Indepth for SQL Server continuously monitors the SQL Server environment—typically providing three samples per second—and captures performance data that enables current, short-term, and long-term performance analysis. To investigate a bottleneck such as a locked session or a runaway process, administrators can use a present-time snapshot of database activity or review the performance data of a recent activity (see Figure 2).

Indepth for SQL Server stores the performance data it collects in an advanced, self-managed performance warehouse database. The performance warehouse not only allows administrators to investigate previous performance problems, but also helps administrators

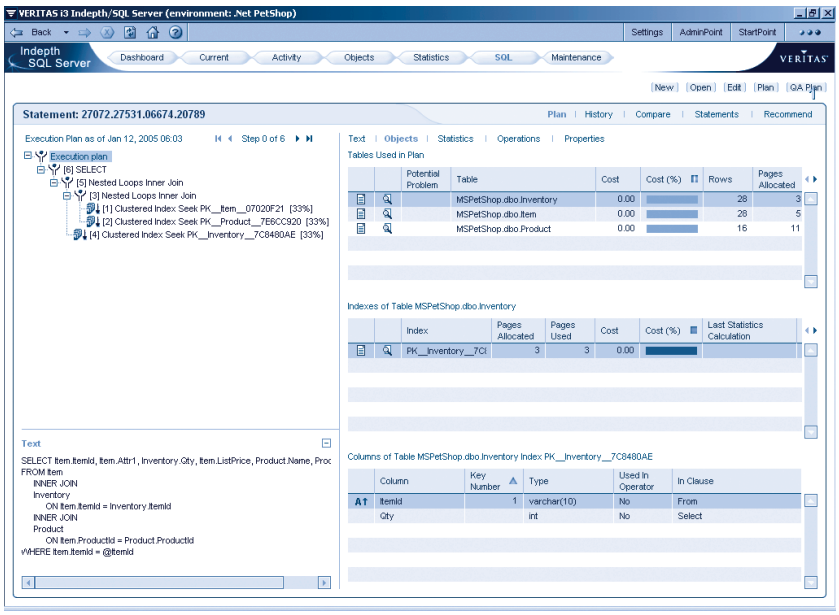


Figure 2. Statement tuning from the SQL workspace within Indepth for SQL Server

[1] Symantec offers Storage Extension for Dell/EMC CX series arrays as an add-on option to Indepth for SQL Server. This option is designed to extend the performance management of a SQL Server environment to the networked Dell/EMC storage.

identify performance trends that may require attention before they become problems.

Integrate third-party ERP and CRM applications

Symantec Indepth for SQL Server can be integrated into SAP, Siebel, and PeopleSoft software as well as other ERP and CRM solutions using specialized extensions (see Figure 3). These extensions are designed to reach beyond the borders of the SQL Server database and allow administrators to easily correlate database activity with application entities such as application users, dialog steps, views, panels, and so forth. Indepth for SQL Server also supports Microsoft COM+ (the enhanced version of the Microsoft Component Object Model) in a similar manner.

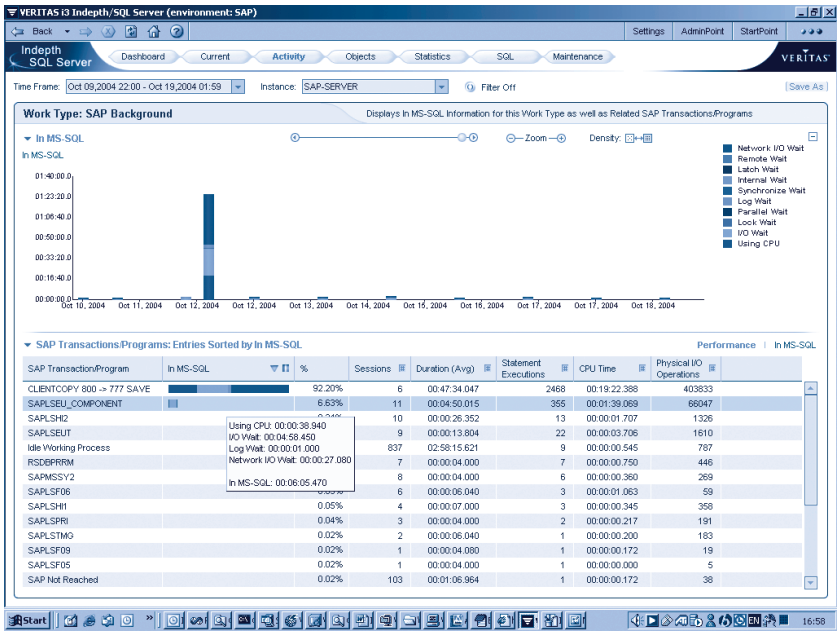


Figure 3. Indepth for SQL Server interface showing breakdown of SAP software transactions

Deploy Indepth for SQL Server: An example scenario

This section provides an example deployment scenario for Symantec Indepth for SQL Server. In this example, an enterprise within the railroad industry reported experiencing severe performance problems in its SQL Server environment—problems that threatened to interfere with national train traffic. At first, the IT team was convinced the performance issue resulted from CPU limitations of the four-processor Intel® architecture–based servers.

As due diligence before requisitioning eight-processor servers, the IT team performed a performance analysis of the SQL Server environment using Indepth for SQL Server. The analysis showed that two problematic SQL statements were consuming more than 90 percent of the CPU resources. By using Indepth for SQL Server, administrators were able to quickly identify and correct a sorting problem, tuning the problematic SQL statements to boost performance dramatically. In this example scenario, using Indepth for SQL Server enabled the IT team to free considerable CPU resources on the servers and resolve the performance problem without having to purchase eight-processor Intel architecture–based servers. In addition, administrators reported that they were able to further optimize resources by relocating additional databases onto the four-processor servers.

Achieve peak performance with Indepth for SQL Server

Symantec Indepth for SQL Server performance management software is designed to help improve and manage application response times by proactively monitoring, analyzing, and tuning the Microsoft SQL Server environment. This software tool provides a comprehensive view of application performance by capturing, measuring, and correlating performance metrics from each supporting tier of a SQL Server–based application infrastructure, including Web server, application server, database, and storage tiers. When problems are detected, Indepth for SQL Server helps pinpoint the cause and identify an effective course of action, while innovative SmarTune technology is designed to quickly correct the problem to enable peak system performance. Indepth for SQL Server helps streamline SQL Server operations, optimize CPU utilization, enhance application response time, and improve efficiency by providing administrators with a comprehensive view of application performance.

Ron Gidron is a regional product manager at Symantec Corporation in the Europe, Middle East, and Africa (EMEA) division for the Application Performance Management Group. Ron has more than eight years of experience in application performance management, working on global projects in numerous industry sectors and providing performance consulting to large corporations worldwide. Additionally, Ron has a software engineering and sales engineering background.

FOR MORE INFORMATION

Dell and Symantec:
www.dell.com/symantec

Dell and Microsoft SQL Server:
www.dell.com/sql

Symantec Indepth for SQL Server:
www.veritas.com/Products/www?c=product&refId=317

Planning for Blade Server Deployment in the Data Center

The modular Dell™ PowerEdge™ 1855 blade server is designed to optimize rack environments by integrating up to 10 server blades as well as storage, networking, power, and management components within the 7U Dell Modular Server Enclosure. While this approach enables significant efficiency enhancements, blade server deployment requires careful planning of data center resources such as power, networking, infrastructure fabric, cooling, and management access.

BY NARAYAN DEVIREDDY AND MICHAEL BRUNDRIDGE

Related Categories:

Best practices

Blade servers

Data center density

Dell PowerEdge blade servers

System deployment

Systems management

Visit www.dell.com/powersolutions for the complete category index.

To meet the demands of growing data centers, many IT administrators are turning to modular server architectures that are optimized for rack environments. A modular server architecture comprises a set of hardware components and an integrated management environment. The Dell PowerEdge 1855 blade server is based on an industry-standard modular server architecture. Its 7U chassis—known as the Dell Modular Server Enclosure—contains up to 10 server blades; a midplane; common infrastructure modules for storage, networking, power, and cooling shared by all blade servers in the chassis; and a management module that enables an integrated management environment.

Benefits of a modular server architecture include improved rack-space density compared to traditional rack-mountable servers, server aggregation, I/O resource consolidation, cable consolidation, integrated chassis management, rapid deployment and provisioning, and enhanced manageability. Built on industry-standard technology, the PowerEdge 1855 blade server helps achieve high efficiency in the data center by providing an optimized rack environment.

Although blade servers are not fundamentally different from traditional rack-mountable servers, upfront planning specifically for blade servers can save administrators time later on in the deployment process. Systems deployment involves several steps that must be carefully and accurately executed to achieve the expected results. Advance planning of the following data center resources can help ensure a smooth deployment of blade servers:

- **Power requirements:** Planning for power requirements requires a review of the overall system configuration (number of server blades, number of processors, amount of memory, number of I/O modules, and so forth); available configurations for system power supply units (PSUs); power redundancy; and data center power configurations.
- **Network infrastructure:** IP addresses must be allocated to the individual server blades as well as to the chassis management modules.
- **Cooling and thermal considerations:** The data center's cooling architecture must provide adequate

PSU type	2,100 watts
Voltage	200–240 VACrms
Frequency	47–63 Hz
Input current	13.1 Arms at 200 VAC
Input power	2,620 watts at 200 VAC
AC input cord	IEC320-C13 at PSU IEC320-C20 at PDU

Figure 1. Power supply specifications for the Dell Modular Server Enclosure

airflow and cooling in and around the rack where the blade servers will be deployed.

Power requirements

The Dell Modular Server Enclosure can contain two to four power supply modules. The PSUs can be arranged in various policy configurations, ranging from $2+n$ to $4+0$. Figure 1 shows power supply specifications for the Dell Modular Server Enclosure. The power supplies must be plugged into a Dell-approved power distribution unit (PDU).

The Dell Modular Server Enclosure is configured with two 2,100-watt power supplies capable of 200–240 volts AC (VAC). Additional 2,100-watt power supplies can be configured in the chassis for redundancy. *Note:* The power supply wattage rating discussed throughout this article (2,100 watts) is based on the amount of 12-volt DC output from a single power supply. It does not represent AC wattage requirements for the entire blade server because those depend on specific configurations.

Calculating how much power is required

Because new modules are likely to be released over the course of a chassis life cycle, best practices recommend using the Dell Product Configuration Calculator to determine power needs before introducing changes in the chassis. This calculator is designed to provide the power and airflow requirements for a specified configuration. The calculator can be found on the Dell Web site at www1.us.dell.com/content/topics/topic.aspx/global/products/pedge/topics/en/config_calculator?c=us&cs=RC968571&l=en&s=hea.

Configuring power redundancy

Power redundancy is the ability to provide reserve power if power supply hardware fails or an external AC power source fails. Power supply redundancy policies are commonly referred to as $n+y$, where n is the number of power supplies needed to power the chassis and y is the number of supplies reserved for redundancy.

In addition to redundant power supplies, data centers can implement redundant AC power grids. If the data center has more than one AC power grid being supplied from the power company, then the power supplies can be wired to be AC grid redundant. This configuration enables the chassis to remain fully powered and operational even if one power grid fails. However, if a data center has only one power grid—a configuration known as DC redundant—all power supplies will lose power and the chassis will power down if the power grid fails, in the absence of any other power protection such as an uninterruptible power supply (UPS).

The Dell Modular Server Enclosure was initially offered with 1,200-watt power supplies; however, in the first half of 2005, Dell replaced the 1,200-watt components with 2,100-watt power supplies, which allows for additional wattage as well as additional policies. Dell is developing a new firmware release for the Dell Remote Access Controller/Modular Chassis (DRAC/MC) that is expected to enhance power policies for the Dell Modular Server Enclosure.[1] Figure 2 shows the power policies that will be possible with this firmware release. The $2+2$ AC power policy is a redundant configuration designed to protect the chassis against power supply hardware failure as well as against AC grid outages (assuming only a single grid fails). *Note:* If a mixed power supply configuration is used—that is, a 1,200-watt supply and a 2,100-watt supply residing in the same chassis—the power policy will be $3+n$ because the smaller supply dictates the power policy in such a configuration.

Using parallel power supplies

The power supplies in the Dell Modular Server Enclosure use a parallel design, which enables the power needed to sustain the chassis to be drawn equally from all the power supplies installed in the chassis. This type of design offers several benefits such as expandability, the ability to replace a failed power supply without interrupting the flow of power to the chassis, and the ability to support redundancy.

The parallel power supplies rely on a sophisticated, highly accurate current-sharing method known as active current sharing. This method uses a closed-loop feedback circuit, or *load share bus,* between the power supplies. The load share bus is designed to synchronize the power supplies so that they provide the same power output on the 12-volt rail. However, even with such technology, it is virtually impossible for each power supply to output exactly the same amount of wattage because of slight tolerance differences between supplies. To account for output wattage differential when more than one power supply is installed,

Policy	1,200-watt supplies	2,100-watt supplies	Mixed configuration
4+0	Yes	Yes	No
3+0	Yes	Yes	Yes
3+1	Yes	Yes	Yes
2+0	No	Yes	No
2+2	No	Yes	No

Figure 2. Power policies planned for the Dell Modular Server Enclosure (pending DRAC/MC firmware release expected in the first half of 2006)

[1] Dell expects to make this DRAC/MC firmware release available in the first half of 2006.

I/O module	External port speed	External ports	Fabric type	Media type
Dell PowerConnect 5316M switch	10/100/1,000 Mbps	6	Ethernet	RJ-45
Ethernet pass-through	1,000 Mbps	10	Ethernet	RJ-45
Fibre Channel pass-through	1 or 2 Gbps auto-sensing	10	Fibre Channel	Small form-factor pluggable (SFP)
Brocade SilkWorm 3014 switch	1 or 2 Gbps auto-sensing	4+1*	Fibre Channel	SFP
McDATA 4314 switch	1 or 2 Gbps auto-sensing	4+1*	Fibre Channel	SFP
Topspin InfiniBand pass-through	4X InfiniBand (10 Gbps)	10	InfiniBand	InfiniBand copper

**A single dedicated 10/100 Mbps Ethernet management port*

Figure 3. I/O module types and associated port, fabric, and media specifications

the DRAC/MC compensates for this condition and displays the potential differential as "Load Sharing Overhead" on the Power Budget Status page.

