DELL POWER SOLUTIONS · FEBRUARY 2005

DELL™

FEBRUARY 2005 · $12.95

POWER SOLUTIONS

THE MAGAZINE FOR DIRECT ENTERPRISE SOLUTIONS

Inside This Issue:

Driving data center density with the new Dell PowerEdge 1855 blade server

Deploying and managing Oracle Database 10*g* clusters

Building high-performance computing clusters with NPACI Rocks

Tooling Up to Manage Change

How Dell OpenManage 4 streamlines change management through industry-standard technologies

Tooling Up to Manage Change

Gigabit Ethernet
Multi-Gigabit Ethernet

DELL POWER SOLUTIONS
THE MAGAZINE FOR DIRECT ENTERPRISE SOLUTIONS
FEBRUARY 2005

EDITOR'S COMMENTS

EXECUTIVE INSIGHTS

NEW-GENERATION SERVER TECHNOLOGY

SYSTEMS MANAGEMENT

COVER STORY | PAGE 10

Managing Change Through Industry-Standard Technologies

By Paul Laster

Dell OpenManage 4 infrastructure and the recently announced Dell OpenManage change management capability utilize industry-standard technologies to provide extended systems management functionality across the enterprise.

Dell PowerEdge 1855 blade server

TABLE OF CONTENTS

STORAGE

SCALABLE ENTERPRISE

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

EDITORIAL

EDITOR-IN-CHIEF | Tom Kolnowski
MANAGING EDITOR | Debra McDonald
SENIOR EDITORS | Liza Graffeo, Cathy Luo, Anthony Poliseno, Vicki Van Ausdall
CONTRIBUTING AUTHORS | Tim Abels; Rizwan Ali; Rudramuni B; John Benninghoff; Suri Brahmaroutu; Michael E. Brown; Alan Brumley; Michael Brundridge; Casey Burns; Onur Celebioglu; Kalyana Chadalavada; Balasubramanian Chandrasekaran; Narayan Devireddy; Yung-Chin Fang; Bill Fitzgerald; Mike Garcia; Roger Goff; Steven Grigsby; Manoj Gujarathi; Baris Guler; Rinku Gupta; Eric Harless; Jenwei Hsieh, Ph.D.; Ruoting Huang; Munira Hussain; Saeed Iqbal, Ph.D.; Dave Jaffe, Ph.D.; Rajiv Kapoor; Sandeep Karandikar; Monica Kashyap; Garima Kochhar; Paul Laster; Jim Lathan; J. Craig Lowery, Ph.D.; Arrian Mehis; Balaji Mittapalli; Todd Muirhead; C.S. Prasanna Nanda; Srivathsa NS; Michael O'Hara; Michael Parker; Komal Patel; Ron Pepper; Jimmy Pike; Joe Pollock; Ryan Putman; Paul Rad; Ramesh Radhakrishnan, Ph.D.; Anusha Ragunathan; Ramesh Rajagopalan; Mike J. Roberts; Amina Saify; Bharat Sajnani; Ananda Sankaran; David Schmidt; Uday Datta Shet; Lance Shuler; Simone Shumate; Pramada Singireddy; Sudhir Srinivasan, Ph.D.; Scott Stanford; Bob Ure; Doug Wallingford; Gong Wang; David Weber; Yi-Ming Xiong, Ph.D.; Jianwen Yin, Ph.D.; and Haihong Zhuo

ART

ART DIRECTOR | Mark Mastroianni
DESIGNER AND ILLUSTRATOR | Cynthia Webb
COVER DESIGN | Mark Mastroianni
CONTRIBUTING PHOTOGRAPHER | Lee Kirgan

ONLINE

WEB PRODUCTION | Brad Klenzendorf

SUBSCRIPTION SERVICES

CIRCULATION | Sheryl Williams

Subscriptions are free to qualified readers who complete the online subscription form or the subscription reply form on page 74 in this issue. 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**.

No part of this publication may be reprinted or otherwise reproduced without permission from the Editor-in-Chief. Dell does not provide any warranty as to the accuracy of any information provided through *Dell Power Solutions*. Opinions expressed in this magazine may not be those of Dell. The information in this publication is subject to change without notice. Any reliance by the end user on the information contained herein is at the end user's risk. Dell will not be liable for information in any way, including but not limited to its accuracy or completeness. Dell does not accept responsibility for the advertising content of the magazine or for any claims, actions, or losses arising therefrom. Goods, services, and/or advertisements within this publication other than those of Dell are not endorsed by or in any way connected with Dell Inc.

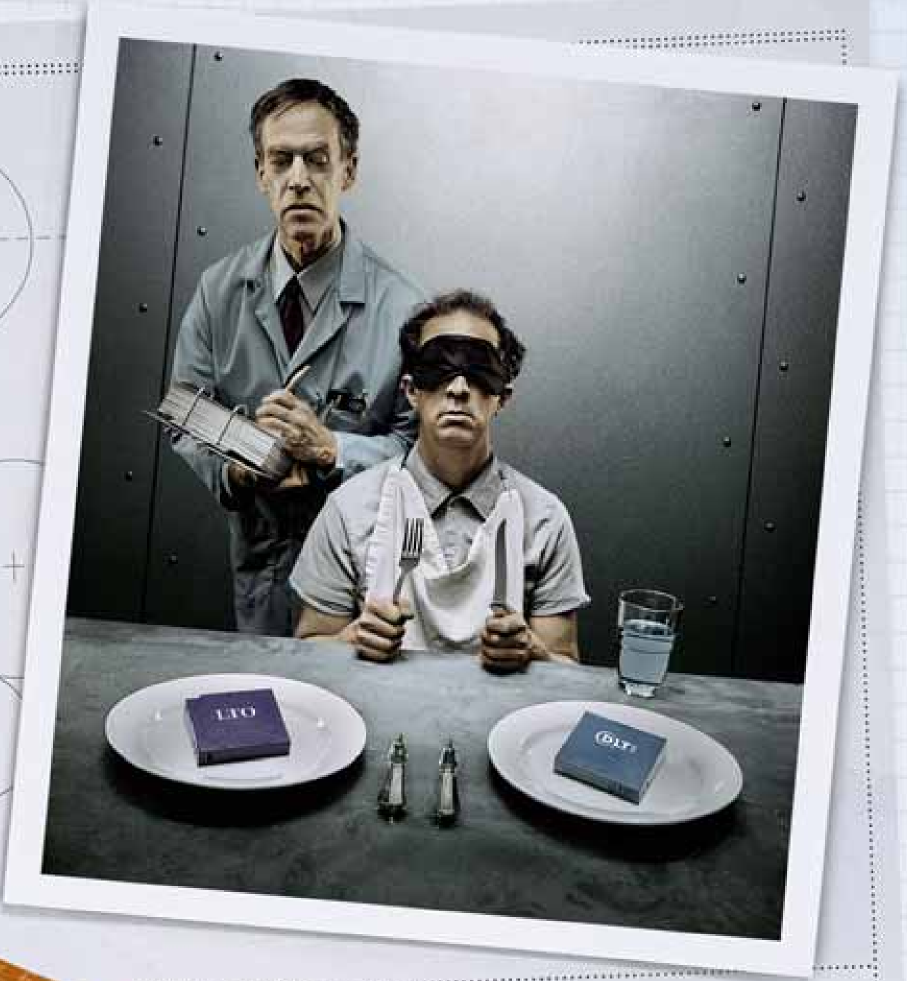
ON THE FRONT COVER | The Dell PowerEdge 1855 blade server.

February 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.

SDLT 600 vs. LTO-2



TEST #320

COMMENTS:

	SDLT 600	LTO-2	TIE
Taste test :	□	□	☑
Highest capacity :	☑	□	□

TESTED BETTER
DLT TAPE

In blind taste tests between the SDLT 600 and LTO-2, neither tape was ever able to gain a statistical advantage. Test subjects' comments ranged from, "both tapes equally offensive to the gastrointestinal system," to "If there is a hell, this is the food." Scientists have agreed to conduct the next round with condiments. As for data backup abilities, however, it was no contest. **The SDLT 600 has 50% more capacity than LTO-2, not to mention up to 20% more speed.** How do we know? It's been tested. For more info and to see the whitepaper, visit DLTtape.com.

DLT TAPE

TABLE OF CONTENTS

HIGH-PERFORMANCE COMPUTING

ADVERTISER INDEX

ONLINE XTRA

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

WWW.DELL.COM/ POWERSOLUTIONS

Troubleshooting Enterprise Production Environments Using Dell Server Diagnostic Tools

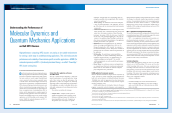
By Komal Patel and Pramada Singireddy

Dell server diagnostic tools are engineered to augment fault isolation and root-cause analysis in enterprise production environments. This article discusses various tools available for server diagnostics at the tier 1, tier 2, and tier 3 levels, along with best practices for the use of server diagnostic tools.

Evaluating File-Serving Performance of the Dell PowerEdge 700 Server

By Yi-Ming Xiong, Ph.D.

The Dell PowerEdge 700 demonstrated a considerable performance gain compared to its predecessor, the PowerEdge 600SC, when Dell engineers measured its file-serving performance using Ziff Davis NetBench. This article explains the findings of this comparison test, including results for CPU frequency, cache size, and memory size tests performed on the PowerEdge 700 server.



Understanding Performance of Molecular Dynamics and Quantum Mechanics Applications on Dell HPC Clusters

By Kalyana Chadalavada and Srivathsa NS

High-performance computing (HPC) clusters are proving to be suitable environments for running a wide range of parallel-processing applications. This article discusses the performance and scalability of two domain-specific scientific applications—NAMD (for molecular dynamics) and DFT++ (density-functional theory)—on a Dell PowerEdge HPC cluster running Linux.

Considering Middleware Options in High-Performance Computing Clusters

By Rinku Gupta, Monica Kashyap, Yung-Chin Fang, and Saeed Iqbal, Ph.D.

Middleware is a critical component for the development and porting of parallel-processing applications in distributed HPC cluster infrastructures. This article describes the evolution of the Message Passing Interface (MPI) standard specification as well as both open source and commercial MPI implementations that can be used to enhance Dell HPC cluster environments.

Tooling Up for Change Management

A recent Web search for the phrase *change management* returned over 30,800,000 results. That fact alone elevates change management to the status of a bona fide buzzword—with its abundant references to personal and organizational development, business process reengineering, and countless other disciplines. But for the IT professional, change management in the data center translates to much more than a buzzword. It is a way of life. In this first 2005 issue of *Dell Power Solutions,* we'll help you get a handle on new systems management tools that can alleviate some of the effort required to keep pace with constant change while ensuring server health across the enterprise.

In our cover story, "Managing Change Through Industry-Standard Technologies," author Paul Laster helps you tackle change management head-on, with a technical drill-down on the Dell™ OpenManage™ 4 systems management infrastructure. Among the latest standardization developments in this area is Dell's recent announcement with Microsoft regarding the availability of integrated patch management using Microsoft® Systems Management Server (SMS) 2003 and Dell OpenManage 4. Through the SMS 2003 Inventory Tool for Dell Updates, IT administrators can harness the power of SMS to seamlessly apply patches across distributed Dell PowerEdge™ server hardware via the same processes they use for updating operating systems and application stacks.

For a broader perspective on Dell's systems management strategy, read our Executive Insights interview with Rhonda Holt, vice president of systems management software at Dell. This interview articulates Dell's focus on standardizing systems management elements in the data center to continually add value for enterprise customers, including a discussion of recent enhancements to the Dell OpenManage portfolio and tight integration with key industry partners and standards bodies.

In-depth guidance on systems management tools and techniques appears in the New-Generation Server Technology section. Among the highlights are articles covering remote configuration of the standard baseboard management controller (BMC) in the latest Dell PowerEdge servers and best practices for systems management with the new Dell PowerEdge 1855 blade server. You'll find even more technical guidance for managing Dell blade servers in articles that explore remote management, configuration for optimum networking availability and performance, and how to take scalability to new levels using server virtualization software such as VMware® ESX Server™ software.

In addition to general systems management topics, this issue explores leading-edge content in the Scalable Enterprise and Storage sections, and concludes with a continuation of our high-performance computing (HPC) coverage, featuring seven fresh articles that range from streamlining Beowulf cluster deployments on Linux with the NPACI Rocks toolkit to designing HPC clusters with the new Dell PowerEdge SC1425 server.

To move beyond our roster of 30 articles in the print edition, point your browser to www.dell.com/powersolutions to get our Web edition, which includes the latest Web-exclusive content. Turn to page 4 for a preview of the four online-extra articles featured in this issue.

Tom Kolno

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

Maximizing Manageability Across the Scalable Enterprise

As vice president of systems management software at Dell, Rhonda Holt leads the development and integration of Dell™ OpenManage™ infrastructure for Dell's family of standards-based PowerEdge™ servers. Here she shares her thoughts on the importance of integrated change management in a dynamic business environment.

Need a winning combination for systems management in the scalable enterprise? Look no further. Not only are eighth-generation PowerEdge servers Dell's most manageable systems to date, but the latest release of Dell OpenManage infrastructure—Dell OpenManage 4—is also an industry standards–based platform that has been designed for enhanced efficiency in deploying, monitoring, and updating vital technology across the enterprise.