Invoking the DRAC/MC power budget algorithm

The DRAC/MC power budget algorithm is responsible for determining whether the Dell Modular Server Enclosure has sufficient power available to support the devices within it. When an additional device is installed or a request is submitted to power up the device, that device sends a command to the DRAC/MC, which in turn will determine whether there is sufficient power in the chassis to support the device. If sufficient power is available, the DRAC/MC will power up the device and adjust the available remaining power by subtracting the amount of power the device needed to power up. If a request to power down the device is submitted, the DRAC/MC will power down the device and then adjust the available power by adding back the amount of power the device had previously requested. The same process occurs if the device is removed before a power-down request is submitted. If a device submits a request to power up and sufficient power does not exist to do so, the DRAC/MC puts the device in a queue and powers up the device only when sufficient power is available.

If a power supply fails, the DRAC/MC recalculates the power budget, and if sufficient power does not exist to maintain power to all devices because of a lack of redundancy, the DRAC/MC will attempt to power down enough server blades to maintain the chassis. For the DRAC/MC to power down a server blade in such an event, the blade OS must support Advanced Configuration and Power Interface (ACPI) commands and be properly configured. In a nonredundant power supply configuration, the DRAC/MC will attempt to power down the server blades starting at the blade in the highest-numbered slot to help prevent the chassis from overloading the remaining power supplies. However, the OS controls the shutdown process and may not allow the DRAC/MC to accomplish shutdown efficiently. To help protect against overloading the remaining power supplies, administrators should configure the Dell Modular Server Enclosure with redundant power supplies.

Dell is currently enhancing the DRAC/MC power budget algorithm. The upcoming DRAC/MC firmware release, discussed in the "Configuring power redundancy" section in this article, is being designed to enable administrators to set the power policy and view the results in a concise display. The power policies expected to be available in this firmware release also will be designed to allow for all redundancy policies: $n+0$, $n+1$, and $n+2$.

Following best practices

Dell best practices recommend configuring the Dell Modular Server Enclosure with redundant power supplies to help prevent a power supply or grid failure from affecting the entire chassis. The power supplies should be configured, at a minimum, in an $n+1$ configuration to help prevent a power supply event from causing a complete chassis shutdown. Best practices also recommend using the DRAC/MC redundant option to enable device power management and remote chassis management if the DRAC/MC or management network fails.

Network infrastructure

The Dell Modular Server Enclosure provides various I/O module options. I/O modules installed in the chassis will determine what media type is required. Figure 3 lists these I/O module options and the associated port, fabric, and media specifications.

Using network ports on the blades

Each server blade within the Dell Modular Server Enclosure has two on-board 1 Gbps[2] network controllers. These network controllers, commonly referred to as LAN on Motherboards (LOMs), are routed to I/O modules in bays 1 and 2 located at the rear of the chassis. I/O module bays 1 and 2 support only Ethernet fabric types.

The Dell Modular Server Enclosure also supports a second fabric. Each server blade has a provision for a fabric daughtercard, which can be any of three supported fabrics: Ethernet, Fibre Channel, or InfiniBand. However, the chassis can support only one fabric type installed in this second fabric; therefore, if a blade has a daughtercard installed, it must be of the same fabric type as the I/O module. Some blades can have daughtercards while other do not, but blades cannot have different types of daughtercards in the same chassis.

In addition to the I/O modules, the Dell Modular Server Enclosure includes ports for the DRAC/MC and KVM (keyboard, video,

[2] This term does not connote an actual operating speed of 1 Gbps. For high-speed transmission, connection to a Gigabit Ethernet server and network infrastructure is required.

Device	Network	Video	Other
DRAC/MC	10/100 Mbps RJ-45	n/a	9-pin serial
Digital KVM switch	10/100 Mbps RJ-45	DB15 VGA	6-pin PS2 keyboard 6-pin PS2 mouse
Analog KVM switch	n/a	DB15 VGA	6-pin PS2 keyboard 6-pin PS2 mouse

Figure 4. Port specifications for the Dell Modular Server Enclosure

mouse) switches. Figure 4 shows specifications for these ports.

The Dell Modular Server Enclosure houses several chassis modules that provide graphical user interfaces (GUIs) and command-line interfaces (CLIs) for management over the network. Chassis monitoring, out-of-band management, and KVM configuration can be performed through the DRAC/MC interface. The Dell PowerConnect™ network switch and the Brocade and McDATA Fibre Channel switches each provide an interface for switch configuration. Best practices recommend establishing a dedicated management subnet to access the chassis components for management purposes. The following tasks can help IT administrators define the management subnet:

- Identify a range of IP addresses and a subnet mask.
- Provide access to a Dynamic Host Configuration Protocol (DHCP) server from the management network if dynamic IP addresses are used. The DRAC/MC, Dell PowerConnect 5316M switch, Brocade SilkWorm 3014 Fibre Channel switch, and McDATA 4314 Fibre Channel switch support DHCP-based IP addressing and static IP addressing.
- Provide access to a Dynamic Domain Name System (DDNS) server from the management network. The DRAC/MC supports DDNS-based host name identification, and DDNS works with dynamic IP addresses assigned by DHCP servers.
- Provide access to a Microsoft® Active Directory® domain controller from the management network. The DRAC/MC supports an Active Directory–based user authentication mechanism.
- Provide access to FTP or Trivial FTP (TFTP) servers from the management network. The Dell Modular Server Enclosure firmware update process for the chassis modules is performed via an FTP or TFTP server.

Accessing the DRAC/MC management interface is one of the first steps involved in chassis management. The DRAC/MC provides a serial interface and a network interface designed to perform all configuration and systems management functions using the DRAC/MC Web-based interface or the serial/telnet console. The DRAC/MC default IP address is 192.168.0.120, which is configurable by the administrator. Refer to the DRAC/MC product documentation for the default username and password.

Cooling and thermal considerations

Cooling within the Dell Modular Server Enclosure involves two fan modules located at the rear center of the chassis and four power supplies divided between the left and right sides of the chassis. Both the fan modules and the power supply modules supply cooling to the entire chassis. The fans, when running at their maximum cooling capacity, are designed to consume 520 cubic feet per minute (CFM) of air from the front of the chassis. Data center administrators should help ensure that enough cool air is supplied to the front of the rack to support this airflow.[3]

Best practices for blade server deployments

Modular server environments, such as those based on the Dell PowerEdge 1855 blade server and the Dell Modular Server Enclosure, can maximize data center space while helping to streamline and centralize server management. However, to optimize an IT environment based on modular data center components, administrators should heed best practices for configuring power, networking, and cooling components to help ensure smooth server deployment and operations.

Narayan Devireddy is a development manager in the Dell Enterprise Systems Management Software organization. He has 14 years of systems management product development experience. Before joining Dell, he worked for Novell, Compaq, Cheyenne Software, and Computer Associates in different capacities. He has an M.S. in Computer Science from Alabama A&M University.

Michael Brundridge is a strategist in the Dell OpenManage Development organization. Before joining Dell, he worked as a hardware engineer for Unisys. Michael attended Texas State Technical College and has a technical degree from the Southwest School of Electronics.

FOR MORE INFORMATION

Dell PowerEdge 1855 blade server:
www1.us.dell.com/content/products/productdetails.aspx/pedge_1855?c=us&cs=555&l=en&s=biz

[3] For more information about airflow configuration, see "PowerEdge 1855: Best Practice Recommendations to Aid in Data Center Deployment" by the Dell Enterprise Product Group, January 2005, www.dell.com/downloads/global/products/pedge/en/PowerEdge%201855%20DC%20Whitepaper.pdf.

Deploying Dell PowerEdge 1855 Blade Servers

Using DRAC/MC Virtual Media

The Dell™ PowerEdge™ 1855 blade server has an optional digital KVM (keyboard, video, mouse) module that can be installed in the Dell Modular Server Enclosure housing the server blades and shared systems components. This KVM module provides virtual media capability, enabling administrators to assign a remote CD or DVD drive, ISO image, or floppy drive to individual server blades. The remote media device then appears as if it is attached directly to the server blade. This virtual media feature can be used to deploy and provision a blade server remotely.

Related Categories:

Blade servers

Data center density

Dell PowerEdge blade servers

Keyboard, video, mouse (KVM)

Remote management

System deployment

Systems management

Virtual media

Visit www.dell.com/powersolutions for the complete category index.

BY JAKE DINER AND ALAN BRUMLEY

In blade server environments, cable consolidation and pooled resources for power and network I/O can help reduce costs and streamline management while enabling high availability through redundancy. Systems such as the Dell PowerEdge 1855 blade server are designed to offer cost-effective, high-density computing power. The 7U chassis of the Dell PowerEdge 1855 blade server—known as the Dell Modular Server Enclosure—accommodates an Avocent Digital Access KVM (keyboard, video, mouse) module. This module resides within the chassis and offers two features: a remotely accessible virtual console and virtual media.

Accessing server blades through the remote console

The Avocent Digital Access KVM module provides a remote console by emulating a keyboard and mouse at the hardware level as well as probing the video signal and then processing the image. This allows the remote console feature to function without drivers regardless of the OS or graphics mode used. Administrators can access the remote console via the Web-based interface of the Dell Remote Access Controller/Modular Chassis (DRAC/MC)—which also resides in the Dell Modular Server Enclosure. On the Console page of this interface,

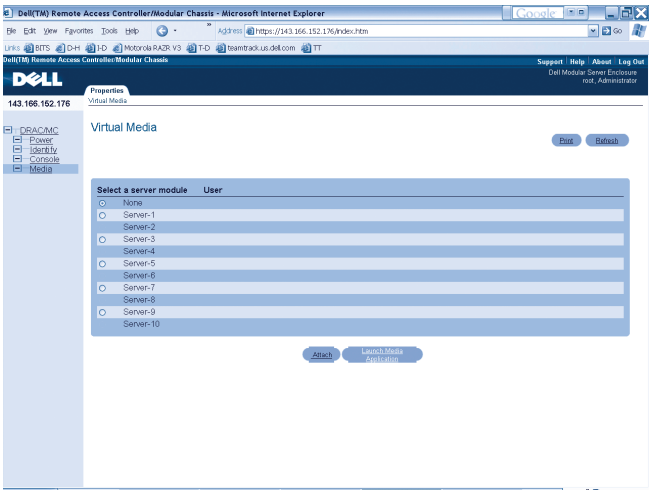


Figure 1. Connecting server blades to virtual media through the DRAC/MC interface

administrators can select the appropriate server blade to view. After selecting the server blade, administrators can launch the Java-based application that decodes the video and encodes the keyboard and mouse movements.

Connecting to remote storage devices with virtual media

A key feature provided by the Avocent Digital Access KVM module is virtual media, which allows administrators to remotely connect a mass storage device to a server and use the device as if it were present on the server. Like the remote console, virtual media uses hardware to emulate USB devices with native USB mass storage drivers, helping ensure no other drivers are necessary. Virtual media is accessed in a similar manner as the remote console—administrators navigate to the DRAC/MC interface and select the server blade that needs to be controlled. From there, a console is launched that facilitates selection of the floppy or optical device (or supported image file). Once the device is selected, it is virtually connected to the OS. After the device is connected, the OS treats it as if the administrator had physically inserted a USB device, floppy disk, or CD into the remote server.

Booting from a virtual CD

Server administrators can use the Avocent Digital Access KVM module to boot from the Dell OpenManage™ Server Assistant CD, which is provided with the Dell PowerEdge 1855 blade server. Before starting, administrators must install Java Runtime Environment (JRE) version 1.4.2 or later. *Note:* Earlier versions of JRE have known issues that may reduce the reliability of the virtual media applet that runs on the client workstation. This update is required for both Linux® OS–based and Microsoft® Windows® OS–based clients.

After installing the prerequisite JRE, administrators can use the Web browser to connect to the DRAC/MC and log in using administrative credentials. The account used when logging in to the DRAC/MC must have sufficient security privileges to access virtual media and the remote console. After connecting and logging in, administrators should expand the DRAC/MC tree view and select "Media" (see Figure 1).

Next, administrators should select the server blade they wish to connect to virtual media and click the Attach button. The OS will then detect an insert notification on the USB interface. At this time, two new USB devices will be "present" on the server blade: the CD/DVD device and the mass storage device. However, these devices will appear to the OS without any media inserted. The DRAC/MC interface page should reload to show the username next to the appropriate server. Administrators should then click the Launch Media Application button.

A key feature provided by the Avocent Digital Access KVM module is virtual media, which allows administrators to remotely connect a mass storage device to a server and use the device as if it were present on the server.

Virtual media application

The client machine will download and launch a Java-based application. The system configuration determines which items are listed in the application as well as which drives are available. Figure 2 shows an example Java-based application screen for a system that has one floppy drive available and one CD drive (D:). Using such an application, administrators can select a media image instead of a physical drive for each media type. If using a physical drive, administrators should ensure that the media is inserted before

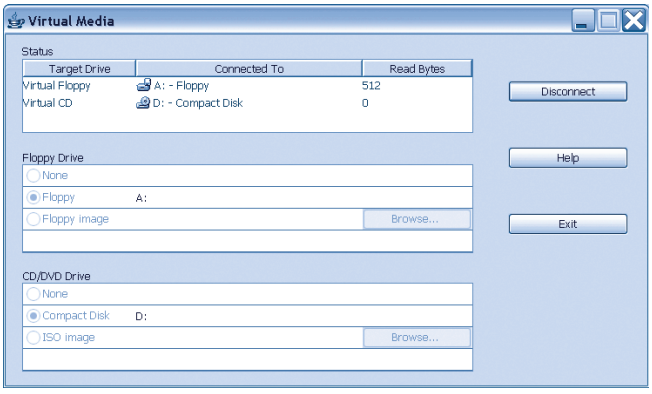


Figure 2. Connecting virtual media to target devices with the Java-based application

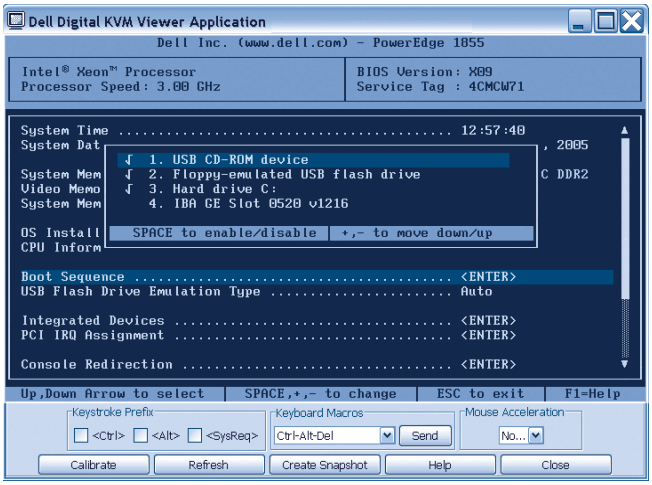


Figure 3. Accessing the server blade BIOS setup through the remote console

selecting the drive and clicking the Connect button. After this step, the OS may detect media has been inserted. Depending on the OS configuration, the OS may mount the device and begin reading from it. The top portion of the screen shown in Figure 2 displays status information about which devices are connected and how many bytes have been transferred through the network.

Once the application has initiated the connection, the media may not be changed until administrators click the Disconnect button. At this time, the media may be changed, or a different image file may be selected. Best practices strongly recommend against trying to swap media or change files without disconnecting. Failure to do so may result in the remote server receiving corrupted information because of client OS caching, for example. If a virtual floppy disk is being used, it may be read and written to as if the floppy disk were present on the server. When using an image file or virtual CD drive, the media will always be read-only even if a CD-RW drive is connected.

Virtual media boot

Once virtual media is connected to the server blade, administrators can remotely boot the system using virtual media. To do this, administrators should select "Console" from the DRAC/MC tree. From there, a Web-based interface similar to the one shown in Figure 2 will load; this interface allows administrators to access a server blade and then launch a remote console application. The process is very similar to launching the virtual media application.

Once the remote console is launched, it will show whether the OS on the remote server is running, which depends on the server state. If the server is not powered up, administrators can simply select "Power" to access an interface that will let them power up the server blade being managed. From this point on, the management experience will be as if the administrator were physically present at the server.

Administrators can press F2 during the boot process to access the BIOS setup (see Figure 3). Under the Boot Sequence menu, "USB CD-ROM device" and "Floppy-emulated USB flash drive" will appear, and these items can be moved up and down in the boot sequence. Administrators can simply move one or both devices to the top of the list to enable booting from virtual media.

Installation and configuration of the OS will also proceed as if the administrator were physically present at the server. The only difference is that administrators must use the virtual media applet to disconnect and connect the media when changing CDs or floppy disks.

Enabling remote management for modular server environments

Modular server systems such as the Dell PowerEdge 1855 blade server enable enterprises to concentrate computing power into a minimal form factor—allowing administrators to use data center space efficiently. However, such dense rack configurations can make systems difficult to access. The 7U Dell Modular Server Enclosure housing the PowerEdge 1855 blade server accommodates optional components that provide remote management capabilities to help ease server administration. In particular, the Avocent Digital Access KVM module and the DRAC/MC enable administrators to access server blades from a remote console and use virtual media devices to install software and configure server blades from virtually anywhere across the enterprise.

Jake Diner is a software engineer in the Dell Enterprise Systems Management Software organization. His interests involve public speaking and wireless communication. Jake has a B.S. in Computer Science from Michigan State University.

Alan Brumley is a software engineer and senior consultant in the Dell Enterprise Systems Management Software organization. He is the lead engineer on the Dell OpenManage Deployment Toolkit and has worked at Dell for more than five years. His interests involve embedded system design and radio-control aviation. Alan has a B.S. in Computer Engineering from the University of South Carolina.

FOR MORE INFORMATION

Dell Remote Access Controller/Modular Chassis User's Guide:
support.dell.com/support/edocs/software/smdrac3/dracmc/index.htm

Avocent Digital Access KVM Module for the Dell PowerEdge 1855 Blade Server

Expanding enterprises require increasingly dense data centers. Consequently, more IT managers are adopting flexible, high-performance platforms such as the Dell™ PowerEdge™ 1855 blade server. To help minimize space, reduce management costs, and restrict cable clutter, system administrators can incorporate an Avocent® Digital Access KVM (keyboard, video, mouse) switch in the Dell Modular Server Enclosure that houses PowerEdge 1855 server blades. Avocent Digital Access KVM switches permit single-console control of the 10 individual server blades within the Dell Modular Server Enclosure—whether administrators are across the hall or halfway around the world.

BY ROBERT LESIEUR AND GREG KINCADE

Related Categories:

Avocent

Data center density

Data center technology

Dell PowerEdge blade servers

Keyboard, video, mouse (KVM)

Remote management

Systems management

Visit www.dell.com/powersolutions for the complete category index.

The Dell PowerEdge 1855 blade server can be an optimal platform for dense data centers. It comprises up to 10 server blades and is housed in the 7U Dell Modular Server Enclosure, which also supports built-in power and cooling modules, storage, switches, and systems management components such as the Dell Remote Access Controller/Modular Chassis (DRAC/MC). The Avocent Digital Access KVM (keyboard, video, mouse) switch module is an optional component that can be included in the Dell Modular Server Enclosure. When used in conjunction with the DRAC/MC, this switch module enables remote console and virtual media capabilities that are designed to enhance manageability of the Dell Modular Server Enclosure.

Options for connecting the KVM console

The Avocent Digital Access KVM module includes a single port for a local KVM cable, which provides a direct connection to an external keyboard, monitor, and mouse (see Figure 1). This approach is designed to provide full

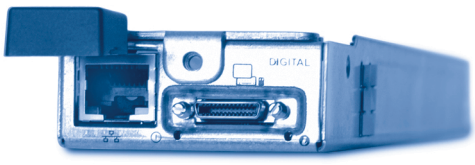


Figure 1. Avocent Digital Access KVM connector for the Dell Modular Server Enclosure

control of server functions, including local OS installation, blade server configuration, maintenance, and troubleshooting.

An Avocent Digital Access KVM switch module can be installed in the Dell Modular Server Enclosure to replace the default Avocent Analog KVM module. Incorporating the KVM switch into the Dell Modular Server Enclosure enables organizations to dramatically reduce cable requirements for PowerEdge 1855 server blades, compared to the cabling that is required for the same number of traditional 1U servers. Integrated switching capability allows all 10 server blades in the PowerEdge 1855 blade server to be connected with only one Category 5 (Cat 5) cable—and in a dense data center, this reduction of cable sprawl can conserve valuable real estate.

The Dell Modular Server Enclosure offers three options to connect a KVM console to the server blades using the Avocent Digital Access KVM module. Administrators may choose to:

- Connect the KVM cable directly to the PS/2 and video ports of the local KVM cable on the Avocent Digital Access KVM module. The KVM switch's on-screen menu lets administrators select any one of the 10 server blades in the chassis.
- Connect the PS/2 and video ports of the local KVM cable to the PS/2 and video ports of a server interface pod (SIP), and connect a Cat 5 cable between the SIP and an external digital KVM over IP switch from Dell or Avocent. Individual server blades then appear in the menu for the external switch just as any server would if it were directly connected.
- Connect a Cat 5 cable between the Avocent Digital Access KVM module and the same IP management network to which the DRAC/MC is connected. In this case, administrators simply use a browser to connect to the DRAC/MC module and log in with a password, and then they select the Console or Media tab (to access the KVM console or virtual media, respectively) and the server blades they wish to access. From there, a viewer appears (if the Console tab was selected) or a virtual media applet appears (if the Media tab was selected).

Security and compression

With remote virtual KVM access, a server and its controlling keyboard, monitor, and mouse can be miles apart. To prevent performance degradation and to maintain session security, the Avocent client software is designed to work with the embedded Avocent KVM switch hardware to implement RC4-compatible data encryption and compression along the entire path. Keystroke and mouse movements are protected by Secure Sockets Layer (SSL) encryption; keys expire with the termination of each KVM session. Permissions and access levels can be specified for individual users and devices, and detailed user logs can provide audit trails for added security. And because the server's KVM activity is routed over the existing TCP/IP network, industry-standard encryption methods can be used to enable fast, secure access around the world.

Video data is highly compressed and packetized using the Avocent Dambrackas Video Compression™ (DVC) algorithm. Specifically optimized for interactive user sessions of administrative consoles, the 7-bit DVC algorithm is designed to provide extremely high compression rates. The Avocent DVC algorithm helps minimize control/response latency, while encryption and compression functions are embedded in the Avocent KVM switch hardware. This combination is designed to provide a near-real-time response that can create a "virtual presence experience" similar to using a keyboard, monitor, and mouse that are directly attached to the target server.

Virtual media capability

Besides commanding server blades via a KVM switch, IT administrators often must load system software remotely. To permit this, the Avocent Digital Access KVM switch supports simultaneous virtual connection to two remote USB devices. These devices may include one generic mass-storage device or file system (such as a floppy drive, USB flash drive, or a floppy image) and one CD/DVD storage device or file system (such as a CD, DVD, or ISO image). This virtual connection to remote media, a capability known as virtual media, simulates devices such as hard drives, CD drives, or USB devices and allows mass-storage functions to be performed without a physical connection to individual server blades.

Virtual media capability can help organizations maintain secure physical access to data centers. Administrators can install software remotely from virtual media devices—and can even provision an OS on a server. Virtual media contributes to low total cost of ownership by allowing administrators to stay at their desks and perform tasks remotely, helping to eliminate unnecessary trips to the data center. Because the Avocent Digital Access KVM switch operates independently of the OS, individual server blades within the Dell Modular Server Enclosure can run whichever OS is required. The user interface provides independent mechanisms for controlling both KVM and virtual media switching between server blades.

Administrators may launch simultaneous virtual media and virtual KVM console application sessions independently from a remote client. For example, administrators can begin loading a system update or security patch on one server blade while running a virtual KVM session on another, and then revert to the first system to check its progress.

Integration into the existing KVM infrastructure

The Avocent Digital Access KVM switch in the Dell Modular Server Enclosure can be integrated into a data center's existing KVM infrastructure. When the built-in switch is connected to an external KVM switch such as the Dell 2161DS KVM over IP switch, the interface recognizes the server blades in the Dell

Modular Server Enclosure and allows administrators to select them from the server list.

Because the integrated KVM switch can control all 10 server blades in the Dell Modular Server Enclosure, fewer external KVM switch ports are consumed when compared to traditional servers. This can lead to significant savings in data center space and costs. Administrators can treat server blades and stand-alone servers identically.

Management interface options for KVM switches

Depending on switching access—traditional KVM or virtual KVM—administrators have two primary means to enable KVM functionality. The Avocent On-Screen Configuration and Activity Reporting (OSCAR®) overlaid menu system provides access to traditional KVM switching functions. Virtual KVM functionality is accessed through a standard Web browser over the DRAC/MC interface.

Using OSCAR

OSCAR provides administrators with intuitive menus to configure the KVM switch and to select servers to control via the KVM console. The OSCAR graphical user interface (GUI) can be used to navigate among connected servers (see Figure 2). Keyboard hot-key sequences such as Print Screen allow administrators to access the OSCAR interface at any time, regardless of the OS on the server being managed.

Figure 2. OSCAR GUI

From the OSCAR menu, an administrator can select a specific server blade from a list of the server blades in the Dell Modular Server Enclosure that are connected to the built-in KVM switch. The list displays up to 15 characters of the server name that the network administrator assigned at installation. The OSCAR menu is designed to address every server blade connected to that switch. Flags on the on-screen display indicate the server blades that are currently connected to the KVM switch.

Using the DRAC/MC interface

Virtual KVM functionality is available through a standard browser pointed to the IP address or URL of the DRAC/MC Web server. Administrators can connect to the server blades using the interface provided by the embedded Web server of the DRAC/MC. The Web console interface looks and operates similar to the OSCAR interface (see Figure 3). A login screen provides secure authentication to the server blades in the Dell Modular Server Enclosure. As with the OSCAR interface, once authenticated to the DRAC/MC Web console

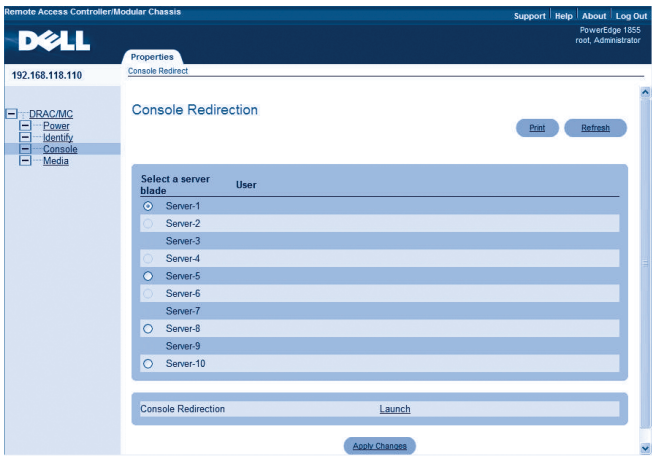


Figure 3. DRAC/MC Web console interface

with a password, the administrator has access to every server blade in the enclosure. To secure server blades, network administrators should enable OS security for each server blade.

From the Web console, administrators can select the server blades to control, configure the digital KVM settings of the Dell Modular Server Enclosure, and launch virtual media sessions to provide access to remote storage. The Web console supports Microsoft® Internet Explorer browsers on Microsoft Windows® platforms and Mozilla browsers on Linux® platforms. As long as administrators employ a supported browser, the target server blade can run either Windows or Linux.

Remote server management in high-density data centers

In the past, IT managers installed KVM switches to simplify local access in dense server environments. The Dell Modular Server Enclosure with the Avocent Digital Access KVM switch helps further simplify server management in dense environments by enabling remote access and control. Regardless of the connection method, communication with server blades is designed to be fast, secure, and reliable because embedded Avocent virtual media and virtual KVM technologies leverage legacy IP connectivity and industry-standard security models.

This approach enables administrators to load software—patches, updates, applications, or even a new OS—from a remote location virtually anywhere in the world. Combining a Dell Modular Server Enclosure with an Avocent Digital Access KVM switch module can help ease the burden of management and keep total cost of ownership low by providing unparalleled access to data center systems.