In particular, the recently announced Dell OpenManage change management capability, which works in concert with Microsoft® Systems Management Server (SMS) 2003 and Altiris Server Management Suite, for the first time enables highly automated patch management across hardware, operating systems, and applications—on the latest Dell servers as well as on several previous-generation Dell servers.

How does Dell's systems management infrastructure facilitate change management?

Change management refers to the manner in which an organization plans for and deploys technology upgrades across the enterprise. It sounds simple, but change management can be extremely tricky—particularly given the escalating rate of change brought about by mounting security concerns. Change management requires a thorough awareness of all existing enterprise assets—from how they are being utilized to where they are dispersed. It also requires consistent attention to available IT resources so that transition processes can be as efficient as possible. Moreover, change management demands that administrators keep their eyes on enhanced technology, so they can seize new opportunities to improve return on IT investments.

We know that administrators face serious limitations on time and resources. Managing thousands of cross-enterprise security and applications updates is extremely challenging. By consolidating important change management tools, Dell OpenManage 4 helps make technology transitions simpler, so administrators can accomplish upgrades in less time and with less effort than was possible using previous versions of Dell OpenManage. Ultimately, this reduced complexity can lead to increased employee productivity, longer-term IT investment return, and lower overall total cost of ownership. These are compelling advantages for the scalable enterprise.

What is the significance of the Dell OpenManage–based approach to change management?

Dell's approach is innovative. While other vendors continue to invest in and develop proprietary, hardware-only change management applications, Dell is again focusing on standardization. We know standards are the key to enabling more functional, more efficient change management applications. And we know interoperability is essential to seriously easing the burden on IT administration.

Toward that end, Dell recently co-announced with Microsoft's Steve Ballmer the SMS 2003 Inventory Tool for Dell Updates. This is a watershed event for our customers.[1] For the first time, this combined solution gives enterprises a single view of all Dell PowerEdge servers on the network and helps them easily locate update requirements for everything from the operating system to system software to applications. Moreover, Dell is the *only* hardware vendor to enable comprehensive hardware patch management for its servers with

[1] For more information, see the Dell news release at www1.us.dell.com/content/topics/global.aspx/corp/pressoffice/en/2004/2004_11_15_par_000?c=us&l=en&s=corp.

leading third-party enterprise management applications such as Microsoft SMS and Altiris Server Management Suite.

Dell Update Packages are the basic building block in all Dell update solutions, allowing administrators to deliver hardware updates with a single mouse click. In addition, these intuitive, self-extracting files let administrators concurrently apply updates across many systems for the operating system, applications, and components such as BIOS, firmware, and drivers. Dell Update Packages can be scripted for unattended installation—which unchains administrators from the chore of babysitting dialog boxes, freeing them to tackle business-critical tasks. This unified, integrated approach is designed to save tremendous amounts of time, effort, and expense. This architecture allows you to integrate Dell updates into your existing software distribution process. You can also look for Dell to offer similar integration with other leading management solutions in the near future.

Other important change management capabilities within Dell OpenManage 4 infrastructure include two e-mail notification services, File Watch and Knowledge Watch. These services help administrators conveniently track available updates for Dell products. File Watch sends information on all newly posted product downloads, while Knowledge Watch sends notifications only for critical product alerts.

How do industry standards contribute to Dell's management infrastructure and the overall manageability of Dell servers?

Industry standards are the mainstay of the scalable enterprise. They provide a common language that enables communication between system components regardless of vendor. In the case of systems management, this common language helps administrators manage many system tasks in a single operation. In other words, standards can enhance productivity, streamline systems management, and maximize IT resource efficiency.

Dell has been committed to standards-based technology for many years, and this has been a driving force in the development of our scalable enterprise strategy. In fact, our most recent server launch has centered strongly on standards-based platform commonality. Dell's powerful eighth-generation servers feature a common system image, common platform elements, and standard Intelligent Platform Management Interface (IPMI)–accessible baseboard management controllers. In addition, these servers can be equipped with the standards-based Dell Remote Access Controller 4 (DRAC 4), which is designed to provide operating system–independent operation for uninterrupted manageability of remote Dell servers. Dell OpenManage 4 infrastructure is designed to provide the critical framework that enables organizations to leverage these platform commonalities to dramatically improve systems management capabilities.

The release of Dell OpenManage 4 infrastructure accentuates our commitment to leading and supporting such critical standards as the Common Information Model (CIM) and the Systems Management Architecture for Server Hardware (SMASH) specification from the Distributed Management Task Force (DMTF), Desktop Management Interface (DMI), Simple Network Management Protocol (SNMP), and Wired for Management (WfM).

And Dell's dedication to standards-based systems management does not stop with our servers. Dell/EMC storage area network (SAN) solutions offer a centralized, standardized management console through EMC® VisualSAN® Standard Edition software, which is designed to be integrated with Dell OpenManage Server Administrator for maximum productivity. Plus, our line of Dell PowerConnect™ network switches can be managed through the Dell OpenManage Network Manager or a standard command-line interface.

Does Dell offer other integrated management solutions?

Certainly. Besides enabling the distribution of Dell Update Packages through a number of leading software solutions, Dell facilitates interoperability across other critical areas of systems management. For example, Dell eases deployment challenges with an enabling technology called the Dell OpenManage Deployment Toolkit. Used in conjunction with solutions such as Microsoft Automated Deployment Services and Altiris Deployment Solution, the Dell Deployment Toolkit helps leverage automated, standardized methodologies to enable accelerated group deployments of PowerEdge servers. Dell also offers a holistic monitoring approach for the PowerEdge server family that integrates our hardware management solution, Dell OpenManage Server Administrator, with industry-leading management solutions such as Microsoft Operations Manager (MOM), Computer Associates Unicenter, and BMC PATROL.

Additionally, with Dell OpenManage 4 infrastructure, we offer integration with Microsoft Active Directory® advanced network services to strengthen remote access and management security for a wealth of Dell enterprise assets, including Dell PowerEdge servers; Dell OptiPlex™, Dell Latitude™, and Dell Precision™ client systems; Dell PowerVault™ storage systems; and Dell PowerConnect switches. In particular, the DRAC 4 card can help leverage this integrated management solution by enabling secure remote access to servers outside the physical grasp of administrative IT staff virtually anywhere, virtually anytime. And the DRAC 4 is designed to provide complete, one-touch management—without requiring specific operating system services or drivers.

How does Dell envision the future of systems management?

We plan to continue to add value by strongly driving standardization in the area of systems management—an area where we feel our customers should accept nothing less. And look for ever more functional, ever more manageable products. For example, the recently released Dell PowerEdge 1855 blade server demonstrates a unique understanding of the needs of the scalable enterprise—enabling organizations to significantly increase density, improve price/performance, and lower power consumption compared to traditional Dell 1U servers.

Managing Change

Through Industry-Standard Technologies

Continuing its strategic focus on shifting technology toward industry standards, Dell recently broadened its enterprise systems management offerings. Dell™ OpenManage™ 4 infrastructure and the recently announced Dell OpenManage change management capability utilize industry-standard technologies to provide extended systems management functionality across the enterprise. This article introduces these extensions to the Dell scalable enterprise initiative.

BY PAUL LASTER

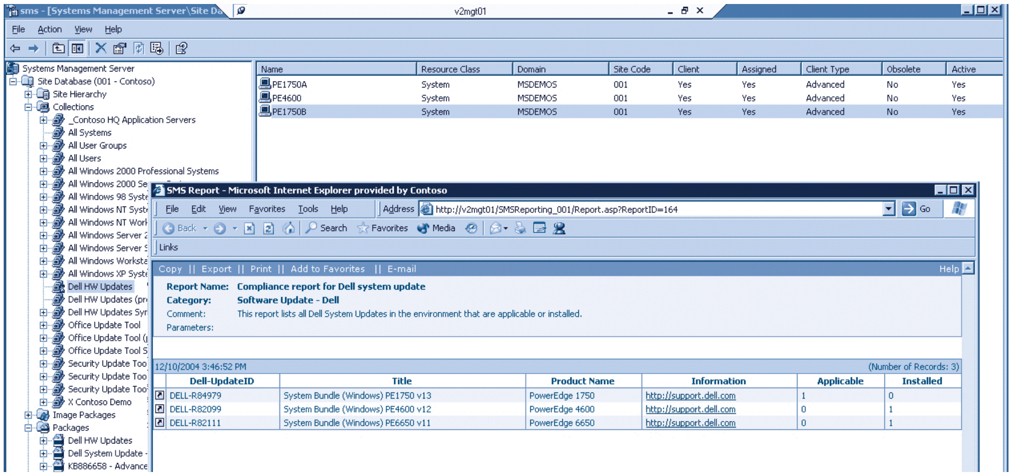
The value of enterprise systems management goes only as far as an organization's ability to integrate all of its relevant assets—which often include heterogeneous server and storage systems—wherever they may be located. However, many IT organizations are hamstrung by the challenge of managing increasing volumes of data and maintaining enterprise systems with little, if any, growth in IT staffing to handle the task.

Dell's baseline systems management infrastructure, Dell OpenManage 4, is designed to help alleviate IT resource concerns. Released concurrently with the first wave of eighth-generation Dell PowerEdge™ servers,[1] the Dell OpenManage 4 infrastructure is backward compatible with several previous generations of Dell PowerEdge servers. Dell OpenManage 4 is designed to streamline the way organizations handle their change management, deployment, and monitoring processes. In particular, the Dell OpenManage change management capability adds innovative functionality to the Dell OpenManage infrastructure to help leverage existing patch management applications. While other server hardware vendors approach change management by providing proprietary, hardware-only capabilities, Dell's emphasis on standardization with Dell OpenManage enables seamless distribution of Dell hardware updates—through the same familiar tools and processes administrators use for updating their operating systems and application stacks.

Eliminating proprietary, hardware-only change management tools

IT organizations can consume a great deal of time and budget ensuring that systems have the latest operating system (OS) and application patches. Offerings such as Microsoft® Systems Management Server (SMS) 2003 can help organizations automate the way they track, stage, and distribute these patches. The Dell OpenManage change management capability integrates with Microsoft SMS 2003 through the SMS 2003 Inventory Tool for Dell Updates to help administrators manage both server software and hardware through a single console (see Figure 1). The technology behind the SMS 2003 Inventory Tool for Dell Updates allows Dell PowerEdge system software—such

[1] For more information about eighth-generation Dell PowerEdge servers, see "Dell Extends the Scalable Enterprise with Eighth-Generation PowerEdge Servers" by John Fruehe in *Dell Power Solutions,* October 2004.



Name	Resource Class	Domain	Site Code	Client	Assigned	Client Type	Obsolete	Active
PE1750A	System	MSDEMOS	001	Yes	Yes	Advanced	No	Yes
PE4600	System	MSDEMOS	001	Yes	Yes	Advanced	No	Yes
PE1750B	System	MSDEMOS	001	Yes	Yes	Advanced	No	Yes

http://v2mgt01/SMSReporting_001/Report.asp?ReportID=164

Report Name: **Compliance report for Dell system update**
Category: **Software Update - Dell**
Comment: This report lists all Dell System Updates in the environment that are applicable or installed.
Parameters:

12/10/2004 3:46:52 PM (Number of Records: 3)

Dell-UpdateID	Title	Product Name	Information	Applicable	Installed
DELL-R84979	System Bundle (Windows) PE1750 v13	PowerEdge 1750	http://support.dell.com	1	0
DELL-R82099	System Bundle (Windows) PE4600 v12	PowerEdge 4600	http://support.dell.com	0	1
DELL-R82111	System Bundle (Windows) PE6650 v11	PowerEdge 6650	http://support.dell.com	0	1

Figure 1. Dell OpenManage change management integration with Microsoft SMS 2003 through the SMS 2003 Inventory Tool for Dell Updates

as BIOS, firmware, and driver updates—to be distributed via industry-standard patch management applications such as Microsoft SMS 2003.

Compared to systems management tools that focus on proprietary hardware, the SMS 2003 Inventory Tool for Dell Updates enables administrators using Microsoft SMS to manage Dell PowerEdge servers just as they manage the OS and application software. Dell is the only hardware vendor that enables server hardware updates to be delivered via third-party applications such as Microsoft SMS.

Integrating updates for server system software such as BIOS, firmware, and drivers into patch management applications allows system administrators to use a single process to update their OS, applications, and system software. As a result, distribution of hardware updates can be scheduled and delivered to target servers at the same time as OS and application patches, which helps streamline operations and minimize any potential business impact resulting from the systems maintenance process.

Applying updates with Dell OpenManage 4

The Dell OpenManage change management integration enhancements can empower administrators with a single view of software and hardware inventories—including PowerEdge servers and a list of the latest Dell system software updates—to determine what version of system software resides on each server and what needs to be updated. Integrating seamlessly with Microsoft SMS 2003, the SMS 2003 Inventory Tool for Dell Updates can proactively access systems management updates posted on the Dell Web site and provide automatic alerts of available upgrades. This integrated approach helps expedite the process of change management, saving valuable administration time by minimizing the demands placed on IT resources. At the click of a mouse, administrators can automatically download BIOS, firmware, and drivers from Dell's Web site. In addition, Microsoft SMS is designed to proactively search Dell's Web site and download the latest system software.