Robert Lesieur is the product manager for Dell PowerEdge blade server systems.

Greg Kincade is a senior applications engineer for Avocent Corporation.

The Basics of Application Packaging:

Best Practices for Enabling Reduced Software Management Costs

Application packaging can help enterprises manage growing volumes of software for desktop and server systems efficiently. By streamlining software configuration and deployment, application packaging can help reduce application management costs. The information in this article pertains to OS migrations for desktop systems, including best practices for implementing application packaging techniques.

BY JUKKA KOULETSIS

Related Categories:

Altiris/Wise

Altiris

Application development

Application packaging

Microsoft Systems Management Server (SMS)

Systems management

Visit www.dell.com/powersolutions for the complete category index.

Maintaining desktop and notebook systems has become an expensive proposition for many corporate networks. New application management techniques are being developed to help enterprises administer their existing PC investments more efficiently, reduce end-user support costs, and minimize end-user business disruptions. One such development is application packaging, which is emerging as a critical component of an overall software configuration management strategy.

Application packaging involves the preparation of standard, structured software installations targeted for automated deployment. Automated installations, or *packages,* must meet the installation requirements for a specific environment: corporate standards for software usage and desktop design, multiple languages, regional issues, and software-related support issues. In addition, packages must be prepared for both commercial software and applications developed in-house.

To enable this level of application management, Microsoft now provides the Microsoft® Windows® Installer (MSI) service as a part of its desktop operating systems. Starting with Microsoft Windows 2000, the Windows Installer has been included in the base desktop OS. For earlier OS versions, Windows Installer can be downloaded from Microsoft's Web site. This database-driven service resides on workstations and controls the installing, uninstalling, patching, and repairing of software. Input into the Windows Installer is an .msi formatted file, which is further explained in the "Understanding Microsoft Windows Installer" section in this article.

When deploying new applications, many administrators consider an automated approach to software installation as an attractive option to help save time and help reduce compatibility issues. However, the testing needed to verify that a new application will not cause an old application to fail can be difficult, if not impossible, in an uncontrolled environment. Although business disruptions caused by a new application deployment often cannot be measured directly, impaired end-user productivity can be costly. This article examines how application packaging practices—and in particular, Windows Installer—can help address common application management concerns.

Understanding Microsoft Windows Installer

The Windows Installer service was designed to support every phase of the application management life cycle, providing a service to support each step involved in managing a desktop application from deployment through retirement (Figure 1). To support these functions, the Windows Installer needs to receive instructions from an

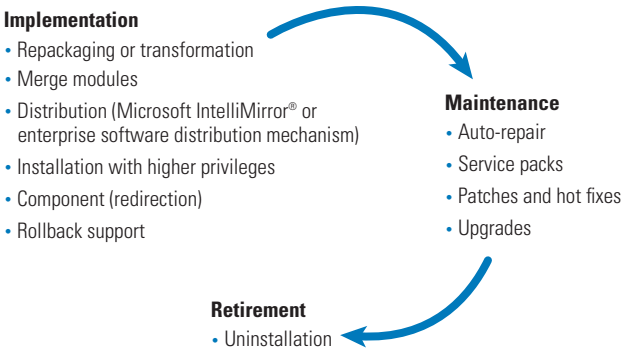


Figure 1. Application life-cycle tasks supported by Windows Installer

installation package. Previously, installation packages took the form of a setup.exe file. Unfortunately, inconsistencies in the way independent software vendors and internal software development groups created these installation files sometimes led to complications when administrators attempted to manage automated installations.

The emerging standard is for Windows Installer to use the msiexec.exe program to process the installation packages at an end user's PC. The packages follow a standardized database structure containing the information that Windows Installer requires to install or uninstall an application and to run the user interface for the setup. Each installation package includes an .msi file containing the installation database, a summary information stream, and data streams for various parts of the installation. The .msi file can also contain one or more transforms, internal source files, and external source files or cabinet files required by the installation. Figure 2 shows Windows Installer file extensions.

This approach enables Windows Installer to determine components that belong to an application, and to safely remove application components and restore a system to a working state. Furthermore, because Windows Installer is a service, it is designed to support software installations as the local Administrator role in locked environments, enhancing the application process.

Windows Installer can support applications installed from a network share—referred to as an administrative installation—or locally on an end user's PC. The downside to using a network share can be that systems receive patches or repairs only when they are connected to the network, which may be a consideration for organizations supporting many notebook users.

Creating MSI files

To build the installation package in the correct MSI database format, developers must collect information about each application. Items include executable files, installation instructions, configuration parameters, test instructions, and hardware and software dependencies. Once this information is gathered, the packaging effort is designed to produce a Windows Installer–based installation package for each application, including the installable MSI as well as any deployment files and wrappers required to distribute the package electronically to a PC.

Best practices recommend that installation packages be created by experienced packaging engineers, using tools specifically developed for that purpose. For example, Dell engineers have used Altiris® Wise Package Studio® software for efficient development of MSI installation packages. At minimum, the tools used to build installation packages should provide the following capabilities:

- OS certification analysis
- Capture of the installation snapshot
- Custom scripting and the creation of transforms
- Checking for compliance with packaging standards
- Conflict resolution
- System and end-user testing
- Integration with various software distribution engines

A completed package file should conform to standards and successfully install, uninstall, update, and function on the targeted OS. One package must be produced for each unique application configuration. Successfully packaged applications can be delivered to the project test team in a format ready for deployment on software distribution engines such as Microsoft Systems Management Server (SMS) 2003, the Altiris Software Delivery Solution™ tool, or Computer Associates Unicenter. Any department-specific requirements for installation can be handled through the creation of transforms within an existing MSI file.

Quality assurance (QA) testing should be performed during the creation of installation packages to verify that the MSI database is operating correctly to install, uninstall, and repair an application. In addition, peer reviews help ensure a high first-pass yield through the application distribution process by not requiring special permissions each time an application is installed.

Conflict management is a key aspect of package creation. Because of the way in which dynamic-link libraries (DLLs) are built

Extension	Description
.msi	Windows Installer database
.msm	Windows Installer merge module
.msp	Windows Installer patch
.mst	Windows Installer transform
.idt	Exported Windows Installer database table
.cub	Validation module
.pcp	Windows Installer patch creation file

Source: Microsoft Developers Network, msdn.microsoft.com/library/default.asp?url=/library/en-us/msi/setup/windows_installer_file_extensions.asp.

Figure 2. File name extensions used in the Windows Installer service

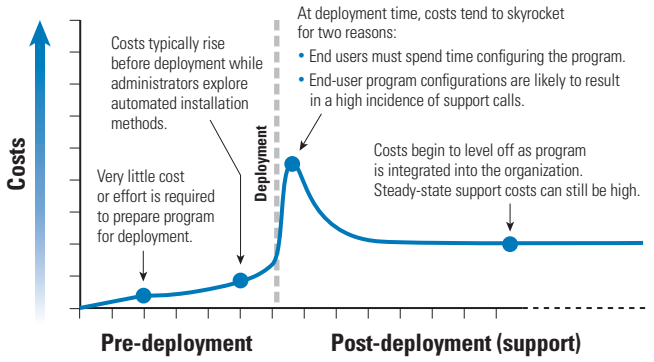


Figure 3. Deployment costs without an application packaging strategy

into a Microsoft OS, packaging engineers must determine whether a new application introduces conflicts. The conflict management process is designed to ensure that DLLs do not interfere with each other and that any required application isolation is configured into the MSI database.

Incorporating application packaging into software management

The application packaging process must fit within an organization's overall software configuration management strategy. Associated software configuration management tasks include application inventory and asset management, profile management, requirements gathering, user acceptance testing, and electronic software distribution.

Application inventory and asset management. This task involves compiling a list of applications that reside on each desktop system, grouping the applications by line of business, and determining which applications are necessary. Subsequently, administrators may load inventory information into a database that tracks assets and licenses on an ongoing basis.

Profile management. In this task, applications are logically grouped into *bundles* and entered into a profile management system. Each bundle is created within the context of a tier—for example, tier 0 is for the base OS, tier 1 is for company-wide applications, tier 2 is for division-wide applications, tier 3 is for departmental applications, and tier 4 is for individual user applications.

Requirements gathering. This task gathers the requisite technical data about each application, including the installation source code, test scripts, installation and uninstallation instructions, and any configuration details.

User acceptance testing. During this task, testing performed on an application package is designed to ensure that the application is functioning properly and to avoid business disruptions caused by application failure.

Electronic software distribution. This task involves moving the application package onto the software distribution server, performing QA to help ensure accurate delivery to a PC, and transitioning into steady-state support.

Minimizing application management costs

By implementing an application packaging strategy, organizations can help reduce administrative costs while providing business benefits. For example, this approach enables administrators to set and enforce corporate software configuration standards and to minimize the frequency of administrative errors during installation. The Windows Installer service also offers features that help streamline application deployment, including powerful self-repair and rollback capabilities that are designed to dramatically reduce the occurrence of deployment-related desktop software problems.

Figures 3 and 4 compare generalized application deployment scenarios. The Figure 3 example shows a typical pattern of deployment costs that are incurred when application packaging is not used, while the Figure 4 example shows a typical pattern of deployment costs when application packaging is used. *Note:* These are general findings; actual figures would be based on an organization's specific software configuration management cost structure.

Adhering to best practices for application packaging

To realize the benefits of application packaging, IT organizations should adhere to best practices for the following activities: gathering requirements, using Windows Installer, building a stable core image, managing conflicts, evaluating application suitability for packaging, establishing a centralized packaging process, creating a structure to group applications, and implementing a formal request process.

Requirements gathering. Those responsible for creating the application packages must collect the needed details about each application. Best practices recommend that an engineer experienced in creating MSI packages be responsible for gathering the technical data about an application.

Use of Windows Installer. Microsoft's Windows Installer technology is designed to simplify the process of adding applications to

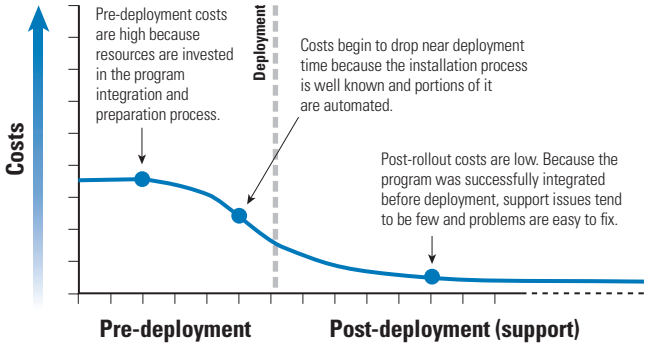


Figure 4. Deployment costs with an application packaging strategy

STREAMLINING APPLICATION PACKAGING WITH WISE PACKAGE STUDIO

Wise Package Studio is an application life-cycle management solution used by deployment and desktop management teams to prepare applications for the enterprise. Based on a structured application management, packaging, and QA process known as enterprise software packaging, Wise Package Studio helps administrators migrate to MSI-compatible application and patch packages while enabling high-quality, reliable deployments that support corporate standards. In turn, organizations can benefit from quick software rollouts, streamlined Windows Installer migration, and high return on Windows 2000 and Windows XP investments. The Wise Package Studio product family includes Professional Edition, Quality Assurance, Enterprise Management Server, and Standard Edition.

Wise Package Studio also offers advanced integration with Altiris IT life-cycle management products, including Altiris Patch Management Solution™ software. For more information, visit www.wise.com or www.altiris.com.

the desktop and to minimize support costs by helping to eliminate errors associated with those installations.

Stable core image. Engineers should not begin packaging until the core OS image has been defined and stabilized. Attempts to build and test packages on early versions of a core image typically lead to significant reworking of the packages on the final core image.

Conflict management. Even when IT organizations identify conflicts between old and new applications, they do not always have a structured method for resolving the conflicts. Packaging tools can help to quickly identify common conflicts and then aid in establishing policies for conflict resolution as well as for automating the conflict-management process.

Application suitability for packaging. Although application packaging offers several benefits, certain applications should be deployed outside an automated software distribution. Rarely is an installation package created for every application that is deployed across an organization.

Centralized packaging process. Although enterprises typically use numerous Windows-based applications, many enterprise IT organizations do not have a common organization or methodology for software packaging and deployment. Selection of a packaging tool set can facilitate this centralization.

Structured application grouping. While some applications are used across an organization, others may be useful only to a certain business unit, geographic location, or group of specialized users. By instituting a packaging process, IT organizations can establish a structure for classifying and managing diverse software used throughout a large enterprise. Typical groupings include applications core to the business, those used primarily for departments or business units, and those deployed ad hoc to small groups of users.

Formal application request and approval process. Packaging technology can help organizations implement a formal process for requesting, approving, and distributing applications. A simple workflow process can provide a way for users to request a specific application and obtain approval from the appropriate business and IT managers. This process can significantly aid in eliminating unnecessary application deployments while helping to ensure that end users receive the appropriate, predefined versions of the requested applications. This approach can lead to a reduction in the cost and complexity of support services.

Achieving efficient, cost-effective application deployment and software management

Application packaging can be an important component for efficiently managing the increased volume of software on desktop and notebook systems. By streamlining software installation, uninstallation, repair, and patching, application packaging can help reduce costs associated with each phase of the application deployment and support life cycle. In particular, application packaging is designed to reduce costs and improve efficiency during the deployment and post-deployment phases. Such benefits depend on having a stable environment from which to automatically distribute packages to PCs, using tools such as Microsoft SMS, Altiris Client Management Suite™ software, and Novell® ZENworks® software.

By enabling fast, standardized software installations, application packaging is designed to minimize desk-side visits by support staff and avoid business disruptions caused by software failure—thereby helping to reduce costs for IT support and business operations, respectively. When implemented as an IT best practice, application packaging can help create a cost-effective software repository that is in line with overall business priorities.

Jukka Kouletsis is a senior consultant with Dell Services. He has been involved in technology migrations for more than 20 years, including seven years with Dell. Jukka has a B.S. in Computer Science from Hofstra University in New York.

FOR MORE INFORMATION

Wise Resource Center:
www.wise.com/rescenter.asp

Microsoft Windows Installer:
msdn.microsoft.com/library/default.asp?url=/library/en-us/msi/setup/roadmap_to_windows_installer_documentation.asp

Deploying Dell OpenManage Server Administrator from Dell OpenManage IT Assistant 7

Dell™ OpenManage™ IT Assistant (ITA) is a central management console that can be used to monitor and manage networked systems. ITA communicates with Dell OpenManage Server Administrator (OMSA) agents installed on individual managed nodes. Hence, the first step in setting up a centralized IT Assistant management environment is the installation of these OMSA agents. IT administrators can remotely deploy the agent software from the central ITA console using Windows Management Instrumentation scripts and tasks.

BY THE DELL OPENMANAGE ENGINEERING TEAM

Related Categories:

Command-line interface (CLI)

Dell OpenManage

Dell PowerEdge servers

Microsoft Windows

Scripting

Windows Management Instrumentation (WMI)

Visit www.dell.com/powersolutions for the complete category index.

Dell OpenManage IT Assistant (ITA) supports a wide variety of advanced systems management tasks such as device discovery, software inventory, software updates, device control, reporting, event management, and system health monitoring. ITA provides these capabilities by interoperating with Dell OpenManage Server Administrator (OMSA), which is installed on each managed node.

To install OMSA on each managed node, IT administrators can remotely deploy the agent software from the central IT Assistant console using Windows Management Instrumentation (WMI) scripts and tasks. This method can also be used to deploy standard Microsoft® Windows® OS–based software or service packs that are packaged as a Microsoft Windows Installer (MSI) package or an MSI patch (MSP) from the central ITA console.

Creating a WMI installation script

The first step in deploying Dell OpenManage Server Administrator is to develop a WMI script that installs the Dell OpenManage agent MSI package. For example, Figure 1 shows a generic WMI script—written in Visual Basic Script (VBScript)—that can run a given process on a system with the proper credentials.

The script shown in Figure 1 accepts four input arguments. The first argument is the host name or IP address of the remote node on which the process must be executed. The second argument is the process that is to be executed remotely. The third and fourth arguments correspond to the credentials to be used to spawn the process. This script either outputs the process ID of the remote process upon successful execution, or it returns an error code. In the example scenario discussed in this article, this script is used to silently install the MSI package for the Dell OpenManage agent.

Note: This script assumes that the Dell OpenManage agent installation files (the MSI package and the cabinet file contents of the Dell OpenManage Installation


```
strComputer = WScript.Arguments.Item(0)
strCommand = WScript.Arguments.Item(1)
strUserName = WScript.Arguments.Item(2)
strPassword = WScript.Arguments.Item(3)
Set objSWbemLocator = CreateObject("WbemScripting.SWbemLocator")
Set objSwbemServices = objSWbemLocator.ConnectServer(strComputer, "root\cimv2", strUserName, strPassword)
Set objSWbemObject = objSWbemServices.Get("win32_Process")
errReturn = objSWbemObject.Create(strCommand,null,null,intProcessID)
if errReturn = 0 Then
   Wscript.Echo strCommand & " was started with a process ID of " _
   & intProcessID & "."
Else
   Wscript.Echo strCommand & " could not be started due to error " & errReturn & "."
End If
```

Figure 1. Example WMI script

and Server Management CD) already exist on the target system in the directory C:\tmp. If administrators need to copy the files onto the target system, they should refer to the "Copying the MSI package onto the target node" section in this article before proceeding further.

Creating a generic CLI task to run the WMI installation script

The next step for IT administrators is to execute the script shown in Figure 1 on a remote node by creating a task in IT Assistant. They must create a generic command-line interface (CLI) task to execute this script with the appropriate parameters. ITA provides the Task Creation Wizard to guide administrators through the process of creating a task, which involves the following steps.

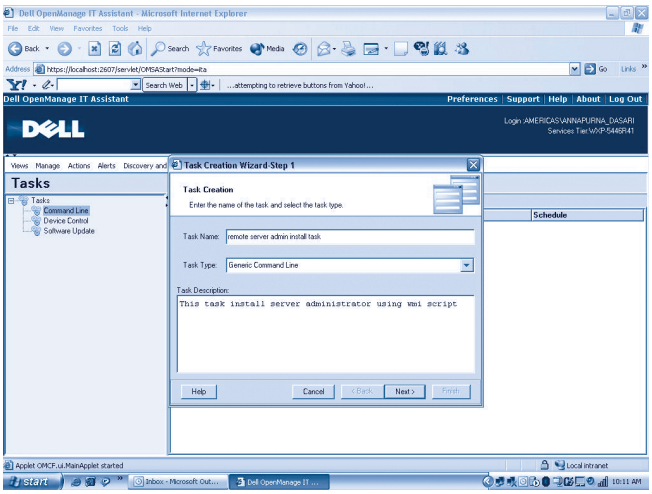


Figure 2. Entering task name and type in IT Assistant Task Creation Wizard

Step 1. Enter a name for the task, select "Generic Command Line" as the task type, and provide a brief description of the task (see Figure 2).

Step 2. Enter the name of the executable as well as the name and location of the script to be executed, followed by the arguments to the script. In the example scenario, the executable is cscript,

WINDOWS MANAGEMENT INSTRUMENTATION

Windows Management Instrumentation provides a consistent way to access comprehensive systems management information. It does so by using WMI infrastructure, which consists of the WMI providers, the WMI service, the WMI repository, and the WMI consumers.

The WMI providers act as intermediary between the WMI service (Common Information Model Object Manager, or CIMOM) and a managed resource. They request information from and send instructions to WMI-managed resources on behalf of consumer applications and scripts. The WMI service handles the interaction between the WMI consumers and the WMI providers. The WMI repository stores the schema that defines the management information exposed by WMI. The WMI consumers are scripts and applications that access and control management information available through the WMI infrastructure.

A WMI class exists to encapsulate the properties and the actions that WMI can perform to manage each manageable resource. The managed resource that is accessed in the example installation script in this article is the win32_Process class, which encapsulates properties and actions that can be performed on a process in a Microsoft Windows OS.

TYPES OF TASKS IN DELL OPENMANAGE IT ASSISTANT

The task management component in Dell OpenManage IT Assistant is designed to provide a feature-rich, robust, and secure systems management interface that can be used by system administrators to efficiently manage systems in their enterprise. IT Assistant supports execution of various types of tasks, which can be classified based on their functionality. These task types include:

- **Generic command line:** Executes generic remote command-line tasks that are not specific to any agent.
- **Remote Server Administrator command line:** Executes Dell OpenManage Server Administrator CLI commands remotely.
- **IPMI command line:** Executes Intelligent Management Platform Interface (IPMI) CLI commands remotely.
- **Remote client instrumentation:** Executes Dell OpenManage Client Connector CLI commands remotely.
- **Shut down device:** Performs a shutdown operation on a selected device.
- **Wake up device:** Wakes up a selected device via the user-specified port number. This task works by sending a "magic packet" to the target device.
- **Software update:** Performs a software update on one or more remote devices.

and it executes the WMI-based remoteexec.vbs script located in C:\scripts. The WMI script takes the following four arguments:

- **Target device:** The first argument is the host name or IP address of the target system on which the Dell OpenManage agent must be installed. Administrators can use $IP or $NAME to select devices from the device selection pane.
- **MSI installation command:** The second argument is the silent installation command for the SysMgmt.msi package located in C:/tmp on the target device. The silent installation command—`msiexec.exe /i c:\tmp\SysMgmt.msi /l*v "C:\install.log" /qn`—installs the complete MSI package and writes the log to the C:\install.log file. Administrators can pass options to install specific components of the MSI package.[1]
- **Username:** The third argument specifies the name of the user running the WMI script on the remote node.
- **Password:** The fourth argument is the password of the user running the WMI script. Administrators can use $USERNAME and $PASSWORD to pass the credentials securely using the ITA authentication pane.

Step 3. Select the target device by either choosing from the Devices tree or running a query.

Note: Administrators can select more than one target device for the task. In this case, the script will be executed on each of the target nodes with the same parameters except for the first argument, which will change depending on the device. Also, the MSI should be placed in the same location on all target devices, and the same credentials should apply.

Step 4. Select the time at which the task should run.

Step 5. Enter the credentials—username and password—with which the task needs to be run on the remote node.

Step 6. Review the summary of the task and click the Finish button if no changes are required (see Figure 3).

Task execution results

Once the task executes, administrators can view the execution results by selecting the Execution Log tab. This screen displays the standard output and the standard error of the process that executed the task on the remote node. Figure 4 shows the Execution Log screen after a task was successfully executed.

Verifying the installation

The output of the task reveals only whether the installation started successfully—it does not provide information about how and when the installation completed. Because the task is asynchronous, it spawns only the MSI installation and does not wait for the process to complete. To verify whether the installation ended successfully,

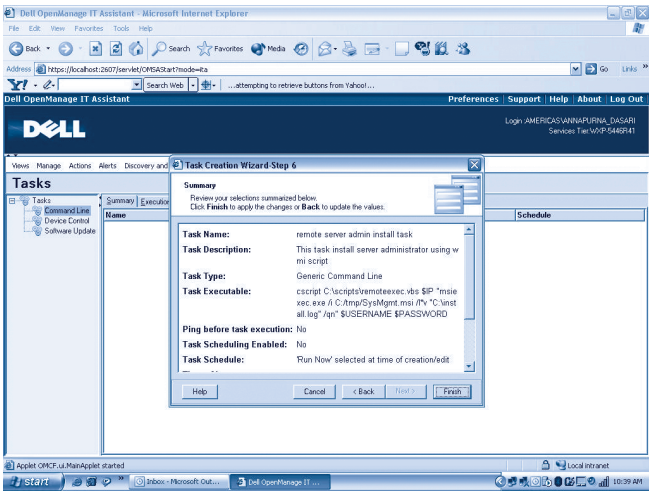


Figure 3. Task summary information from IT Assistant Task Creation Wizard

[1] For more information about other silent installation options, refer to the *Dell OpenManage Server Administrator User's Guide* at support.dell.com/support/edocs/software/svradmin.

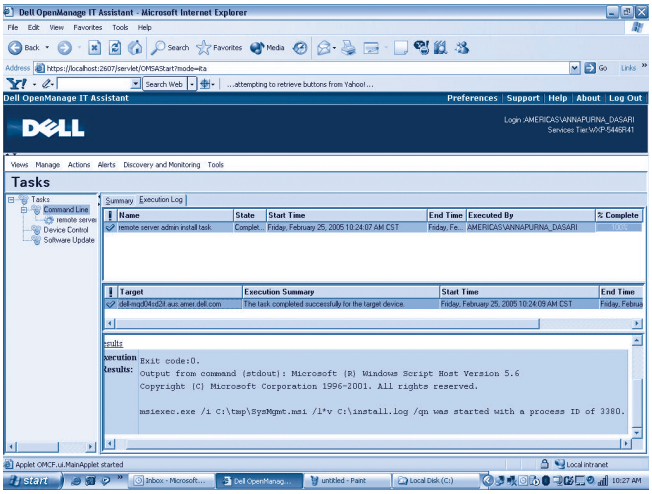


Figure 4. Execution Log screen in IT Assistant showing task execution results

administrators can wait for the installation to complete and then perform an inventory of the target node or the range that includes the target node.

Once the inventory task has been performed, the target node is moved to the Servers group in the Devices tree of the ITA console, indicating successful installation of Dell OMSA on the remote node. If the remote node does not appear in the Servers group, it is likely that the installation failed. Common reasons for installation failure include the following:

- **Invalid credentials:** The credentials passed to the WMI script are invalid or do not have the sufficient privileges to run the intended WMI task of installing the Dell OpenManage agent software.
- **Remote WMI problem:** The WMI service on the target device has not started or is not configured properly. (The default WMI configuration is sufficient for the Dell OpenManage agent installation script to work.)
- **Connectivity issues:** The management station is unable to communicate with the target node because of networking problems. Running the ping connectivity test from the troubleshooting tool can help identify and resolve such issues.
- **MSI errors:** MSI errors were generated while the SysMgmt.msi package was being installed. Such errors are written to the installation log file when administrators use the `/l` option in the silent installation command. Administrators can diagnose the failure by obtaining information about the MSI error code that is written to the installation log file on the target machine.
- **Simple Network Management Protocol (SNMP) configuration problems:** If SNMP is not configured properly on the target node, the node will not be identified as a server even if the installation is successful. Administrators can run the SNMP connectivity test from the troubleshooting tool to check whether SNMP is configured properly on the target node. They can adjust SNMP settings on the target node remotely from the management station by using a WMI script. Visit www.dell.com/powersolutions for an example script that configures SNMP settings on a target node by changing the appropriate registry keys.

Copying the MSI package onto the target node

In the installation scenario discussed in this article, the installation files exist locally on the managed node. If this is not the case in a real-world deployment, administrators need to copy the installation files from the management station (a local system) onto the remote managed node. This step is mandatory because the MSI installer may run into problems while installing a remote package. Administrators can run a WMI script that copies files onto the remote system. Visit www.dell.com/powersolutions for an example WMI script that can be used to copy files from the local system to a remote node.

This type of WMI script can also be run from ITA using a generic CLI task as shown in the "Creating a generic CLI task to run the WMI installation script" section in this article. The script takes the IP address of the target node as the single argument, which must be entered as $IP in the arguments section. Next, administrators should select the target remote node and then enter credentials corresponding to the user who has file-copying rights on the remote system. Typically, the credentials supplied are those of a domain user who has appropriate privileges on both the source and the target nodes.

Note: The WMI script discussed in the preceding example uses WMI impersonation and thus can be run only from IT Assistant 7.1 or later. If systems are running IT Assistant 7, best practices recommend running the copy script from the command prompt.

Deploying Dell OpenManage across the enterprise

Dell OpenManage IT Assistant can be used to execute a wide variety of systems management tasks such as software installation and upgrade, status monitoring, and event management. By enabling standards such as WMI, IT Assistant can help provide a viable and comprehensive systems management environment for enterprise data centers.

FOR MORE INFORMATION

Dell OpenManage:
www.dell.com/openmanage

Protecting the Microsoft Exchange Production and Development Environment with CommVault Software

Six years ago, Microsoft was looking for powerful, reliable restore capabilities for its internal Microsoft® Exchange test lab. By implementing CommVault® data management software—which is based on the Microsoft Windows® OS platform—the Microsoft team was able to perform seamless restores and enhanced data management in the lab. This article explores the challenges the Microsoft team faced and how implementing CommVault software helped address those challenges while providing significant technical and business benefits.