To help ensure that administrators use the latest updates, Dell provides File Watch and Knowledge Watch services to monitor newly posted downloads for products and critical product alerts, respectively. The Dell OpenManage Subscription Service also helps administrators stay up-to-date with the latest versions of the Dell OpenManage infrastructure.

Dell Update Packages are the fundamental building blocks that enable administrators to apply updates to PowerEdge servers in a scalable, nonintrusive manner. For example, Dell Update Packages can be used as one-to-one, stand-alone executable applications to apply a hardware update to a single server, while ensuring that specific validation criteria are met before the application is installed. Alternatively, Dell Update Packages can be distributed to servers using industry-leading patch management applications such as Microsoft SMS as well as several other custom-developed IT change management tools. In addition, Dell Update Packages can be deployed remotely in a one-on-one manner with Dell OpenManage Server Administrator.

Deploying new systems with Dell OpenManage 4

Dell OpenManage 4 provides automated deployment tools to help minimize the time, cost, and productivity drain of implementing technology upgrades. Administrators can install Dell OpenManage 4 through Dell OpenManage Server Assistant—which comes on a bootable, standalone CD that provides a series of intuitive and easy-to-follow interview screens to guide administrators through the installation. The Dell OpenManage Server Assistant CD also includes the latest drivers optimized for use on Dell PowerEdge servers—and is designed to configure Dell-provided RAID controllers with a consistent interface. Dell OpenManage Server Assistant also assists with OS installation, including installation of Microsoft Windows®, Novell® NetWare®, and Red Hat® Enterprise Linux® operating systems.

For highly automated server deployments, the Altiris Deployment Solution for Dell Servers provides comprehensive server provisioning from bare-metal servers for both Windows and Linux environments, including the configuration of BIOS, RAID, and Dell Remote Access Controller (DRAC) components. Prebuilt scripts are included for deploying Dell Update Packages and common server applications concurrent with the server build process. The solution also supports deployment of server blades within the Dell PowerEdge 1855.

HIGH-DENSITY COMPUTING: DELL POWEREDGE 1855 BLADE SERVER

Dell's latest blade server offering, the Dell PowerEdge 1855 blade server, breaks new ground in high-density computing (see Figure A). This high-performance blade server helps solve many of the challenges that organizations face today: intensified scrutiny on capital expenditures, limited data center space, increased workload for IT resources, inadequate power and cooling, complex integrations into existing environments, and unmanageable cable sprawl.

The PowerEdge 1855 blade server is designed to deliver true server functionality, featuring dual Intel® Xeon™ processors with Intel Extended Memory 64 Technology (EM64T); an 800 MHz frontside bus; dual hot-plug Ultra320 SCSI 3.5-inch hard drives; double data rate 2 (DDR2) memory scalable to 16 GB with 4 GB dual in-line memory modules (DIMMs);* Peripheral Component Interconnect (PCI) Express; and integrated server management.

Eighth-generation PowerEdge technology

The PowerEdge 1855 blade server not only offers many of the same features that the PowerEdge 1850 server provides, but it also shares a similar system architecture with the PowerEdge 1850 server. Organizations that require the performance and manageability of the PowerEdge 1850 1U rack server can easily integrate the PowerEdge 1855 blade server into their data center. The common baseboard management controller in both systems means that the PowerEdge 1855 server is designed to appear the same way to management applications as stand-alone eighth-generation PowerEdge servers.

Dell designed the PowerEdge 1855 blade server chassis based on Dell's eighth generation of PowerEdge technologies. In addition, the PowerEdge 1855 blade server was designed with the foresight to support future generations of PowerEdge blade servers and I/O technologies. The modularity of server blades and I/O bays in the rear of the chassis was designed specifically to allow future-generation server blades to drop into the chassis, helping to protect IT investments.

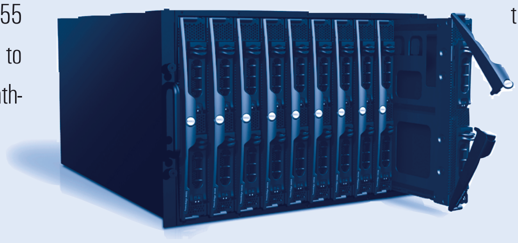


Figure A. Dell PowerEdge 1855 blade server

Simplified systems management

Dell's integrated approach to systems management allows IT organizations to manage blade servers in the same way they manage stand-alone servers. Just like other PowerEdge servers, the PowerEdge 1855 blade server is easy to manage using the intuitive Dell OpenManage 4 suite for deployment, change management, and monitoring. Once the server blade is installed in the chassis, it can easily be provisioned with existing tools, including those from Dell, Microsoft, and Altiris. A Universal Serial Bus (USB) port on the front of each server blade facilitates provisioning with external CD media, floppy drives, or USB key devices.

Benefits of reduced form factor

The blade server chassis greatly simplifies the cabling process compared to stand-alone servers by enabling administrators to aggregate cabling for up to 10 server blades at once. The PowerEdge 1855 helps reduce network; power; and keyboard, video, mouse (KVM) cabling required for 10 1U rack servers up to 70 percent by implementing server blades that are configured in a single chassis. This configuration enables administrators to reduce redundant power cables from 20 cables (two per server) to 4 cables per chassis. A single KVM controller for the blade server chassis allows one connection to replace the 30 connections required for 10 individual rack-mounted servers. These cabling reductions continue across networking, storage area network (SAN) connectivity, and management controllers. Reducing the number of cables not only helps speed deployment of servers, but it can also help prevent problems by presenting fewer physical connections to the servers.

Physical deployment of servers can be greatly accelerated by utilizing blade server systems. With individual rack servers, each server must be installed separately. With blade server systems, once the blade chassis has been installed, adding a server is as simple as sliding the server blade into the chassis and powering it up. The consolidation of power supplies and fans into a single chassis also helps to reduce the power requirements on a per-server basis.

By building its scalable enterprise strategy upon industry-standard technology, Dell enables choice and flexibility as well as integration with leading enterprise management applications within existing management infrastructures.

For more information, visit www.dell.com/blades.

*Availability of the dual-ranked 4 GB DIMMs is scheduled for Q1 2005.

Another Dell OpenManage 4 installation tool is the Dell OpenManage Deployment Toolkit. It provides a suite of enabling technologies designed to help with pre-OS configuration, remote server deployment, and reprovisioning of PowerEdge servers. In the form of command-line utilities and sample DOS batch scripts, the Deployment Toolkit can be downloaded from support.dell.com. Whether administrators have developed proprietary deployment solutions or use a third-party deployment approach, the Deployment Toolkit can help deploy Dell servers cost-effectively in an unattended fashion.

Monitoring servers with Dell OpenManage 4

Dell OpenManage 4 provides various standards-based tools for monitoring server configuration and health. The monitoring capability of Dell OpenManage 4 is designed to protect IT investments by working within existing environments—even when those environments comprise heterogeneous, geographically dispersed systems. Dell OpenManage 4 tools can monitor system configuration and health, help upgrade BIOS and firmware, and alert administrators to potential system problems.

The Dell PowerEdge 1850, PowerEdge 2800, and PowerEdge 2850 servers each include a baseboard management controller (BMC) based on the Intelligent Platform Management Interface (IPMI) 1.5 standard for remote sensor monitoring, fault logging and alerting, sensor and log displays, and server power control. The embedded BMC monitors critical components and environmental conditions such as fans, temperature, and voltage within the system. It generates alerts and helps administrators to detect and remedy problems at local and remote locations.

Remote access capabilities

By complying with the IPMI 1.5 industry management standard, the BMC can provide enhanced functionality such as enabling remote access to the server via the serial port or network interface card (NIC). This remote access allows administrators to power on, power off, and power reset servers independent of their OS or application status. In this way, administrators can manage servers without having to access them physically—whether or not the servers are operational. By providing administrators with secure access to remote servers, Dell can help organizations increase systems availability, enhance security, and improve IT productivity.

For example, Dell OpenManage Server Administrator is a highly secure, Web-based tool that enables administrators to manage an individual server from virtually anywhere and at virtually any time, without requiring console access. Server Administrator enables administrators to determine how much space is left on a system's disks, and what OS and OS version is installed; check for the latest firmware, drivers, and BIOS; diagnose a server without shutting it down; identify who made the latest modification to a server; and determine what is installed in a server's slots—all from the Microsoft Internet Explorer browser or a command-line interface (CLI). Additionally, Server Administrator is compatible with Linux, NetWare, and Windows operating systems.

Dell OpenManage Server Administrator software resides completely on the managed server. As a result, administrators do not have to access a central console when they are notified that a server needs attention. Instead, they can use a Web browser to select the server and log in using their domains or OS passwords to perform a wide range of systems management actions.

Network-wide systems management

Administrators can manage Dell servers and clients distributed throughout a network by using Dell OpenManage IT Assistant. From this central, standards-based console, administrators can gain a high level of control over the availability of Dell servers. IT Assistant identifies systems that are experiencing problem conditions and alerts administrators proactively, helping to reduce the risk of system downtime. Its Web-enabled graphical user interface (GUI) allows administrators to access IT Assistant from anywhere within the organization's network.

For organizations that require additional remote console capabilities, the DRAC family enables administrators to view real-time remote server activity from anywhere in the network. The latest enhancements, available in the DRAC 4 version for eighth-generation Dell PowerEdge servers, include continuous video capability, virtual media, and integration with the Microsoft Active Directory® directory service.[2]

Achieving comprehensive systems management

By consolidating change management tools, the Dell OpenManage 4 infrastructure helps minimize management complexity and eliminate the need for two separate processes—one to update the OS and applications, and another to update the system software. By using the Dell OpenManage change management capability and various other tools available with the Dell OpenManage 4 infrastructure, IT organizations can streamline their systems management procedures to make the most of limited IT resources.

Paul Laster is a marketing communications advisor for Dell Global Brand Interactive Services. He is a veteran of the Dell online community, working both as a developer and an online designer for the past seven years. Paul has a B.S. in English and Business from The University of Texas at Austin.

FOR MORE INFORMATION

Dell OpenManage:
www.dell.com/openmanage

[2] For more information about the DRAC 4, see "Exploring the Next-Generation DRAC 4 Dell Remote Access Controller" by Jon McGary and Donnie Bell in *Dell Power Solutions,* October 2004.

Systems Management Techniques

for the Dell PowerEdge 1855 Blade Server

Systems management capabilities built into the Dell™ PowerEdge™ 1855 blade server allow administrators to manage individual server blades using the Dell OpenManage™ infrastructure, while also enabling management of the entire system through the Dell Remote Access Controller/Modular Chassis management module. This article provides a high-level overview of the deployment, monitoring, and change management capabilities of the PowerEdge 1855 blade server.

BY STEVEN GRIGSBY AND NARAYAN DEVIREDDY

The Dell PowerEdge 1855 blade server is a modular system consisting of a chassis, up to 10 independent server blades, and the infrastructure components shared by the server blades in the chassis. The shared infrastructure components include power supply modules; cooling modules; network I/O modules; the Dell Remote Access Controller/Modular Chassis (DRAC/MC) remote access management module; and a keyboard, video, mouse (KVM) switch module. Systems management of the PowerEdge 1855 blade server can be categorized into two key areas:

- **Server blade management:** Server blade management is accomplished through individual management features on the server blades and in the I/O modules.
- **Chassis management:** System-wide management of shared infrastructure components is accomplished through the DRAC/MC.

Managing Dell PowerEdge 1855 blade servers is similar to managing stand-alone Dell PowerEdge servers. Each server blade provides traditional in-band[1] management features using the Dell OpenManage suite and out-of-band, agentless management using a built-in baseboard management controller (BMC)—and each server blade can be managed independently of the other server blades in the chassis.

Because all the server blades in the modular system share the common chassis infrastructure components, chassis management plays a critical role in overall Dell PowerEdge 1855 systems management. The Dell PowerEdge 1855 blade server chassis and its common infrastructure components are managed through the DRAC/MC component. The DRAC/MC monitors power supplies, power allocation, blade presence, and I/O module presence. The DRAC/MC provides the following management access points:

- Web-based graphical user interface (GUI) accessible through a browser
- Command-line interface (CLI) accessible through a Telnet connection
- CLI accessible through the DRAC/MC's integrated serial port

[1] In-band features are used when the operating system is present.

The DRAC/MC runs on low-wattage power and has a separate network connection to help ensure remote access to the system. The separate network connection provides a redundant path for the routing of network alerts.

Server blade configuration and deployment

Manually configuring the server blade hardware and deploying an operating system (OS) image on multiple blades may seem simple, but can become a daunting task when a large number of servers are involved. Several local and remote deployment options are available—ranging from local, attended deployment with CD media to remote, unattended deployment using Preboot Execution Environment (PXE) and custom scripts.

Local configuration

Manually configuring the server blades and deploying the OS requires administrators to physically interact with the blade server system. This may be practical in organizations that have relatively few servers if all the servers are physically accessible. One local deployment method is to repeat the configuration and deployment steps for each server blade. However, it may make more sense to perform the steps once, save the configuration data, and use the data to configure and deploy the remaining systems.

Because all the server blades in the modular system share the common chassis infrastructure components, chassis management plays a critical role in overall Dell PowerEdge 1855 systems management.