BY RANDY DE MENO

Related Categories:

Business continuity

CommVault

Dell PowerEdge servers

Disaster recovery

IT management

Microsoft Exchange

Microsoft

Storage software

Storage

Visit www.dell.com/powersolutions for the complete category index.

To protect critical Microsoft Exchange environments, enterprise IT organizations need reliable software tools such as CommVault Galaxy™ and QiNetix™ software. This article describes how CommVault data management software has been successfully used in the Exchange test lab at Microsoft Corporation—and thus, how IT organizations can implement CommVault software in their own Exchange environments.

Introducing the Microsoft Exchange lab

In the late 1990s, Microsoft coined the term *DogFood* to refer to code that was still being developed. In an effort to help improve the quality of Exchange code, the Exchange development team was challenged to use its own code in production before anyone else used it—a practice known as "eating their own dog food." Since then, the concept has crystallized and Microsoft Exchange undergoes extensive production testing long before a new Exchange version becomes a release candidate, which in turn benefits all Exchange users.

Hardware and software components in the lab

The Microsoft Exchange lab uses a mix of servers, network attached storage (NAS), and a storage area network (SAN) that includes Dell, HP, EMC, and Xiotech hardware. The Microsoft software used in the lab includes the Microsoft Windows OS as well as SharePoint®, SQL Server™, Microsoft Operations Manager (MOM), and Live Communications Server (LCS) software.

Before implementing CommVault software in the Exchange lab, the Microsoft team had to manually transport the NTBackup, EMC® Legato®, and VERITAS® Backup Exec™

tapes back to the test lab and manually load them each morning when the new, up-to-date data was required. This process resulted in questionable reliability and unacceptable work hours for test engineers dedicated to this task. It also reduced the time other team members could work with the software in production. In addition, administrators in the Exchange lab typically perform 5 to 30 full database restores per week, and before implementing CommVault software, the lab experienced a high failure rate for these restores.

Users and test environments in the lab

The Microsoft Exchange Products Team, the Microsoft Office Products Team (which has thousands of production mailboxes), and various Microsoft employees and executives store their live production data and e-mail in the DogFood lab on the Microsoft campus in Redmond, Washington. Microsoft simulates numerous customer environments for implementing Exchange and conducts in-depth testing in these environments. The DogFood lab is a combination of the following environments:

- Systems running mixed releases of the Windows OS and Exchange
- Servers with numerous mailboxes (some greater than 10 GB) and a large amount of data
- Servers with numerous mailboxes and a small amount of data
- Servers with a few mailboxes and a large amount of data
- Servers with a few mailboxes and a small amount of data
- Remote management (from the Redmond campus) of servers that reside in Europe and other locations throughout the world

Examining Microsoft's goals for improving the lab

In 1999, Microsoft invited CommVault to test its Windows-centric data management software in the Exchange lab. CommVault had recently made a strategic decision to base its software on a Windows platform because CommVault software was designed to provide the performance and reliability required to manage large data centers. The challenges that Microsoft invited CommVault to address were as follows:

- Create a robust storage environment
- Enable reliable backups and restores
- Use software that is flexible enough to leverage new hardware and software technologies
- Enable extensive reporting capabilities
- Provide global access to the Exchange environment from anywhere in the world

In addition, CommVault faced the challenge of managing Exchange in an environment that serves key Microsoft executives as well as the engineers that develop the Exchange software. The following sections explain how CommVault addressed Microsoft's challenges.

Requirement 1: Remote management

The Microsoft administration team needed a robust enterprise data management solution designed to provide world-class remote management capability and the flexibility to support Microsoft's frequent Windows OS and Exchange builds. New, cutting-edge Microsoft technologies are tested so often that they may include versions of operating systems and applications that are 6 to 18 months ahead of when they are expected to be available as release candidates to customers. Over the years, these products have included:

- New versions of the Microsoft Windows Server™ 2003 and Windows 2000 Server operating systems
- New versions of Windows Server and Windows Storage Server
- Exchange Server 2003 databases with support for Virtual Shadow Copy Service (VSS) snapshot software
- The upcoming version of Exchange (code-named "E12"), SQL Server 2003, and SQL Server 2005
- Microsoft Active Directory® information
- SharePoint data
- MOM servers
- LCS data
- Windows File System data

Requirement 2: Replication

Based on the performance of the CommVault software during its initial 12 months of deployment in the DogFood lab, Microsoft requested that CommVault replicate the production data to Microsoft's DogFood simulation testing lab—known as the SIMULATED lab. This replication was designed to help the Exchange team's testing efforts by making live production data available to the various Microsoft engineers who test Exchange software. The SIMULATED test lab helps accelerate the overall build process for Exchange.

Before using CommVault software, the Microsoft Exchange team was unable to restore production data to the SIMULATED test lab without manual intervention. The SIMULATED lab is connected to the DogFood lab via a private network, and the test machines in the SIMULATED lab have the same Domain Name System (DNS) names, Microsoft Active Directory forests, and Exchange Server names as their cloned cousins in production. Before CommVault software was implemented, the client names in the two labs could not be differentiated. Microsoft administrators tested numerous backup solutions, but none could address this problem. The only solution at the time was to manually bring tapes into the SIMULATED lab for restores—which required a team of dedicated

administrators working excessive hours, including staying late to change tapes. Time-consuming backups and restores made it hard for administrators to perform their primary responsibilities—maintaining the production environment and finding bugs in Exchange. Consequently, Microsoft needed to find a way to reduce the time required to create the shadowed SIMULATED lab while enhancing the performance and reliability of the restores.

CommVault addressed these challenges with its CommVault QiNetix software, which differentiates the two sets of servers by giving a unique name to the servers in the SIMULATED lab. The dual-homed CommVault CommServe system connects the two large networks. By leveraging these capabilities, the CommVault software enabled administrators to spend less time on backups and restores and more time on their core competencies as test engineers. Furthermore, CommVault software supports VSS without requiring manual intervention by administrators.

Requirement 3: Migration

In late 2004, after five years of successful deployment, the entire DogFood production environment moved to the internal Microsoft IT production network. Today, Microsoft moves production data to the SIMULATED lab using the CommVault CommCell™ migration capability. The CommCell migration capability enables administrators to restore data from one CommServe-based system to another.

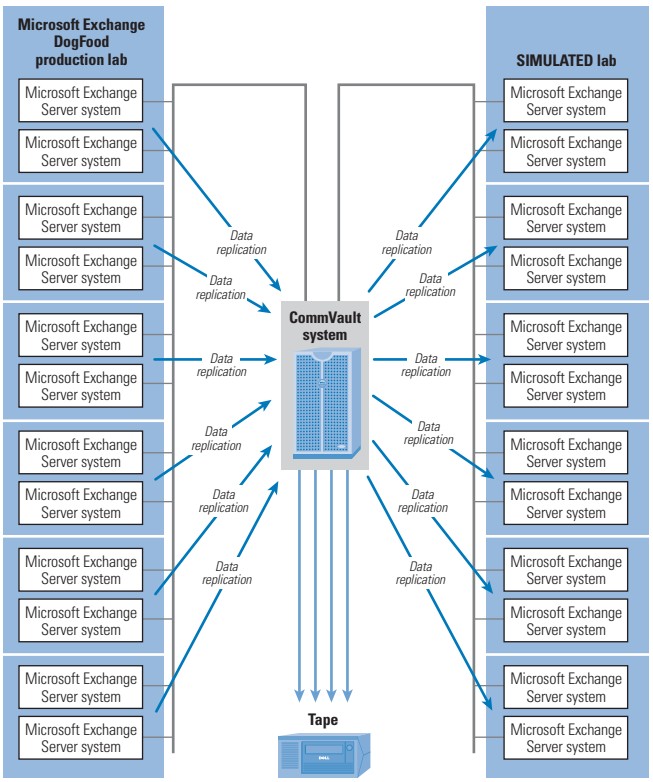


Figure 1. Microsoft Exchange DogFood lab managed with CommVault software

Implementing the CommVault solution

Microsoft enables anytime, anywhere management of its DogFood lab by using one Windows-based server (the CommServe system) with five CommVault MediaAgents. The MediaAgents are a mix of high-performance SAN storage and DLT7000 (digital linear tape) storage. Figure 1 shows the CommVault system's role within the Microsoft lab.

For more than six years, the Microsoft team has been using CommVault Galaxy software to back up data in the lab. Each night over 1 TB of full and incremental backup data—including Exchange Server 2003 databases, Windows Server 2003 data, SQL Server 2000 data, SharePoint data, MOM data, and domain controller data (such as system state and Active Directory service data)—is backed up. Disk-to-disk-to-tape (D2D2T) management is used: backups are stored to a SAN and then migrated to tape. The implementation of D2D2T backups has significantly enhanced the performance and reliability of the lab. Since Galaxy has been deployed, the Microsoft team has not reported experiencing a single instance in which administrators were unable to restore data from a backup. Although initial software and hardware releases have experienced occasional problems, the Microsoft team works with vendors to troubleshoot and resolve any flaws before allowing products to be deployed to joint customers.

Since Galaxy has been deployed, the Microsoft team has not reported experiencing a single instance in which administrators were unable to restore data from a backup.

In addition, CommVault QiNetix enables administrators to hot swap a cloned server to production within minutes if a system fails. This capability is designed to save time and help make data highly available.

CommVault software relies on a fully-unified, modern code base that is powerful yet cost-effective and easy to use. This code base leverages the features and power of the Windows Server OS. Supported by Dell and Microsoft, CommVault software is built on a Common Technology Engine that moves, manages, and catalogs data.

CommVault's policy-based data management software is designed to provide the Microsoft Exchange environment with anywhere, anytime management. This includes support for Microsoft's cutting-edge products and technologies as well as

simple data management using disk as the primary media target. In addition, the CommVault software helps manage the Microsoft DogFood and SIMULATED environments from a single Windows-based server.

Technical benefits of CommVault software

CommVault software enables Microsoft and other environments to benefit from numerous features and capabilities:

- **Simplicity and ease of use:** CommVault software helps make installation, configuration, and management easy for novice administrators to learn.
- **Web browser–based management:** Administrators can perform complete management and maintenance tasks from virtually any location—whether the office, a hotel, or home—that has Web access to the CommServe-based system. This capability is designed to save administrators time and help ensure that data is managed properly.
- **Reliable restores:** CommVault software has a track record of high reliability—the Microsoft lab has not reported experiencing a failed restore in over six years.
- **Cutting-edge hardware and software technology:** CommVault software has continually supported Microsoft operating systems and applications early in the development and beta cycles and throughout production, including tests of many diverse customer environments. CommVault engineers and developers have worked closely with the Microsoft team to provide continuous data management support for all versions—early-build, alpha, beta, release to manufacturing (RTM), and production—of Microsoft's next-generation operating systems and applications, while CommVault software runs in a production environment at Microsoft.
- **Support of combined releases:** Microsoft customers often have mixed revisions in their networks, and Microsoft often runs mixed versions and revisions of Exchange and Windows.

Business benefits of CommVault software

Microsoft administrators can benefit from the reporting capabilities of CommVault software, which enables them to manage the environment from the lab, the office, home, or a text-messaging device. Recent enhancements to the CommVault software also enable administrators to use MOM to monitor the overall enterprise. These reporting capabilities help the Microsoft lab save time and money. Figure 2 explains additional business benefits and key performance indicators for these benefits.

Key benefits	Key performance indicators
Reliable protection	The Microsoft team has not reported experiencing a restore failure during the six years it has been running CommVault software.
Minimized cost	A single Windows-centric CommVault system is designed to manage multiple servers and applications, two disparate networks, and multiple terabytes of data.
Seamless business continuity	Service-level agreements (SLAs) for CommVault software have been met ahead of time for the past six years. Before CommVault software was implemented, replicating the environment was a challenge.
Cross-application support	Exchange, SharePoint, SQL Server, Active Directory, Windows Storage Server, MOM, and other Microsoft applications are used in production.
Ease of use	Web browser management from across the building or from home is possible—with easy-to-read reporting that can be sent to a Windows Mobile phone or a BlackBerry device.
Fast Exchange database restores	Complete, reliable restores can be performed in minutes.
Support for next-generation builds of Windows and Exchange	CommVault supports releases of Windows and Exchange that are often 6 to 12 months away from becoming release candidates.
Assistance in rebuilding and adding Windows-based servers	CommVault software offers the 1-Touch System Recovery option, which leverages Windows Preinstallation Environment—allowing administrators to recover an entire system quickly and easily by inserting a CD that links system information from the CommVault database.

Figure 2. Business benefits of CommVault software

Backing up Exchange in a demanding IT environment

Six years ago, Microsoft was looking for a solution that would provide world-class restore capabilities and a fully integrated user interface. CommVault now serves as the software vendor providing data protection for the Microsoft Exchange DogFood lab and for the Microsoft IT organization. The integration of CommVault software with the Windows platform and CommVault's experience working closely with Microsoft has enabled powerful, reliable data management and restore capabilities in the Exchange environment.

After a six-year record of seamless restores in the Microsoft Exchange DogFood lab, CommVault's cutting-edge data management software has incorporated global management and high-quality reporting features, further enhancing enterprise data management capabilities.

Randy De Meno is the director of advanced applications and the Microsoft partnership at CommVault Systems, Inc.

FOR MORE INFORMATION

CommVault case study for the Microsoft Exchange DogFood lab:
www.commvault.com/microsoft/case_study.asp

Promoting E-Mail Security on Dell Servers Using

Symantec Mail Security 8200 Series Appliances

Symantec® Mail Security 8200 Series appliances are designed to protect e-mail systems and enhance efficiency by integrating volume management, attack prevention, content filtering, encryption, and archiving in a single appliance that is easy to deploy and administer from a Web-based console. The Symantec Mail Security 8200 Series leverages the Dell™ PowerEdge™ 850 and PowerEdge 1850 server platforms.

BY FAREED BUKHARI

Related Categories:

Dell PowerEdge servers

E-mail technology

Network security

Security

Symantec

Visit www.dell.com/powersolutions for the complete category index.