Local, attended configuration and deployment is accomplished by booting the Dell OpenManage Server Assistant (DSA) CD using a Universal Serial Bus (USB) CD drive attached to the front-panel dongle of the server. This requires the administrator to configure the boot order in the BIOS. Once the DSA CD boots, a DSA configuration wizard prompts the administrator with a series of questions to configure the server blade. As part of this configuration, the administrator must select an OS. DSA then prompts for the OS configuration data including network settings and time zone. To complete the installation, DSA prompts the administrator to insert the OS installation media.

To enable local, unattended installation, the administrator can choose to save the configuration data to a USB floppy disk drive attached to the front-panel dongle of the server and use this data to replicate subsequent installations on additional server blades. When the administrator moves the floppy disk containing the configuration data to the next server blade and boots the DSA CD, the configuration data is automatically read by DSA to complete the installation.

Remote configuration

When an organization has many servers to deploy, or when physical access is limited, administrators can opt to configure and deploy server blades remotely. The Dell OpenManage Deployment Toolkit contains several configuration tools for scripting tasks such as partitioning drives and configuring BIOS settings, BMC settings, and RAID controllers.

Using a combination of the Dell OpenManage Deployment Toolkit and third-party remote deployment products such as Altiris Deployment Solution, Microsoft® Automated Deployment Services (ADS), VERITAS OpForce, and Symantec ON iCommand, administrators can rapidly deploy remote PowerEdge 1855 server blades. The Dell OpenManage Deployment Toolkit provides configuration tools and sample scripts to configure hardware devices in a pre-OS environment. Scripts that configure BIOS, BMC, RAID, and disk devices can be leveraged in PXE to remotely configure the hardware and deploy the OS on the server blades.

Using PXE and remote deployment products is an excellent option for remotely deploying numerous servers. Although such techniques require additional time up-front to create the PXE image and script the configuration, the time investment is rewarded with the automatic, remote configuration and deployment of the server blades. Dell OpenManage Deployment Toolkit tools can also be leveraged in any custom deployment environment.

PowerEdge 1855 blade server monitoring capabilities

Hardware monitoring and alerting on a PowerEdge 1855 blade server can be accomplished with the following components:

- On-board BMC on each server blade
- Dell OpenManage Server Administrator (OMSA) running on each server blade
- DRAC/MC management module in the chassis

The embedded BMC is designed to work in conjunction with the DRAC/MC (which resides in the chassis) and OMSA (which runs on the server blade OS) to log and send alerts on the network. The BMC is responsible for monitoring the status of the voltage and temperature probes on individual server blades. When the BMC detects an event, the event is written to the BMC hardware log and sent to the DRAC/MC using the Intelligent Platform Management Interface (IPMI). The DRAC/MC then writes a corresponding event to the system event log (SEL) and sends a corresponding Simple Network Management Protocol (SNMP) trap to the configured recipient. The DRAC/MC can also be configured to send an e-mail alert using Simple Mail Transport Protocol (SMTP).

If Dell OpenManage software agents are installed on the server blade OS, the event will also be handed off to OMSA for user notification. OMSA can be configured to send SNMP traps and SMTP e-mail alerts to multiple recipients.

The combination of the DRAC/MC, the server blade's BMC, and Dell OpenManage agents is designed to provide a redundant notification path and redundant logging. The DRAC/MC can send an SNMP trap, an SMTP e-mail alert, or both through its own network connection; at the same time, Dell OpenManage agents can send an SNMP trap and an SMTP e-mail alert through the server blade's network connection. The event is logged in both the DRAC/MC SEL and the BMC hardware log.

In addition, Dell OpenManage agents can provide enhanced logging capabilities in the OMSA Alert log and Command log—the same level of logging provided by Dell OpenManage agents on stand-alone Dell servers.

The DRAC/MC also monitors the presence and health of the I/O modules, cooling modules, KVM switch module, and power supply modules in the chassis. When an event occurs on one of these modules, the DRAC/MC handles the event exactly as it would handle an event on a server blade. The event is written to the DRAC/MC SEL and—if configured—an SNMP trap, SMTP e-mail alert, or both are sent over the DRAC/MC out-of-band network.

When an organization has many servers to deploy, or when physical access is limited, administrators can opt to configure and deploy server blades remotely.

The PowerEdge 1855 blade server can also be remotely monitored and managed with Dell OpenManage IT Assistant (ITA), which is included in the Dell OpenManage suite. When ITA discovers the PowerEdge 1855 system, it groups PowerEdge server blades, Dell PowerConnect™ switches, and the DRAC/MC into a Chassis group—providing an overall system view that allows administrators to survey the individual server blades and components.

Access to the logs from the DRAC/MC and the server blades is available from the ITA console. When OMSA is installed on the server blades, OMSA can be configured to send server blade–related SNMP traps to the ITA console. Additionally, the DRAC/MC and PowerConnect switches can be configured to send SNMP traps to the ITA console. ITA can decode and report the traps and update the component and system status accordingly on the Status page of the system view.

PowerEdge 1855 blade server change management

Keeping server blade drivers, firmware, and BIOS updated can be challenging in organizations that have many servers to update. Using Dell Update Packages, administrators can update these components one at a time, or simultaneously apply the updates to many servers using management frameworks such as Microsoft Systems Management Server (SMS) or Altiris Deployment Solution. Dell Update Packages for the PowerEdge 1855 blade server are available for download from the Dell support Web site at support.dell.com. For single-server updates, a Dell Update Package can be either downloaded and executed or run from the Update tab of the OMSA GUI.

DRAC/MC firmware can be updated using the Trivial FTP (TFTP) protocol over the local area network. To update the firmware in this manner, administrators must install a TFTP server on a system that resides on the same network segment as the DRAC/MC. The firmware image can then be updated from the TFTP server using either the DRAC/MC CLI or GUI.

I/O module firmware can be updated through the interfaces provided by the I/O modules themselves. For example, the Dell PowerConnect 5316M Gigabit Ethernet switch provides for firmware updates from the switch GUI. In addition, firmware can be updated using the TFTP protocol in a manner similar to that previously described for DRAC/MC firmware.

A different management paradigm for modular blade servers

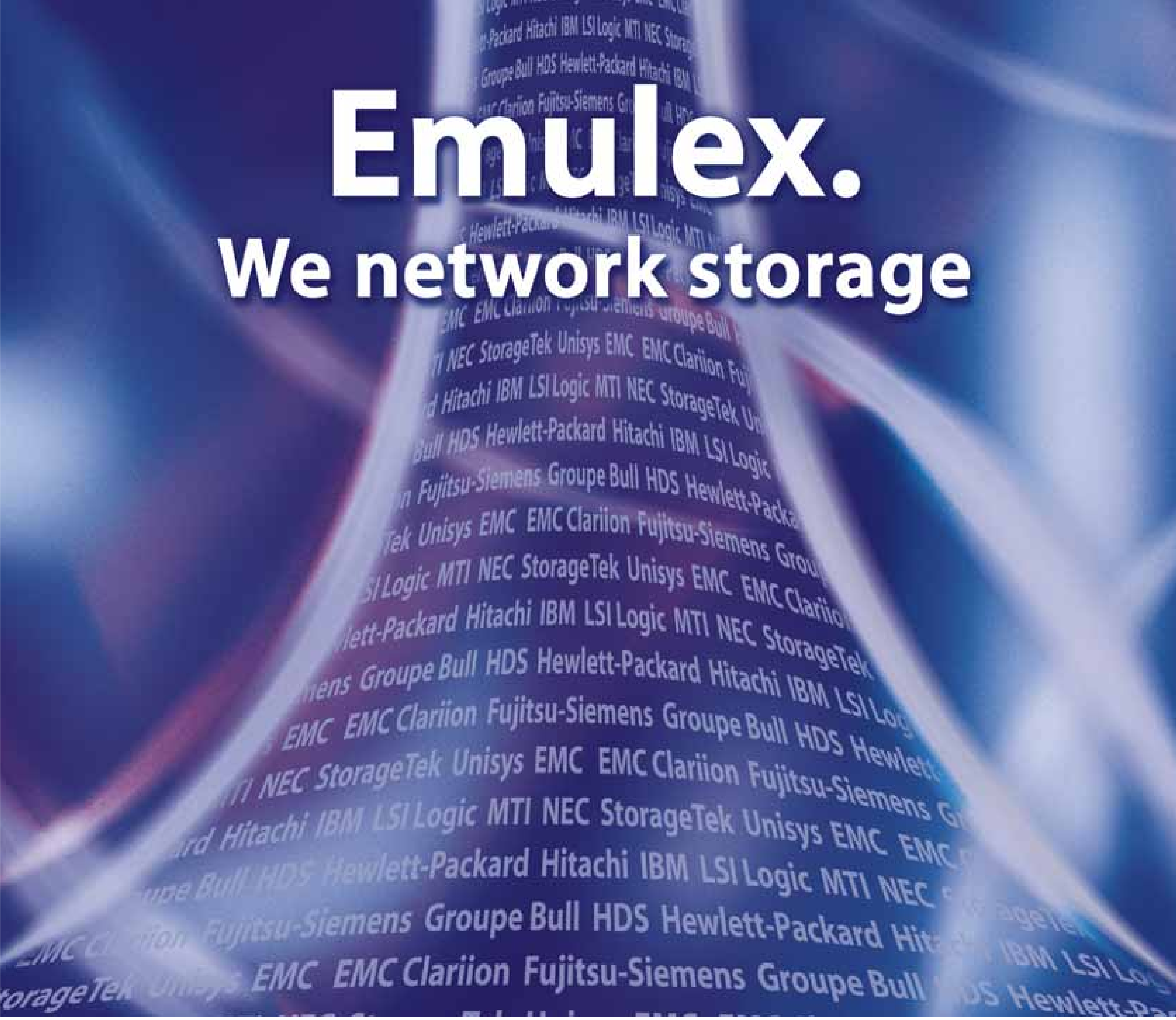
Managing the modular Dell PowerEdge 1855 blade server requires a somewhat different management paradigm compared to the traditional method of managing stand-alone servers. While managing a server blade is similar to managing a stand-alone server, management of the overall blade server system—including chassis infrastructure components and individual server blades—requires a combination of the DRAC/MC, the server blade BMC, and Dell OpenManage software agents. These components work together to provide a robust, redundant management environment—providing a holistic view of the entire system for optimal monitoring as well as redundant notification paths to help ensure that administrators are alerted when events occur.

Steven Grigsby is a test engineer in the Dell OpenManage Product Test organization. He has a B.S. in Computer Science from the University of Oklahoma.

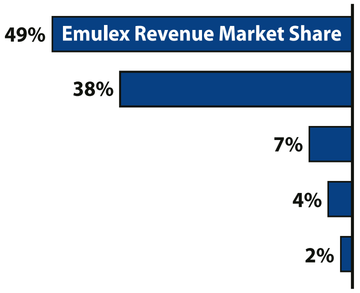
Narayan Devireddy is a development manager in the Dell OpenManage Development organization. He has an M.S. in Computer Science from Alabama A&M University.

FOR MORE INFORMATION

Dell OpenManage:
www.dell.com/openmanage



The industry's preeminent source for a broad range of advanced storage networking infrastructure solutions.



Source: Gartner Dataquest June 2004

Emulex's award-winning products are trusted and demanded by end users worldwide. That's because data integrity and access to their business critical information isn't something they can afford to take lightly. So it's no surprise that Emulex products are also an integral part of the world's largest server and storage providers' solutions. Emulex's market leadership position in the storage industry is an outgrowth of twenty-five years of commitment to continuous technology innovation and product excellence. And that's a commitment that we can't afford to take lightly.

Emulex: World's #1 supplier of Fibre Channel HBAs and embedded storage switches, and a leading supplier of SAN Switches and I/O controller products.

EMULEX™
We network storage

05-106

WWW.EMULEX.COM NYSE: ELX 800 EMULEX1

Remotely Managing the Dell PowerEdge 1855 Blade Server

Using the DRAC/MC

The Dell™ PowerEdge™ 1855 blade server provides two primary out-of-band interfaces for remote management: the Dell Remote Access Controller/Modular Chassis (DRAC/MC) for the blade server chassis, and a baseboard management controller for the individual server blades. This article discusses how to manage the blade server chassis using the DRAC/MC.

BY MICHAEL BRUNDRIDGE AND RYAN PUTMAN

The Dell PowerEdge 1855 system is Dell's new-generation blade server. In addition to providing a flexible, modular architecture, the PowerEdge 1855 blade server is designed to support advanced management features. The system enables high performance in a dense form factor and can include up to 10 server modules; the Dell Remote Access Controller/Modular Chassis (DRAC/MC) management module;[1] three or four hot-pluggable, redundant power supply modules; two cooling modules; up to four I/O module bays for I/O switches or pass-through modules; and a keyboard, video, mouse (KVM) switch module.

While the system's DRAC/MC is based on the standard Dell OpenManage™ feature set that is included in the eighth-generation Dell Remote Access Controller 4 (DRAC 4), the DRAC/MC has been enhanced to support the modular blade server system computing environment.

The DRAC/MC is a customer-serviceable module installed in the rear of the server chassis (see Figure 1). The module has a serial port and a 10/100 Mbps RJ-45 port. Two state LEDs are visible from the back of the module. The DRAC/MC is responsible for the management of the chassis and all of its shared components.

Introducing the DRAC/MC architecture and features

The DRAC/MC architecture provides autonomous monitoring of hardware, events, and recovery mechanisms. The DRAC/MC functions by using system standby power and is designed to be operational as long as AC power is available to the chassis, even if the chassis is powered down. Because the DRAC/MC works from standby power and runs its own real-time operating system (RTOS) and out-of-band interfaces, it can be used to manage and monitor the system even when the server module, chassis, or other shared components are powered down.

For example, when a problem occurs with power, temperature, fan speed, or the general health of the chassis, the system generates a system event log (SEL) event. If properly configured, the SEL event can be sent as an alert in the form of a Simple Network Management Protocol (SNMP) trap to a console such as Dell OpenManage IT Assistant or as an e-mail to one or more DRAC/MC users, or both.

The DRAC/MC offers the following key features:

- Remote power-up, shutdown, reset, or generation of a non-maskable interrupt (NMI) for each server module in the blade server chassis

[1] Redundant DRAC/MC modules are planned for a future release.

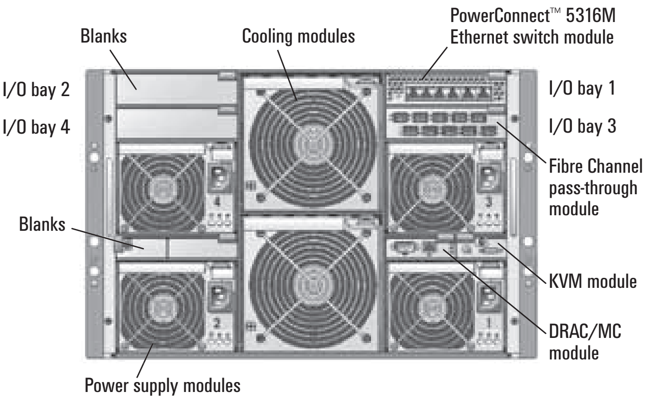


Figure 1. Rear view of the Dell PowerEdge 1855 modular blade server

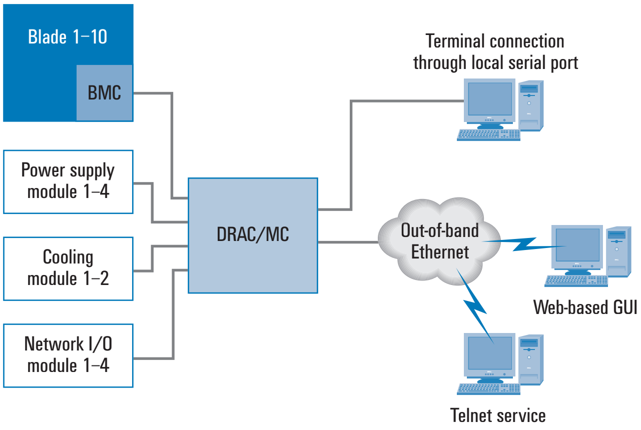


Figure 2. DRAC/MC management architecture

- Remote power-up and shutdown of the complete chassis
- Monitoring of shared components, including power supplies, fans, voltage, and temperature
- Access protection by password and privilege-level security
- Remote upgrade of the DRAC/MC's firmware through Trivial FTP (TFTP)
- Capability to deliver alerts by sending SNMP traps or e-mail messages
- Access to the chassis SEL and remote access controller (RAC) log
- Support for Dynamic Host Configuration Protocol (DHCP) of the DRAC/MC's IP address
- Capability to provide an inventory of the chassis including server modules, switch service tags (if applicable), I/O module types, and Media Access Control (MAC) addresses (if applicable)
- Serial port multiplexing for serial console redirection in which one serial port is multiplexed to the server modules and I/O modules

Figure 2 shows the three ways to access the DRAC/MC:

- **Web-based GUI:** Using the out-of-band, Web-based graphical user interface (GUI) to connect remotely through a Web browser; the GUI supports Secure Sockets Layer (SSL) encryption
- **Telnet service:** Using a Telnet client to connect remotely to the out-of-band, text-based console
- **DRAC/MC serial port:** Attaching directly to the DRAC/MC local serial port and using a virtual terminal such as Hilgraeve HyperTerminal to connect to the out-of-band, text-based console

Setting up the DRAC/MC through the serial port

Using the local serial command-line interface (CLI), administrators can perform all actions on the DRAC/MC with the exception of the items listed in Figure 3. This article assumes the use of the DRAC/MC serial port for initial setup.

Factory defaults for the DRAC/MC include the creation of a default administrator user and settings for many of the database object properties. Refer to Appendix C of the *DRAC/MC User's Guide,* located on the Dell OpenManage Documentation CD, for a quick reference guide to the object default properties.[2]

To view the basic syntax for the serial console commands, enter the following command at the CLI:

`help` (for a list of available commands)

or

`help` *`subcommand`* (for a list of the syntax statements for the specified subcommand)

To view the basic syntax for the `racadm` CLI commands, enter the following command at the CLI:

`racadm help` (for a list of available commands)

or

`racadm help` *`subcommand`* (for a list of the syntax statements for the specified subcommand)

Description	Management interface
Network physical control	Web only
Time zone	Web only
Web GUI time-out	Web only
Date/time format (12/24 hours)	Web only
User group	Web only

Figure 3. Configuration objects not supported in the local serial CLI

[2] The *DRAC/MC User's Guide* is also available online at support.dell.com/support/systemsinfo/documentation.aspx?c=us&cs=04&l=en&s=bsd&~cat=6&~subcat=111.

For detailed information about serial console or `racadm` CLI commands, refer to the *DRAC/MC User's Guide.*

Configuring the DRAC/MC network

The DRAC/MC network configuration consists of a network interface card (NIC), Telnet configuration, and Simple Mail Transport Protocol (SMTP) e-mail alert configuration. The DRAC/MC default network settings are as follows:

- DHCP disabled; static IP address 192.168.0.120; subnet mask 255.255.255.0; gateway 192.168.0.120
- Physical control auto-negotiation (enabling the NIC to automatically detect the correct speed and duplex at which it should be running)
- SMTP enabled; SMTP server IP address 127.0.0.1
- Telnet disabled; port 23 Telnet default

The IP address of the DRAC/MC is used to gain access to its remote interfaces over the out-of-band network. When using DHCP, administrators should configure the DHCP server to use a nonexpiring, MAC-based IP address reservation. The current DHCP-assigned IP address can be determined either by using the Dell OpenManage IT Assistant discovery feature to locate the DRAC/MC or by attaching to the DRAC/MC serial interface.

Configuration using the DRAC/MC CLI

To display the DHCP-assigned IP address, enter the following command at the CLI for the DRAC/MC:

```
racadm getniccfg
```

To enable DHCP, enter the following command at the CLI:

```
racadm setniccfg -d
```

To change the network configuration of the DRAC/MC to use a static IP address, enter the following command at the CLI for the DRAC/MC:

```
racadm setniccfg -s ipaddress subnetmask gateway
```

To enable the Telnet service, enter the following command at the CLI:

```
racadm config -g cfgSerial -o cfgSerialTelnetEnable 1
```

To change the Telnet port, enter the following command at the CLI:

```
racadm config -g cfgRacTuning -o
   cfgRacTuneTelnetPort value
```

Note: In the preceding command, the value for the Telnet port must be entered in hexadecimal format. The hexadecimal format for port 23 is 0x17.

Changing the network configuration does not require the DRAC/MC to be reset. However, users attached to the DRAC/MC over the network will lose the connection and be required to reestablish the connection when the network configuration is changed. Once the IP address is known, DRAC/MC users can connect to the DRAC/MC over the network using the GUI or a Telnet service.

Configuration using DRAC/MC database objects

The network settings can also be set by using the DRAC/MC database objects. To enable or disable DHCP, enter the following command at the CLI:

```
racadm config -g cfgLanNetworking -o
   cfgNicUseDhcp value
```

Note: In the preceding command, the value is either `0` to disable or `1` to enable.

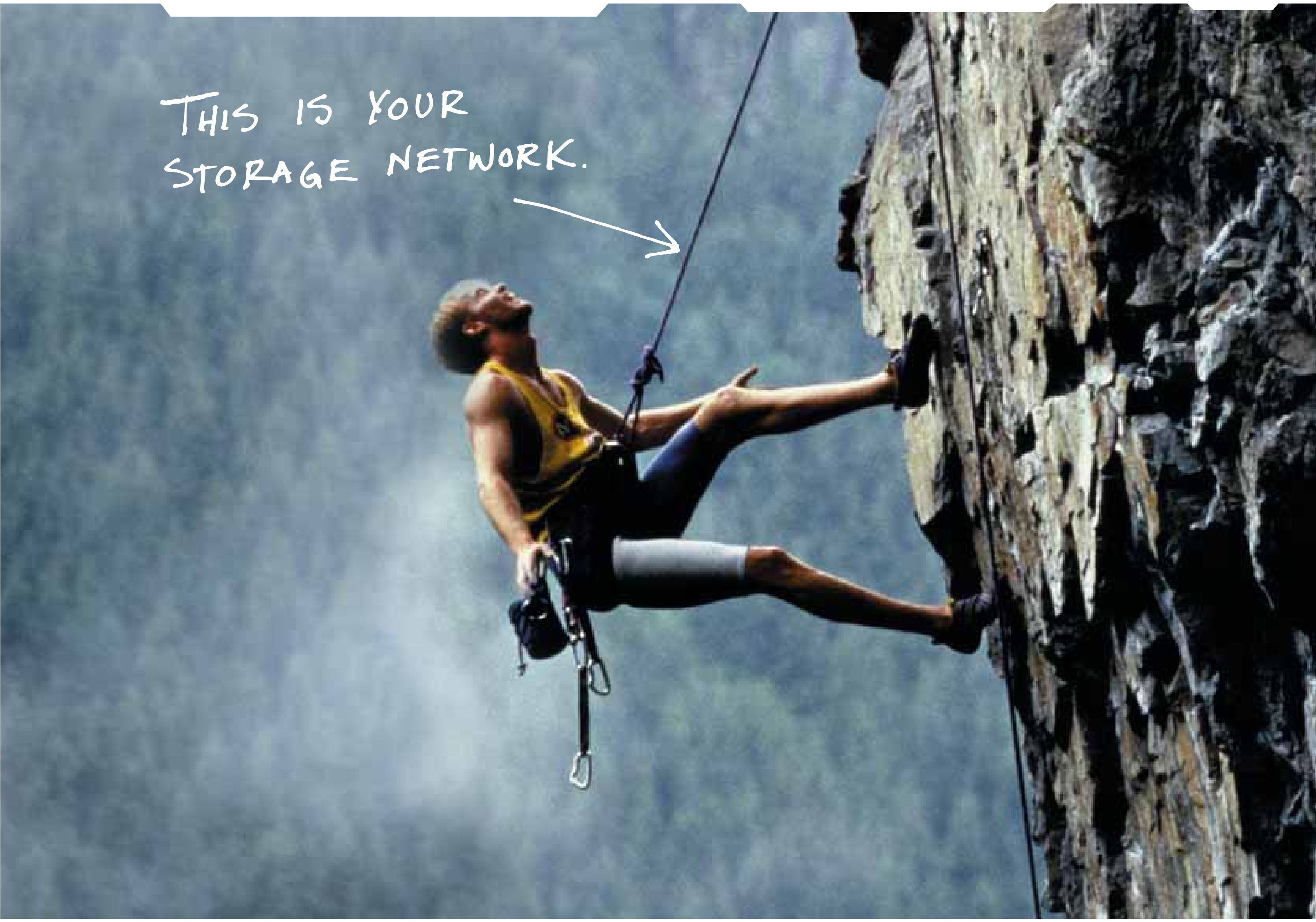
To change the network configuration of the DRAC/MC to use a static IP address, enter the following command at the CLI:

```
racadm config -g cfgLanNetworking -o
   cfgNicIpAddress ipaddress
racadm config -g cfgLanNetworking -o
   cfgNicNetmask subnetmask
racadm config -g cfgLanNetworking -o
   cfgNicGateway gateway
```

For more information on configuring the DRAC/MC network using the DRAC/MC serial interface or DRAC/MC database objects, visit *Dell Power Solutions* online at www.dell.com/powersolutions.

Managing the DRAC/MC through the GUI

The DRAC/MC has an embedded SSL-encrypted Web server from which it delivers an out-of-band GUI for remotely accessing the DRAC/MC. After configuring the network settings, administrators can launch the GUI by entering the DRAC/MC's IP address in a Web browser's URL address field. Then they can securely log in with a valid DRAC/MC username and password. After the username and password have been validated, the DRAC/MC Status page will appear. Because each DRAC/MC user can have a different privilege level, the GUI might appear different to each user based on those privileges. For instance, a user who does not have the Server Actions privilege will not see the Power tab or any of the other associated


THIS IS YOUR
STORAGE NETWORK.

User-level privilege	Bit mask*
Log in to DRAC/MC	0x80000001
Configure DRAC/MC	0x80000002
Configure users	0x80000004
Clear logs	0x80000008
Execute server control commands	0x80000010
Access console redirection	0x80000020
Test alerts	0x80000080
Execute debug commands	0x80000100

*Bit mask 0x80000040 is not currently used.