In many enterprises, e-mail constitutes a business-critical channel for internal and external communication. The key challenge facing administrators is how to preserve the value of e-mail in the face of escalating e-mail security threats, which can compromise valuable information assets, consume mail-server and storage resources, and drain staff productivity. At the same time, IT administrators are under intense regulatory and organizational pressure to monitor and control inbound and outbound e-mail content effectively.

Symantec Mail Security 8200 Series appliances are designed to secure the e-mail gateway and streamline the e-mail infrastructure for medium-to-large enterprises. This approach enables a high level of spam and virus protection through the integration of Symantec Brightmail AntiSpam and Symantec AntiVirus™ technologies. Symantec Mail Security 8200 Series appliances integrate volume management, attack prevention, content filtering, encryption, and archiving in a single appliance that is easy to deploy and administer. Resulting business benefits can include the following:

- **Threat protection:** Sophisticated detection and filtering technologies help keep unwanted content out of the e-mail stream.
- **Simplicity:** Integrated features enable easy installation and centralized management.
- **Rapid response:** Continuous updates from Symantec Security Response centers around the world enable rapid response to help ward off threats 24/7.
- **Content compliance:** Support for compliance helps meet government regulations and enforce enterprise policies.

Comprehensive protection through integration

The Symantec Mail Security 8200 Series is Symantec's response to e-mail security problems. These appliances are designed to offer comprehensive e-mail security, combining threat protection, administration, and ease of use. Employing innovative Symantec e-mail firewall technologies, Symantec Mail Security 8200 Series appliances are designed to reduce e-mail infrastructure costs by helping to restrict high-volume e-mail attacks. In addition, the Symantec Mail Security 8200 Series provides a powerful set of tools that help monitor content compliance.

Spam protection. Accounting for a substantial percentage of e-mail traffic, spam is notorious for clogging the message infrastructure, sapping mail server and storage resources, and cluttering in-boxes. Moreover, offensive and fraudulent spam can create liability issues for organizations. Multilayered antispam protection—maintained by Symantec Security Response—is the cornerstone of Symantec Mail Security 8200 Series appliances (see Figure 1). Powered by Symantec Brightmail AntiSpam technologies, the filtering engine harnesses an arsenal of filtering techniques and targeted filters to help address complex spam attacks. Symantec Mail Security 8200 Series appliances are designed to perform the following functions:

- Trap spam by exploiting more than 20 filtering techniques
- Leverage a global spam-analysis and response infrastructure to help identify and stop spam originating in foreign countries
- Virtually eliminate "false positives"
- Update antispam protection automatically, without frequent tuning and filter training to maintain effectiveness and accuracy

Virus protection. Virus damage can range from e-mail server crashes and system downtime to the destruction of business-critical enterprise data. Countless virus incidents are initiated by Internet-delivered e-mail. Furthermore, the payload of many viruses and worms includes software that installs an open proxy on the target server. In turn, these proxies can convert the computer into a spam relay, which spammers can command to generate even more spam. Given the substantial damage that can result from viruses, many IT organizations are considering ways to employ virus protection at the earliest point of network entry—the e-mail gateway.

Symantec Mail Security 8200 Series appliances scan and detect viruses by integrating Symantec AntiVirus technology. Antivirus protection includes automatic virus-definition updates, flexible policies to handle messages with viruses, and specific defenses against mass-mailing worms and the associated spawned e-mail messages. To provide rapid, around-the-clock virus protection, Symantec Mail Security 8200 Series appliances rely on the Symantec Security Response team. This group of experts works to identify and neutralize viruses before they can enter the network and spread across the enterprise.

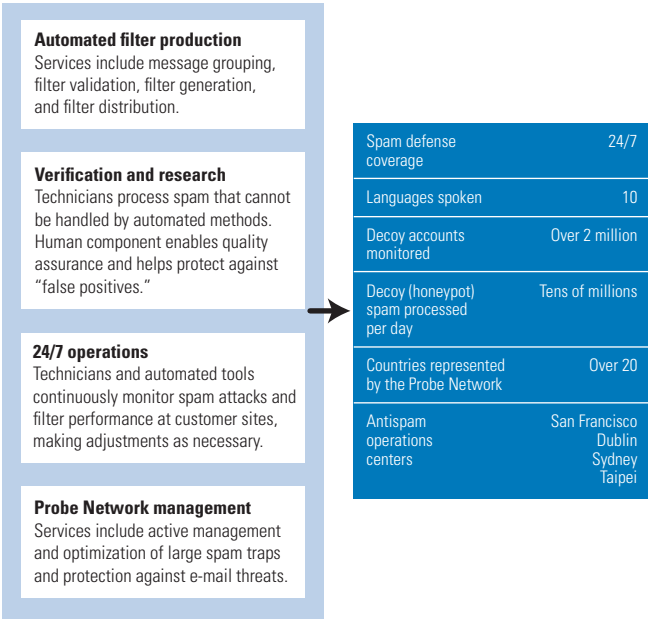


Figure 1. Antispam services from Symantec Security Response

The Symantec Mail Security 8200 Series appliances are designed to include the following antivirus features and technologies:

- A rapid, reliable scanning and repair engine
- Maximum uptime during definition updates
- Heuristics and variable scanning levels for aggressive scanning
- Choice of actions based on policies
- Cleanup of mass-mailer messages

Volume management and attack protection

A solid combination of antispam and antivirus protection is just one part of a comprehensive e-mail security solution. An e-mail security solution must also curtail unnecessary and invalid incoming e-mail traffic. Such initial volume reduction is important because scanning e-mail for spam and viruses is an inherently resource-intensive task that can drain vital system and administrator resources.

Stopping potential attacks is another compelling reason to ensure that certain mail does not breach the gateway. A representative example is the directory harvest attack—an abusive tactic that can result in huge e-mail volumes and compromise an enterprise's e-mail directory information. In these attacks, spammers send thousands of empty messages to mail servers to obtain legitimate e-mail addresses. Tracking which addresses are not rejected enables spammers to deduce valid e-mail addresses that they can use in future spam or phishing campaigns.

Protection against phishing attacks. Phishing refers to the practice of disseminating fraudulent e-mail messages that appear to originate from a legitimate company's Web site or domain

address, but in fact do not. In reality, malicious senders hijack the company's name to entrap existing and potential customers, often to harvest personal information. Symantec Mail Security 8200 Series appliances help mitigate fraudulent e-mail messaging and IP-address spoofing through antifraud and anti-phishing URL filters. The Symantec Mail Security 8200 Series also supports the Sender Policy Framework (SPF) collaborative, wherein organizations publish a list of their approved e-mail servers in the Domain Name System (DNS).

E-mail firewall

Best practices recommend an e-mail security solution that can accurately reject unwanted mail at the gateway based on the IP address. Such Simple Mail Transfer Protocol (SMTP) connection management features can be an effective way to help address issues related to ceaseless increases in e-mail volume.

Symantec Mail Security 8200 Series appliances feature a powerful e-mail firewall—a set of automated and configurable connection-management features that are designed to deploy as soon as an incoming connection is detected. Acting as a first-line defense, the e-mail firewall can be a gatekeeper that helps protect the CPU-intensive components of the filtering engine, including the antispam and antivirus layers. The e-mail firewall can be automatically configured to block spam attacks, directory harvest attacks, connections from senders identified as spammers by Symantec, and more. Symantec draws upon comprehensive resources to help identify and characterize e-mail sources accurately based on the Symantec Probe Network, which leverages more than 2 million decoy e-mail accounts.

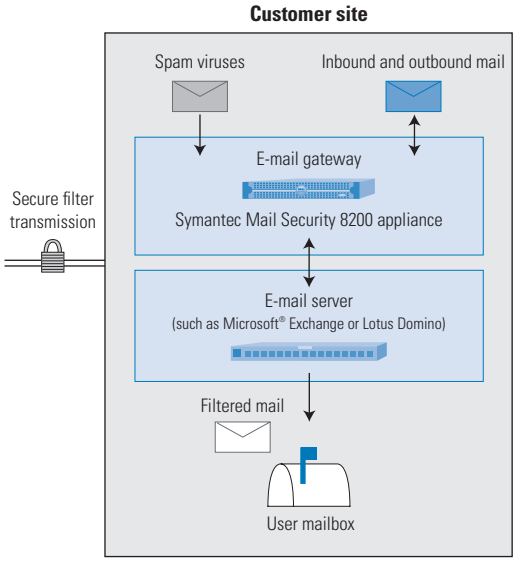


Figure 2. Symantec Mail Security 8200 Series appliance located at Internet gateway between internal mail server and external firewall

Content filtering

If an enterprise lacks an effective scanning system to monitor message content, confidential, offensive, or prohibited outbound e-mail traffic can escape in a matter of seconds. For certain organizations, this can negatively affect competitiveness and business position. On the inbound side, many enterprises operate under mandates to maintain a hospitable workplace. In addition, messaging administrators typically seek to prohibit large multimedia attachments from clogging network bandwidth.

When used as part of the mail policy and management process, the content-compliance features in Symantec Mail Security 8200 Series appliances permit administrators to enforce corporate e-mail policies, helping to reduce legal liability and helping to ensure regulatory compliance. Content-compliance features are designed to enable the following functions:

- Scan mail against predefined or custom dictionaries and keyword lists to help enforce organizational content policies
- Create custom filters to further enforce company policies
- Minimize the load on e-mail servers and filtering resources by managing inbound attachments
- Attach disclaimers or annotations to e-mail messages

Complementary security enhancements

The Symantec Mail Security 8200 Series appliance is designed to reside at the perimeter of an e-mail network and help safeguard existing downstream servers (see Figure 2). The OS, mail transfer agents, and all necessary product software are preinstalled on the appliance. As a result, these appliances are designed to plug seamlessly into existing environments to streamline deployment, administration, and archiving.

For organizations that require encrypted connections between mail servers, Symantec Mail Security 8200 Series appliances support encrypted connections using Transport Layer Security (TLS). Administrators can choose whether TLS is permitted or required for all the appliances in the network.

In addition, Symantec recently launched a program to help ensure that complementary, best-of-breed products are compatible with Symantec Mail Security 8200 Series appliances. For example, solutions for encryption, content security, regulatory compliance, and zero-day threat protection have been tested and validated for compatibility by software vendors.[1] To date, Authentica, Avinti, PGP Corporation, Voltage Security, Vontu,

[1] For more information about these vendors, visit www.symantec.com/partners/technology/list_compatible.html.

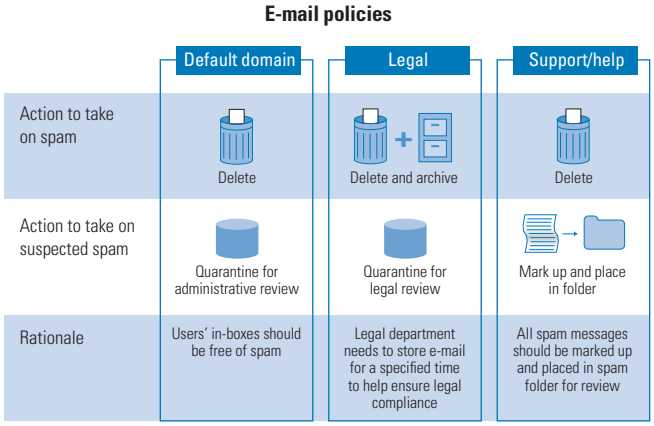


Figure 3. Flexible e-mail policies enabled by the Web-based Control Center console

and ZixCorp have joined the message-security compatibility track within the Symantec Technology Partner Program.

Flexible application of e-mail policies. Powered by automatic filter updates, Symantec Mail Security 8200 Series appliances are designed to run easily in "hands-off" mode. However, the appliances are flexible enough to accommodate administrators who want granular control of message filtering and mail-handling policies. This balance between automation and control is made possible by the Control Center—a Web-based administration, configuration, and reporting console that enables centralized, secure management as well as delegated administration.

As shown in Figure 3, Control Center allows Symantec Mail Security 8200 Series appliances to handle e-mail flexibly using both preset and customizable filtering policies that enable features such as the following:

- Easy group designations
- Multiple mail categories
- Multiple action options
- Web-based quarantines
- Detailed reporting
- Easy software updates
- Robust system monitoring and maintenance

Automatic filter updates from Symantec Security Response. To help ensure comprehensive protection against emerging e-mail threats, Symantec continuously fortifies the defenses of Symantec Mail Security 8200 Series appliances in real time, automatically incorporating the latest antispam filters and antivirus definitions. Symantec can initiate and install filter updates to relieve administrators of this administrative burden. The filter download procedure incorporates two-way validation to help guarantee that updated filters and antivirus definitions originate from Symantec. In addition, the appliance continues filtering e-mail even as it is being updated.

Collaboration between the Dell OEM Industry Solutions Group and Symantec

Symantec e-mail security appliances are designed to operate across Dell PowerEdge 850 and PowerEdge 1850 server platforms. The Symantec products featured in this article are enabled by industry-standard Dell architecture and the comprehensive original equipment manufacturer (OEM) capabilities of the Dell OEM Industry Solutions Group, which include regulatory and compliance assistance, product transition management, and flexible fulfillment and logistics capabilities. The ability to customize Dell servers, as demonstrated by the Symantec e-mail security appliances, is available exclusively through the Dell OEM team and is designed to support a variety of software, telecom, networking, surveillance, and medical equipment applications.

Invoking in-depth technical experience and dedicated product and software design resources, the Dell OEM team provides detailed component-level planning information to help OEM customers align their product release schedules with Dell's product roadmaps. For more information about the Dell OEM Industry Solutions Group, visit www.dell.com/oem/appliance.

Comprehensive e-mail security

Symantec Mail Security 8200 Series appliances integrate a broad range of e-mail security technologies. Symantec Brightmail AntiSpam technology is incorporated into these appliances to combine more than 20 spam-prevention techniques, thereby enabling highly effective and accurate antispam capabilities. Building upon Symantec AntiVirus technology, the real-time scanning feature in Symantec Mail Security 8200 Series appliances helps protect against e-mail viruses, while a mass-mailer cleanup tool is designed to automatically remove e-mail messages associated with worms. Innovative e-mail firewall technologies—which include a powerful suite of SMTP connection-management and attack-detection features—can contribute to a reduction in e-mail infrastructure costs by restricting connections from spam-sending servers. Furthermore, the appliance's all-in-one form factor and intuitive mechanism for updating operating systems, mail transfer agents, and supporting software help ease deployment and management.