Figure 4. User-level privileges and associated bit masks

features that require the Server Actions privilege. Figure 4 indicates the level of permissions that a user must have to be able to perform actions with the DRAC/MC.

Accessing chassis information. The Properties tab of the GUI displays access information about the DRAC/MC, session status, shared components, and server modules. The GUI tabs and their associated links are available only if the user has the corresponding user privilege required for the feature. Information is shown on three separate pages:

- **Chassis Summary:** This page provides chassis, or *enclosure*, information, which includes the DRAC/MC's time, firmware version and date when firmware was last updated, current network settings, IP address, system model, service tag, asset tag, and chassis name and location. This page also provides session status, including the number of unused sessions, preliminary sessions, invalidated sessions, valid sessions, the current session ID, username, user's IP address, and login time.
- **Chassis Status:** This page provides information about all shared modules and server modules. Administrators can use this page to determine which modules are currently running in the Dell PowerEdge 1855 blade server as well as the service tags, power status, and overall health status of these modules.
- **Power Budget Status:** This page provides information about the current power budget status in the system. The page shows the current amount of power available along with the current amount of power currently being used. *Note:* The power values listed in the Power Budget Status page are static, maximum values; they do not reflect the actual power consumption of the system.

Monitoring chassis sensors. The Sensors tab shows sensor readings for the Dell PowerEdge 1855 blade server's temperature, cooling fan speed, and power supply status. This page also provides warning and failure thresholds for temperature and fan speed.

Managing the SEL and RAC log. The DRAC/MC maintains two persistent logs. The RAC log contains a list of user actions such as login and logout as well as alerts issued by the DRAC/MC. The oldest entries are overwritten when the log becomes full. Each log entry includes a brief description and information about severity, date and time, user, and ID of each event.

The SEL displays system-critical events that occur on the DRAC/MC and shared chassis components. This log includes the date, time, and a description of each event. To export both the DRAC/MC and SEL logs, administrators can click the Save As button in the GUI. Refreshing the logs in the GUI by selecting the Refresh button on the Web page before saving helps ensure that the latest logs are exported. Both logs can be cleared by clicking the Clear Log button.

Configuring chassis information. The Configuration tab enables many remote configuration tasks including configuring the chassis, creating and modifying users, creating and modifying alerts, configuring security for the DRAC/MC, configuring the network interface, and configuring the date and time.

Performing diagnostics. The Diagnostic tab allows administrators to display and execute basic network diagnostics, including:

- The Address Resolution Protocol (ARP) button displays the contents of the ARP table.
- `ping` verifies that the destination IP address is reachable from the DRAC/MC with the current routing-table contents.
- `ipconfig` displays the contents of the network interface table.
- `netstat` prints the contents of the routing table.

Updating the system firmware. The Update tab allows administrators to remotely update the DRAC/MC firmware image by using the DRAC/MC Flash function. Before performing the firmware update, administrators must download the latest firmware version from support.dell.com and then upload it to a TFTP server. The controller resets after the firmware update is complete.

Integrating the DRAC/MC with other management applications

The DRAC/MC offers a mechanism to integrate its own management capabilities with those of other management applications such as Dell OpenManage IT Assistant. An SNMP agent is embedded in the DRAC/MC; this agent implements SNMP management information bases (MIBs) and SNMP traps. The SNMP MIB is a hierarchical set of variables that can be read and written over the network. These variables contain information about the managed platform, including status, settings, and so on.

The DRAC/MC implements the MIB-II standard, which defines characteristics of the system and network interface. The DRAC/MC also implements an enterprise-specific MIB that provides management data specific to the server module system and devices such as service tags for the enclosure and server modules, health status for shared components, and so on.

The DRAC/MC uses SNMP version 1.0. Because this version does not provide complete security, the DRAC/MC SNMP agent does not support SNMP `set` operations and is disabled by default. The agent can be enabled by entering the following command:

```
racadm config -g cfgOobSnmp -o cfgOobSnmpAgentEnable 1
```

The MIB objects in the DRAC/MC are read-only. The `get` and `get next` commands can be performed on MIB objects.

SNMP is often used to monitor systems for fault conditions such as voltage failure or fan malfunction. Management applications such as IT Assistant can monitor faults by polling the appropriate object identifiers (OIDs) with the `get` command and analyzing the returned data. However, this polling method has its challenges. Performed frequently, polling can consume significant amounts of network bandwidth. Performed infrequently, this method may not allow administrators to respond quickly enough to the fault condition.

SNMP agents, supported by the DRAC/MC, can overcome such limitations by sending alerts or SNMP traps to designated recipients. The DRAC/MC can notify administrators when a system fails or is going to fail. To receive DRAC/MC SNMP traps at a management station running IT Assistant, the DRAC/MC must be configured for the trap destination, trap community name, and so on.

The DRAC/MC can also be configured to notify different trap destinations for different events by setting the proper SNMP trap filter. When the DRAC/MC detects a new event, the DRAC/MC inspects the event against each destination's trap filter and sends an SNMP trap to the appropriate destination.

Configuring alerts

DRAC/MC alerts consist of e-mail alerts and SNMP traps. The e-mail alert contains the following information: message, event description, date, time, severity, system ID, model, asset tag, service tag, managed system host name, and Embedded Server Management (ESM) version. The SNMP trap provides specific information describing the cause and source of the event. This information includes sensor identification, entity or Intelligent Platform Management Bus (IPMB) slave address, sensor number, sensor ID string (if possible), current sensor reading, range, and threshold values.

Adding a user with alert capabilities. To add a user who can receive e-mail notification, first locate the appropriate user index by entering `racadm getconfig -u` *`username`* command. Then, enter the following commands:

```
racadm config -g cfgUserAdmin -o
   cfgUserAdminEmailEnable -i index 1
racadm config -g cfgUserAdmin -o
   cfgUserAdminEmailAddress -i
   userindex email_address
racadm config -g cfgUserAdmin -o
   cfgUserAdminEmailCustomMsg -i
   userindex Custom Message
racadm config -g cfgRemoteHosts -o
   cfgRhostsSmtpServerIpAddr SMTP_Server_IP
```

Enabling SNMP traps. Up to 16 SNMP trap entries can be stored in the DRAC/MC MIB. To locate an available index to add a new SNMP trap, execute the following command for each index from 1 through 16 until an available index is located:

```
racadm getconfig -g cfgTraps -i trapindex
```

After an available index is located, enter the following command to enable an SNMP trap:

```
racadm config -g cfgTraps -o cfgTrapsEnable
   -i trapindex 1
racadm config -g cfgTraps -o cfgTrapsDestIpAddr
   -i trapindex IP_Address
racadm config -g cfgTraps -o cfgTrapsSnmpCommu-
   nity -i trapindex Community_Name
```

To create a test trap, enter the following command:

```
racadm testtrap -i trapindex
```

Enabling powerful and flexible management of modular systems

Dell provides several methods for accessing the DRAC/MC, enhancing management of the Dell PowerEdge 1855 blade server. Using the serial console CLI, `racadm` CLI, and the Web-based GUI, administrators can configure, monitor, and manage the Dell PowerEdge 1855 blade server both locally and remotely. By offering powerful and flexible management options for the modular Dell PowerEdge 1855 blade server, Dell helps simplify the management of multiple server blades through a single management interface that seamlessly integrates into an existing management network.

Michael Brundridge is a technologist in the Dell Enterprise Software Development Group. He attended Texas State Technical College and has a technical degree from Southwest School of Electronics.

Ryan Putman is a platform developer for the Dell Enterprise Server Group. He has a bachelor's degree in Electrical Engineering from Vanderbilt University and a master's degree in Computer Engineering from North Carolina State University.

Enhancing Network Availability and Performance on the Dell PowerEdge 1855 Blade Server Using Network Teaming

Network interface card teaming and LAN on Motherboard teaming can provide organizations with a cost-effective method to quickly and easily enhance network reliability and throughput. This article discusses network teaming on the Dell™ PowerEdge™ 1855 blade server and expected functionality using different network configurations.

BY MIKE J. ROBERTS, DOUG WALLINGFORD, AND BALAJI MITTAPALLI

The modular Dell PowerEdge 1855 blade server integrates up to 10 server blades into a highly dense, highly integrated 7U enclosure, including two Gigabit Ethernet switch modules or Gigabit Ethernet pass-through modules and an embedded remote management module. Each integrated switch module is an independent Layer 2 switching device, with 10 internal ports connected to the integrated LAN on Motherboards (LOMs) on the server blades and six 10/100/1000 Mbps external uplink ports (a 10:6 ratio). Each integrated pass-through module is an independent 10-port device directly connecting the individual server blade's integrated LOM to a dedicated external RJ-45 port (a 1:1 ratio). Unlike a switch module, a pass-through module can connect only to 1000 Mbps ports on external switches; pass-through modules do not offer support for 10 Mbps or 100 Mbps connections.

Each server blade in the Dell PowerEdge 1855 blade server has two embedded LOMs based on a single dual-port Intel® 82546GB Gigabit[1] Ethernet Controller. The two LOMs reside on a 64-bit, 100 MHz Peripheral Component Interconnect Extended (PCI-X) bus. The LOMs are hardwired to the internal ports of the integrated switches or pass-through modules over the midplane, a passive board with connectors and electrical traces that connects the server blades in the front of the chassis with the infrastructure in the rear. The LOMs provide dedicated 1000 Mbps full-duplex connections (see Figure 1). LOM 1 on each server blade connects to an internal port of switch 1 or pass-through module 1, and LOM 2 on each server blade connects to the counterpart port of switch 2 or pass-through module 2. *Note:* The second switch or pass-through module is optional. However, two installed switches or pass-through modules can enable additional connectivity or network redundancy and fault tolerance provided that the limitations and capabilities of these features are fully understood and implemented correctly.

One important distinction between a blade server and other types of servers is that the connection between the LOM and internal ports of the integrated I/O module (switch or pass-through) is hardwired through the midplane. This design enables the link between the LOM and the integrated switch to be almost always in an up, or *connected,* state—unless either a LOM or a switch port fails. The link can remain active even in the absence of a network connection between the external uplink ports on the integrated switch and the external network.

[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.

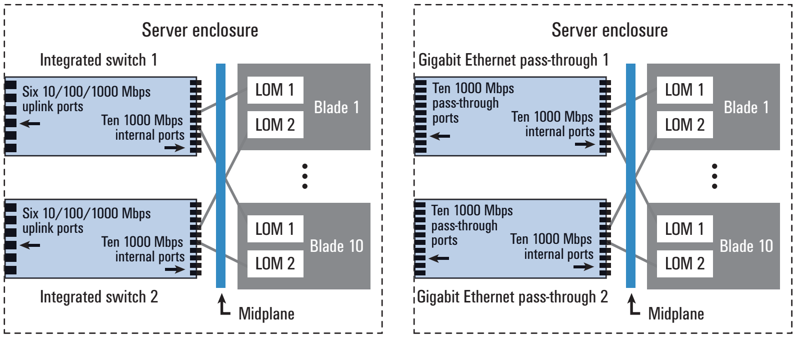

Server enclosure
Integrated switch 1
Six 10/100/1000 Mbps uplink ports
Ten 1000 Mbps internal ports
Six 10/100/1000 Mbps uplink ports
Ten 1000 Mbps internal ports
Integrated switch 2
LOM 1
LOM 2
Blade 1
LOM 1
LOM 2
Blade 10
Midplane
Server enclosure
Gigabit Ethernet pass-through 1
Ten 1000 Mbps pass-through ports
Ten 1000 Mbps internal ports
Ten 1000 Mbps pass-through ports
Ten 1000 Mbps internal ports
Gigabit Ethernet pass-through 2
LOM 1
LOM 2
Blade 1
LOM 1
LOM 2
Blade 10
Midplane

Figure 1. Dell PowerEdge 1855 blade server architecture showing integrated switches and pass-through modules

The connection between the LOM and internal ports of the integrated I/O module is critical in a network teaming scenario where the integrated switch is being used. This is because, to trigger a failover event, the teaming software is looking for *only* the loss of the link between the LOM and the first switch. (Failures to cables or switch ports outside the enclosure do not trigger failover events.)

The scenario is different if a pass-through module is used. With a pass-through module, the link is in a connected state only if a network connection exists between the external ports on the pass-through module and the switch port outside the enclosure—as is true for a stand-alone server. Thus, for the pass-through module, the teaming software triggers a failover event if the LOM, pass-through port, cable, or external switch port fails.

NIC and LOM teaming benefits

Network interface card (NIC) teaming and LOM teaming can help ensure high availability and enhance network performance. Teaming combines two or more physical NICs or LOMs on a single server into a single logical device, or *virtual adapter,* to which one IP address can be assigned. The teaming approach requires teaming software—also known as an intermediate driver—to gather physical adapters into a team that behaves as a single virtual adapter. In the case of the Dell PowerEdge 1855 blade server, the intermediate driver is Intel Advanced Networking Services (iANS). Intermediate drivers serve as a wrapper around one or more base drivers, providing an interface between the base driver and the network protocol stack. In this way, the intermediate driver gains control over which packets are sent to which physical interface on the NICs or LOMs as well as other properties essential to teaming.

> One important distinction between a blade server and other types of servers is that the connection between the LOM and internal ports of the integrated I/O module (switch or pass-through) is hardwired through the midplane.

A virtual adapter can help provide fault tolerance (depending on the type of failure) and bandwidth aggregation (depending on the teaming mode selected). If one of the physical NICs or LOMs—or the internal switch ports to which they are connected—fails, then the IP address remains accessible because it is bound to the logical device instead of to a single physical NIC or LOM.

To provide fault tolerance, when a team is created iANS designates one physical adapter in the team as the primary adapter and the remaining adapters as secondary. If the primary adapter fails, a secondary adapter assumes the primary adapter's duties. There are two types of primary adapter:

- **Default primary adapter:** If the administrator does not specify a preferred primary adapter in the team, iANS chooses an adapter of the highest capability (based on model and speed) to act as the default primary adapter. If a failover occurs, a secondary adapter becomes the new primary. Once the problem with the original primary is resolved, traffic does not automatically restore to the original default primary adapter. The restored adapter will, however, rejoin the team as a secondary adapter.
- **Preferred primary adapter:** The administrator can specify a preferred adapter using the Intel PROSet tool. Under normal conditions, the primary adapter handles all traffic. The secondary adapter receives fallback traffic if the primary fails. If the preferred primary adapter fails but is later restored to an active status, iANS automatically switches control back to the preferred primary adapter.

Teaming features include failover protection, traffic balancing among team members, and bandwidth increases through aggregation. *Note:* Administrators can configure any team by using the Intel PROSet tool, which provides configuration capability for fault tolerance, load balancing, and link aggregation.

Fault tolerance. Fault tolerance provides NIC or LOM redundancy by designating a primary adapter within the virtual adapter and utilizing the remaining adapters as backups. This feature is designed to ensure server availability to the network. When a primary adapter loses its link, the intermediate driver fails traffic over to the secondary adapter. When the preferred primary adapter's link is restored, the intermediate driver fails traffic back to the primary adapter.

The intermediate driver uses link-based tolerance and probe packets to detect network connection failures, depending on the teaming mode selected:

- **Link-based tolerance:** Using this mechanism, the intermediate driver checks the link status of the local network interfaces belonging to the team members. Link-based tolerance provides failover and failback for physical adapter link failures only.
- **Probe packets:** Probing is another mechanism used to maintain the status of the adapters in a fault-tolerant team. Probe packets are sent to establish known, minimum traffic between adapters in a team. At each probe interval, each adapter in the team sends a probe packet to other adapters in the team. Probing provides failover and failback for physical adapter link failures as well as external network failures in the single network path of the probes between the team members.

Note: When only two members exist in the team, the intermediate driver uses link-based tolerance and receive-traffic activity to detect network connection failures. Probing is not used for failover or failback in teams that have only two members.

Load balancing. Adaptive Load Balancing (ALB) mode in iANS provides transmission load balancing by dividing outgoing traffic among all the NICs or LOMs, and can shift traffic away from any NIC or LOM that goes out of service. Receive Load Balancing (RLB) mode in iANS balances receive traffic.

Link aggregation. This feature combines several physical channels into one logical channel. Link aggregation is similar to ALB, and is available in two modes: Static Link Aggregation mode supports Cisco Fast EtherChannel (FEC) and Cisco Gigabit EtherChannel (GEC), while IEEE 802.3ad mode supports the IEEE 802.3ad standard.

Note: Link aggregation is not supported on PowerEdge 1855 blade servers configured with integrated switches because of the server hardware design implementation. For more details, refer to the “Static Link Aggregation” and “IEEE 802.3ad” sections in this article. Other blade server systems that have hardware and network configurations similar to the PowerEdge 1855 blade server tend to function in a similar manner.

Team mode	Integrated switches	Integrated pass-throughs
AFT	Yes*	Yes
SFT	Yes*	Yes
ALB and RLB	Yes*	Yes
SLA**	No	Yes
IEEE 802.3ad**	No	Yes

*This mode cannot detect connectivity failures outside the chassis.

**SLA and IEEE 802.3ad are not supported on Dell PowerEdge 1855 blade servers when using integrated switches because the two NICs or LOMs on the blade are hardwired to different integrated switches. However, these modes are supported when using the pass-through modules.

Figure 2. Teaming support matrix for the Dell PowerEdge 1855 blade server

iANS teaming software architecture

iANS supports the following teaming modes:

- Adapter Fault Tolerance (AFT)
- Switch Fault Tolerance (SFT)
- Adaptive Load Balancing and Receive Load Balancing
- Static Link Aggregation (SLA)
- IEEE 802.3ad

Figure 2 is the teaming support matrix for PowerEdge 1855 blade servers.

Adapter Fault Tolerance

AFT enables automatic recovery from a link failure caused by a failure in a physical NIC or LOM, and internal ports on the integrated switch, by redistributing the traffic load across a backup adapter. Failures are detected automatically, and traffic rerouting takes place as soon as the failure is detected. The goal of AFT is to ensure that load redistribution takes place fast enough to prevent user protocol sessions from being disconnected. AFT supports two to eight adapters (any combination of NICs and LOMs) per team. Only one active team member transmits and receives traffic. If this primary connection fails, a secondary, or *backup,* adapter takes over. After a failover, if the connection to the primary adapter is restored, control passes automatically back to the primary adapter. AFT is the default mode when a team is created using iANS; however, this mode does not provide load balancing.

Teaming combines two or more physical NICs or LOMs on a single server into a single logical device, or *virtual adapter,* to which one IP address can be assigned.

The Intel intermediate driver uses the immediate physical loss of a link—either between the NIC or LOM and the integrated switch or between the NIC or LOM, the pass-through module inside the server enclosure, the cable, and the port of the external switch to which it is connected—to trigger a failover when there are only two adapters in the team. The “X”s in Figure 3a and Figure 3b indicate the failover and failback for link failures. Notice that, in Figure 3a, a failover between the LOM and the external port link does not occur for switches. This is because, as previously discussed, the internal ports of the integrated switches are hardwired to the LOMs through the midplane of the enclosure; therefore, the intermediate driver does not detect a loss of link.

Switch Fault Tolerance

SFT functionality is similar to AFT except that SFT supports only two NICs or LOMs in a team connected to two different switches. In SFT mode, one adapter is the primary adapter and one adapter is the secondary adapter. During normal operation, the secondary adapter is in standby mode. In its standby state, the adapter is inactive and waiting for failover to occur, and the adapter does not transmit or receive network traffic. If the primary adapter loses connectivity, the secondary adapter automatically takes over.

The local network interfaces in Figure 3a show the links between the LOMs and the internal ports of the integrated switches, and Figure 3b shows the links between the LOMs and the internal ports of the pass-through modules.

If a team has more than two members, enabling probe packets can help detect network connection failures, along with the link-based method of fault tolerance. When probes are enabled, the primary adapter will send the probe packets to secondary adapters and vice versa at administrator-defined time intervals. When the primary adapter's probe packets are not received by the secondary adapters, the intermediate driver attempts retry mechanisms before eventually failing over. Probe packets can help detect network failures in the single network path of probe packets between the primary and secondary adapters.

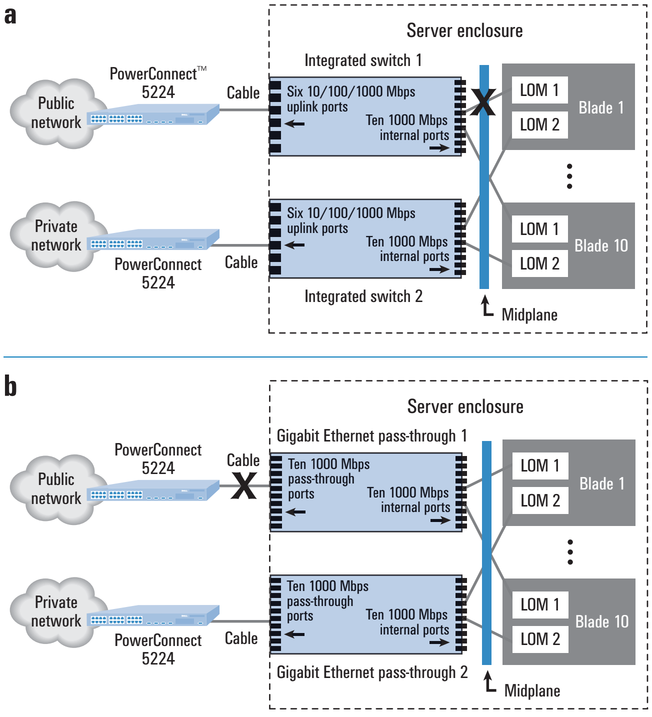


Figure 3. Dell PowerEdge 1855 blade server architecture showing potential failover points

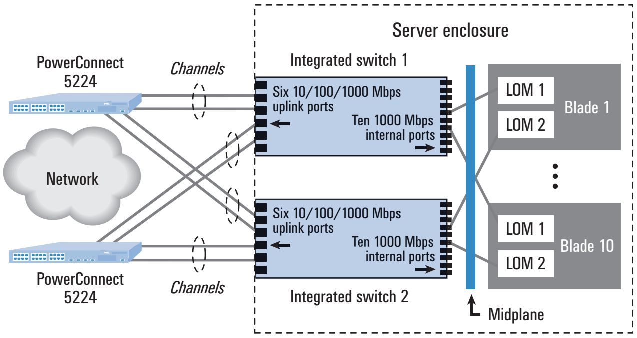


Figure 4. Configuration for redundancy with two external switches

Additional network redundancy. Because the AFT or SFT teams cannot probe the link status of the integrated switch's external uplink ports, or link failures elsewhere in the network, additional protection may be required. Administrators can achieve increased network redundancy by adding redundant links between the integrated switches and external switches or by aggregating at least two of the uplink ports using the Link Aggregate Control Protocol (LACP). (*Note:* The external switches must support and be configured for 802.3ad link aggregation.) When properly configured with link aggregation and adapter fault-tolerance teams, the PowerEdge 1855 blade server is designed to maintain network connectivity if any of the LOMs, internal switches, or switch uplink ports fail.

The PowerEdge 1855 blade server can be configured to provide a highly redundant network environment when plugged into a single external switch. However, administrators may want to provide a further level of redundancy to protect against the failure of an external switch. Figure 4 shows the necessary network connections for the PowerEdge 1855 blade server in such an environment. Administrators can implement this redundancy by creating two dual-port link aggregation groups between each external switch and integrated switches 1 and 2. In this configuration, even if an external switch fails, individual server blades maintain network connectivity with the help of Spanning Tree Protocol (STP) in the switches. Configuration of STP is very important for this network configuration to work reliably.

Adaptive Load Balancing and Receive Load Balancing

ALB is a method for dynamic distribution of data traffic load among multiple physical channels. The purpose of ALB is to improve overall bandwidth and end-node performance. The ALB approach provides multiple links from the server to the switch, and the intermediate driver running on the server performs the load-balancing function.

The ALB architecture uses knowledge of Layer 3 information, such as IP address, to achieve optimum distribution of the server transmission load.

ALB is implemented by assigning one of the physical channels as "primary" and all other physical channels as "secondary." Packets leaving the server can use any one of the physical channels, but incoming packets can use only the primary channel. When enabled, RLB balances the IP receive traffic. The intermediate driver analyzes the send and transmit loading on each physical adapter in the team and balances the rate across all adapters based on the destination address. Adapter teams configured for ALB also provide the benefits of fault tolerance, as described in the "Adapter Fault Tolerance" section in this article.

Static Link Aggregation

Link aggregation is similar to ALB in that it combines several physical channels into one logical channel. The Intel intermediate driver supports link aggregation for FEC and GEC.

FEC is a trunking technology developed by Cisco to aggregate bandwidth between switches working in Fast Ethernet. Using FEC, administrators can group multiple switch ports together to provide additional bandwidth. Switch software treats the grouped ports as a single logical port. Administrators can connect an end node, such as a high-speed server, to the switch using FEC. FEC link aggregation provides load balancing in a similar manner to ALB, including the use of the same algorithm in the transmit flow.

The PowerEdge 1855 blade server can be configured to provide a highly redundant network environment when plugged into a single external switch.

The transmission speed for FEC does not exceed the adapter base speed to any single address; teams must match the capability of the switch. Adapter teams configured for SLA also provide the benefits of fault tolerance and load balancing. When using SLA mode, administrators are not required to set a preferred primary adapter. All adapters in an SLA team must run at the same speed and must be connected to an SLA-capable switch. If the speed capability of adapters in an SLA team is different from the speed of the switch, the speed of the team is dependent on the switch. SLA teaming requires that the switch be set up for SLA teaming and that STP be turned off.

GEC link aggregation is essentially the same as FEC link aggregation, except that GEC supports 1000 Mbps speeds versus 100 Mbps for FEC.

IEEE 802.3ad

IEEE 802.3ad is the IEEE standard for the technology incorporated in the Cisco FEC standard. Intel's intermediate driver support for IEEE 802.3ad is similar to its FEC and GEC support. Administrators can configure a maximum of two IEEE 802.3ad dynamic teams per server, and the configuration must use 802.3ad-capable switches in dynamic mode. Adapter teams configured for IEEE 802.3ad also provide the benefits of fault tolerance and load balancing. The 802.3ad teaming mode allows all network communication protocols to be load balanced.