Fareed Bukhari is a product manager in the Symantec Mail Security group.

FOR MORE INFORMATION

Symantec Mail Security solutions:
ses.symantec.com/esa

Architectural Considerations for Creating High-Availability

VMware VirtualCenter Infrastructures

When implementing and maintaining a high-availability VMware® virtual infrastructure running on Dell™ PowerEdge™ servers and Dell/EMC storage, organizations should follow best practices and use key architectural design models. This approach can enhance management efficiency and help maximize uptime, resource utilization, and scalability in a virtualized data center environment.

BY SCOTT STANFORD, SIMONE SHUMATE, AND BALASUBRAMANIAN CHANDRASEKARAN

Related Categories:

Dell PowerEdge servers

Dell/EMC storage

High availability (HA)

Storage

Storage area network (SAN)

Virtualization

VMware

Visit www.dell.com/powersolutions for the complete category index.

Server virtualization technology is maturing rapidly. As its benefits are demonstrated in a wide range of noncritical test and development scenarios, server virtualization is expected to expand into myriad production environments. When configuring virtual servers in a production environment, IT organizations may benefit from the many best practices that have been developed for traditional IT architectures. In fact, traditional IT best practices are expected to be extremely useful as they evolve to address virtualized IT environments.

Best IT practices for backup and recovery as well as high-availability (HA) strategies can be followed when implementing a virtualized IT environment. For example, in a VMware-based virtual IT infrastructure, systems running VMware ESX Server™ software can be backed up using traditional approaches for backup software, and virtual machines (VMs) residing on ESX Server–based hosts can be clustered using traditional clustering software.

Similar best practices should also be applied when implementing VMware VirtualCenter software, a key component of the VMware virtual infrastructure. VirtualCenter is an application that runs on the Microsoft® Windows® OS and is designed to manage hundreds of ESX Server–based hosts from a central console. Using the VirtualCenter interface, administrators are enabled to create, deploy, and clone VMs; obtain status and performance metrics for ESX Server–based systems; and execute dynamic migration of VMs between different physical ESX Server–based hosts using VMware VMotion™ technology.

Because VirtualCenter relies heavily on an operational database to support and maintain the ESX Server–based hosts and the VM configuration and performance data, implementing this database in an HA configuration can enhance the flexibility, uptime, and scalability that a virtualized enterprise infrastructure is designed to provide. However, the VirtualCenter database is not the only

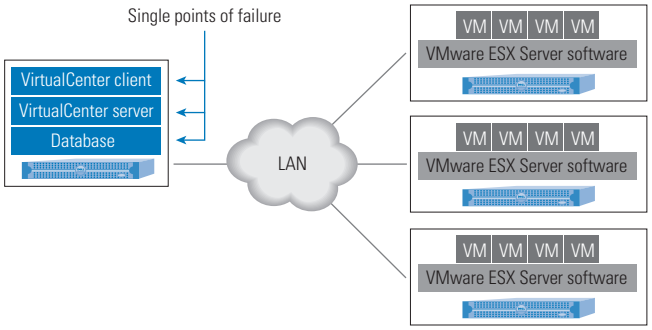


Figure 1. Stand-alone VirtualCenter model

potential single point of failure. Operations and tasks can be scheduled from VirtualCenter for action on VMs or ESX Server–based hosts; if a VirtualCenter server fails, the ensuing downtime may lead to an undesirable impact on business operations. By using key architectural design models and following IT best practices, organizations can enable successful implementations designed to maintain a high-availability VMware virtual infrastructure.

Examining the VirtualCenter design models

IT administrators can choose from four key architectural design approaches for deploying VirtualCenter: the stand-alone model, the distributed model, the HA model, and the virtualized clustered HA model. While the stand-alone model does not represent a true HA system, it provides for a rigorous backup and archiving method that may be well-suited for small data centers or initial virtualization deployments. While the distributed model also does not represent a true HA system, its two-node design for VirtualCenter and its database enables availability enhancements over the stand-alone model.

As business demands grow and virtual infrastructure requirements expand, implementing true HA models can help ensure maximum uptime and resource utilization in a virtual IT infrastructure. For example, one HA VirtualCenter design model uses an active/passive HA configuration for the VirtualCenter component stacks, while another HA model implements a clustered virtual infrastructure model for VirtualCenter and its database.

Stand-alone VirtualCenter model

The stand-alone model for VirtualCenter, shown in Figure 1, is the simplest of the four models to design and implement. In this configuration, VirtualCenter software stack components reside on the same physical server. However, this centralized design means that individual stack components and the underlying server are all potential single points of failure.

The VirtualCenter client can be an important tool for configuring and managing ESX Server–based hosts and VMs, but it does not have to be running for normal VirtualCenter operations to occur. Because the VirtualCenter client is the primary interface for the VirtualCenter server, a failed VirtualCenter server or database component stack can render the VirtualCenter client ineffective until the failed component is recovered.

Another disadvantage of the stand-alone model is the downtime encountered if a VirtualCenter server or database system fails. Depending on the type of failure, recovery time could involve not only the database, but also the systems that host the VirtualCenter software stack components. For example, if a bare-metal VirtualCenter software stack must be restored, it may take administrators several hours to reinstall the OS, database application, and VirtualCenter software and then to restore connectivity to the managed ESX Server–based host and VM resources. During the recovery and restoration period, the ESX Server–based infrastructure could be left in an unmanaged state.

Distributed VirtualCenter model

While supporting the same component stack specified for the stand-alone model, the distributed VirtualCenter approach enables a higher degree of availability than the stand-alone model by creating a tiered relationship between the database server and clients. In this distributed scenario, two servers support the VirtualCenter infrastructure—the VirtualCenter stack components (the client and server) and the database reside on completely separate systems. The VirtualCenter server communicates with the database server using an Open Database Connectivity (ODBC) connection. ODBC is a method that enables applications to communicate with a database server.

A distributed architecture helps reduce the risk of losing the entire VirtualCenter infrastructure, but it does not yet create a highly available infrastructure if hardware or software components fail. As Figure 2 shows, single points of failure still exist in this model. A failed database or VirtualCenter server still leaves the ESX Server–based systems and the VMs they host in an unmanaged state (VirtualCenter will not be operational) until the problem is resolved. In addition, if VM template files reside on a failed VirtualCenter

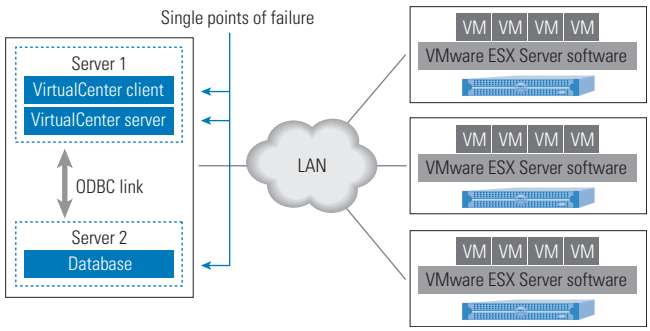


Figure 2. Distributed VirtualCenter model

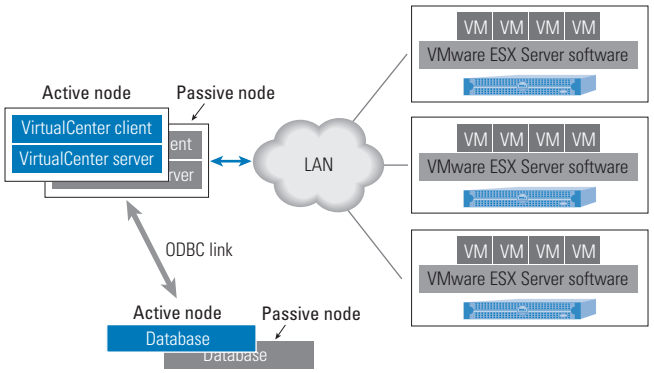


Figure 3. High-availability VirtualCenter model

server, those template files will not be available for scheduled cloning tasks until the VirtualCenter server is recovered. And if the VirtualCenter server is the sole repository for the template files, then master VM template files may not be recoverable unless the VirtualCenter server is regularly backed up.

High-availability VirtualCenter model

In addition to what the stand-alone and distributed VirtualCenter models offer, the HA model is designed not only to reduce the risk of losing the entire VirtualCenter infrastructure, but also to help eliminate the possibility of downtime if certain hardware or software components fail. Figure 3 shows an HA VirtualCenter model that is designed to eliminate single points of failure.

If one or both of the active VirtualCenter or database nodes fail, the corresponding passive node for each is designed to immediately assume operations, enabling the VirtualCenter stack components (the client and server) to maintain connectivity. This resilient configuration helps ensure maximum uptime without interruption in service for ESX Server–based hosts and VMs during the recovery period in which replacement passive nodes are brought online.

For an HA system to provide maximum uptime, however, a robust and fault-tolerant storage system—for example, a storage area network (SAN) based on a Dell/EMC CX series storage array—is required. Dell/EMC storage arrays such as the CX300, CX500, and CX700 can be used as SAN platforms and are well-suited for supporting mission-critical database systems that house data for vital applications such as VirtualCenter. In addition, an IT organization deploying an HA VirtualCenter model can enhance the capabilities of its SAN infrastructure by using the advanced VMotion feature of VirtualCenter. VMotion is designed to move VMs from one physical ESX Server–based host to another physical ESX Server–based host while maintaining VM and application functionality. By deploying an HA VirtualCenter model and using VMotion technology, organizations may enable maximum uptime and functionality from their VMware ESX Server–based virtual server deployments.

Virtualized clustered high-availability VirtualCenter model

Some or all of the VirtualCenter components described in the preceding sections can be virtualized. For example, the VirtualCenter server can be installed on two VMs that are clustered together for high availability. Similarly, database components can be installed on clustered VMs.

Administrators may consider several options for clustering in virtualized data center environments. For example, the cluster can comprise two VMs on a single ESX Server–based physical server, two VMs on two different ESX Server–based physical servers, or one cluster node deployed as a VM on an ESX Server–based physical server and the other cluster node deployed as a physical server. By deploying VMs in a cluster configuration, IT organizations may achieve inherent advantages enabled by virtualization technology—that is, normalized performance, OS and application isolation, resource consolidation, and management efficiency.

Building a resilient virtual infrastructure

A high-availability VMware VirtualCenter configuration can be designed to enable an always-on virtualized enterprise IT infrastructure. By understanding and implementing best-practices methods and key architectural design models for virtualized data center environments, IT organizations may enhance management efficiency and help ensure maximum uptime, resource utilization, and scalability for their enterprise infrastructures.

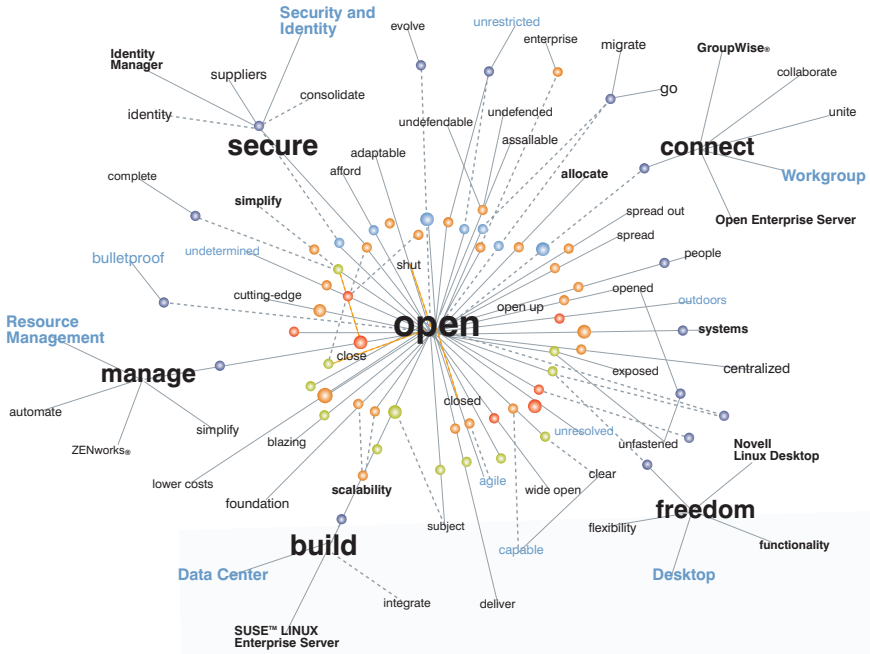
Scott Stanford is a systems engineer in the Scalable Enterprise Computing team within the Dell Enterprise Solutions Engineering Group. His current focus is on performance characterization and sizing for virtualized solutions. Scott has a B.S. from Texas A&M University and an M.S. in Community and Regional Planning from The University of Texas at Austin, and he is pursuing an M.S. in Computer Information Systems at St. Edward's University.

Simone Shumate is a senior systems engineer in the Dell Enterprise Solutions Engineering Group, where she leads the Scalable Enterprise Computing team. She has a B.S. in Computer Engineering from the University of Kansas.

Balasubramanian Chandrasekaran is a systems engineer in the Scalable Enterprise Computing team within the Dell Enterprise Solutions Engineering Group. His research interests include virtualization of data centers, high-speed interconnects, and high-performance computing. Balasubramanian has an M.S. in Computer Science from The Ohio State University.

FOR MORE INFORMATION

Dell and VMware:
www.dell.com/vmware



Define Your Open Enterprise.™

What does Open mean to you? Community? Security? Risk? Reward? Can it leverage legacy systems? Consolidate and simplify? Do you believe in its power and potential?

Introducing Novell® software for the open enterprise™ — the only software that makes Open work for you. From desktop and data center to identity management, resource management and collaboration, our flexible combination of open source and commercial software delivers more than you ever imagined. The power to automate IT asset management. Freedom from single vendor lock-in. Security that keeps the right information safe and the right people informed. And the ability to connect people to performance and business to possibilities. So you can build an open enterprise that makes sense for you — and your future. This is Novell software for the open enterprise. The Open you've wanted all along.

Novell®

This is your open enterprise.™

www.novell.com