Dynamic mode supports multiple aggregators, and adapters should operate at the same speed. Only one team can be active at a time.

The advantages of modular computing

The Dell PowerEdge 1855 blade server can help organizations deliver on the promise of modular computing—namely lower acquisition cost, lower total cost of ownership, rack density, and power efficiency—all without trading off enterprise-class features. Moreover, by understanding the unique aspects of the PowerEdge 1855 blade server's architecture, as described in this article, administrators can create a highly reliable network infrastructure.

Mike J. Roberts is responsible for blade server product planning at Dell. Mike has a B.A. from Marquette University and an M.B.A. from The University of Texas at Austin.

Doug Wallingford is a sustaining engineer in the Enterprise International Product Support department at Dell. Doug has 14 years of experience in network field support, seven years as a Xerox field engineer, and seven years at Dell. He has numerous Microsoft® and Novell® certifications—including Microsoft Certified Systems Engineer (MCSE), Microsoft Certified Systems Administrator (MCSA), and Certified Novell Administrator (CNA).

Balaji Mittapalli is a development engineer in the Server Networking and Communications department at Dell. Balaji has an M.S. in Electrical Engineering from The University of Texas at San Antonio and a B.S. in Electronics and Communication Engineering from S.V. University in India. Balaji's areas of interest are wired and wireless networks.

More data? Less time?

No problem.

Dell | Enterprise

Dell LTO-2 products keep pace with your growing backup needs.

Your organization's data may be increasing exponentially, but your backup window isn't. Chances are, your IT staff must perform backups overnight in only four to five hours—no matter how much data is involved.

Dell helps enterprises meet the challenges of rapid data growth and shrinking backup windows with the Dell™ PowerVault™ series of Ultrium® 2 LTO (LTO-2) tape backup products. Featuring powerful second-generation Linear Tape-Open™ (LTO®) technology, Dell LTO-2 products are optimized to provide both speed and capacity:

Dell™ PowerVault™ 132T tape library with LTO-2 tape drives

- **Outstanding performance:** A data transfer rate of up to 70 MB/sec.[1]
- **High-capacity media:** Up to 400 GB capacity per cartridge can mean fewer cartridges to purchase and store, which can help save you money.[2]
- **Open standards:** LTO open standards foster competitive pricing and a format that has broad industry acceptance.

Best of all, the LTO roadmap projects the potential to double capacity and performance in each successive generation, and supports comprehensive backward compatibility.[3] Dell LTO-2 drives are available in the PowerVault 110T (as a stand-alone unit or configured internally on select PowerEdge™ servers) and in Dell PowerVault 132T, 136T, and 160T LTO-2 tape libraries. To learn more, visit www.dell.com/storage.

Dell™ PowerVault™ 110T LTO-2 tape drive and 136T, 160T LTO-2 tape libraries

ULTRIUM LTO 2

Easy as DELL™

Click www.dell.com/storage

[1] Assumes 2:1 compression. Data compression rate may vary depending on settings, user environment and applications. For more information, visit http://www1.us.dell.com/content/products/productdetails.aspx/pvaul_110tlto2?c=us&cs=555&l=en&s=biz&~tab=specstab#tabtop.

[2] Assumes 2:1 compression. Data compression rate may vary depending on settings, user environment and applications. For more information about LTO format and capacity, visit http://www.lto.org/newsite/html/format.html.

[3] For more information about LTO technology and the LTO roadmap, visit http://www.lto.org.

VMware Virtualization Software

on Dell Blade Servers

Dell blade servers together with VMware virtualization software can help meet the scalability requirements of today's data centers. This article discusses how Dell™ PowerEdge™ 1855 blade servers and VMware® ESX Server™ virtualization software can facilitate server consolidation and supply on-demand server provisioning. In addition, virtual machines running on blade server systems can help reduce total cost of ownership compared to software running on stand-alone systems.

BY BALASUBRAMANIAN CHANDRASEKARAN AND SIMONE SHUMATE

IT organizations must be able to respond quickly to ever-changing business needs—all while accommodating growing system demands with minimal downtime. Consequently, scalability has become a primary requirement for the enterprise data center. Efficient scalability includes two key capabilities: first, to grow easily from n to $n + 1$ resources with minimal disruption to the existing IT configuration; second, to redeploy the workload quickly and conveniently as extra resources are added. Because hardware changes frequently require reconfiguration of the software running the workload, redistributing workloads across newly added resources can be a challenging task.

Other emerging requirements in the enterprise data center are isolation and consolidation. Ideally, applications running on the same system should not interfere with one another. However, as system administrators are well aware, this is not always the case. Today's operating systems do not guarantee that multiple applications running on the same physical server can operate in isolation, unaffected by each other's presence. To provide applications with isolation, data centers often deploy a dedicated server, such as a standards-based Dell PowerEdge server, to run each application.

Consolidation allows administrators to host multiple services or applications on fewer physical systems. Consolidation enables savings in floor space and cabling costs, and enables streamlined systems management, optimal system utilization and cooling, and other environmental benefits. As administrators well know, consolidation and isolation are not mutually exclusive requirements.

Consolidation benefits of Dell blade servers

As a key aspect of its scalable enterprise strategy, Dell has recently released a new generation of Dell PowerEdge blade servers. In blade server systems, the granularity of a computing unit is reduced to a single server blade. Individual server blades can be combined in the same blade server chassis to provide an aggregated computing system, thereby helping to reduce total cost of ownership (TCO) compared to stand-alone servers. Administrators can cost-effectively add a single server blade to an existing blade server system, which helps keep both initial capital investment and ongoing management burden low. Thus, blade server systems can provide the capability to quickly and conveniently scale out using interchangeable, industry-standard components.

Dell PowerEdge 1855 blade servers enable consolidation—a key requirement for many data centers—at the platform level. Blade server systems offer impressive computing power in a small form factor. In addition to housing up to 10 blades in a 7U enclosure, the PowerEdge 1855 blade server integrates Fibre Channel pass-through modules and Ethernet switch modules into the chassis to minimize physical space requirements and facilitate cable management. A single server blade can be dedicated to each application, providing the needed isolation for individual applications.

Dynamic business needs and sudden changes in workload volume can easily be addressed through blade server systems, which enable administrators to add individual server blades as the need arises, facilitating a pay-as-you-grow approach for data centers. However, this approach can still result in underutilized server blades. To make blade server systems truly effective, IT organizations must address the issue of transferring existing workloads as extra server blades are brought online. Such transfers must be accomplished dynamically, with minimal disruption to existing services.

Enhanced manageability through server virtualization

Just as blade servers enable a single chassis to host multiple physical servers, server virtualization software can enable a single physical server to host multiple operating systems. As discussed previously, blade server systems can address critical requirements for today's data centers: consolidation, utilization, and scalability.

Just as blade server systems enable enterprises to grow incrementally by adding server blades into a blade server chassis as the need arises, virtualization software enables administrators to create or delete, as required, a *virtual machine* (VM)—that is, one virtualized server running an operating system and application instance in isolation from other virtualized servers running on the same physical platform.

Consolidation. VMware ESX Server virtualization software enables multiple instances of the same operating system—or a variety of different operating systems—to execute on the same physical server. Under VMware ESX Server, each operating system can host its own application, running the application in isolation even though the underlying hardware is shared. In this way, virtualization software can provide another means to consolidate physical servers. Moreover, by using virtualization software on blade server systems, administrators can achieve an even higher level of consolidation than they could on stand-alone servers by installing multiple operating systems on individual server blades—and multiple server blades in each blade server chassis.

Utilization. The physical consolidation that is possible using virtualization software on blade server systems can help conserve data center space, decrease energy costs, and simplify cabling compared to stand-alone servers while enabling high CPU utilization rates for system resources. In addition to consolidation, high system utilization can be achieved through virtualization in blade server systems.

Scalability. The benefits of consolidation using virtualization software on blade server systems are well known, but organizations can derive additional advantages when other aspects of virtualization are taken into account—namely, replication and relocation of VMs. These benefits address scalability and manageability needs, which can become critical for data centers faced with the requirement to grow and change dynamically in response to changing business needs.

Seamless scalability is a primary benefit of implementing virtualization on blade server systems. As additional server blades are added to the chassis, VMs can be moved to newly added server blades with a few mouse clicks, thereby enabling *dynamic provisioning* of workloads. The chassis of a Dell PowerEdge 1855 blade server is designed to be populated with up to 10 server blades as business needs increase. Administrators can easily manage workloads by moving VMs from existing server blades to newly added server blades using VMware VMotion™ technology. This technology is designed to enable *hot migration,* which allows administrators to move a running application from one server blade to another server blade without stopping either the operating system or applications to allow the migration. This capability helps organizations scale data center resources easily without incurring downtime.

Besides scalability, VMotion enables improved overall manageability. Because operating systems running on the VMs are isolated from the underlying physical hardware by a layer of virtualization software, they are not tied to physical server blades. As a result, server blades can be repaired and upgraded as needed without affecting services, simply by moving the VMs from one server blade to another.

A flexible infrastructure for today's data centers

The Dell PowerEdge 1855 blade server and VMware ESX Server virtualization software can be an effective combination to help achieve data center consolidation, while enhancing scalability and manageability. Advanced features enabled by VMware software—such as the capability to migrate a running application from one physical server to another while maintaining a high level of system availability—can be leveraged by Dell blade server systems to address mission-critical performance and scalability issues affecting today's data centers. Using VMware server virtualization software on Dell blade server systems can help organizations create a flexible IT infrastructure that enables dynamic server provisioning and supports seamless scaling in the enterprise data center.

Balasubramanian Chandrasekaran is a systems engineer in the Scalable Enterprise Lab at Dell. He has an M.S. in Computer Science from Ohio State University.

Simone Shumate is a senior systems engineer with the Dell Enterprise Solutions Engineering Group, where she leads the Scalable Enterprise engineering team. She has a B.S. in Computer Engineering from the University of Kansas.

Efficient BMC Configuration

on Dell PowerEdge Servers Using the Dell Deployment Toolkit

The on-board baseboard management controller (BMC) is a powerful and flexible device that can be used to effectively manage eighth-generation Dell™ servers such as the PowerEdge™ 1850, PowerEdge 2800, and PowerEdge 2850. This article provides guidance on configuring the BMC through the feature set provided by the Dell OpenManage™ Deployment Toolkit.

BY ANUSHA RAGUNATHAN, ALAN BRUMLEY, AND RUOTING HUANG

Eighth-generation Dell PowerEdge servers include an on-board baseboard management controller (BMC) that complies with the industry standard Intelligent Platform Management Interface (IPMI) 1.5 specification. The BMC monitors the server for critical events by communicating with various sensors on the system board, and sends alerts and log events when certain parameters exceed their preset thresholds—thus enabling administrators to configure, monitor, and recover systems remotely.

The BMC is designed to perform the following platform management functions:

- Remote access to the BMC through the system's serial port and integrated network interface card (NIC)
- Fault logging and Simple Network Management Protocol (SNMP) alerting through Platform Event Filters (PEFs)
- Access to the system event log (SEL) and access to sensor status information
- Control of system functions, including power-up and power-down
- Support that is independent of the system's power-up operating state
- Text console redirection for system setup, text-based utilities, and operating system (OS) consoles
- Access to the Microsoft® Emergency Management Services (EMS), Microsoft Special Administration Console (SAC), and Red Hat® Linux® serial console interfaces using serial over LAN (SOL)

Dell provides several options for configuring or managing the BMC. A primary criterion for selecting an application is its scope of execution; Figure 1 lists application tools that are suited to the needs of various types of organizational infrastructures.

Understanding the DTK on eighth-generation PowerEdge servers

One option for configuring the BMC is the Dell OpenManage Deployment Toolkit (DTK). The DTK provides DOS command-line utilities that facilitate manual, automatic, and unattended server deployment. DTK utilities support the configuration of BIOS, RAID, BMC, utility partitions, NICs, and OS installation files through a command-line interface (CLI) or an initialization file. The configuration can be stored to this file and then read back to clone a BMC configuration on a new system.

DTK utilities are typically used in large server environments where automation is highly valued. This automation is accomplished through the use of batch files and unattended boot processes.

The DTK 1.3 release provides support for the BMC on eighth-generation Dell PowerEdge servers by adding functionality to its suite of existing DOS CLI tools—namely, a new utility for BMC configuration called bmccfg. This release also enhances the existing bioscfg BIOS configuration tool to support the BMC, particularly the serial port configuration parameters used with the BMC. In addition, control of the power and non-maskable interrupt (NMI) buttons has been moved to the bmccfg utility.

The bmccfg utility is an executable file (bmccfg.exe) that can be used to perform the following tasks:

- Configure users and access levels
- Configure the local area network (LAN), serial interfaces, SOL, and panel buttons
- Configure alerts and response actions and clear the SEL
- Query globally unique identifiers (GUIDs) on systems and firmware

Using the DTK to configure the BMC to perform management tasks

The bmccfg utility can be used to configure the BMC and perform basic management tasks. Proper configuration of the BMC is necessary so that it can be used by other BMC management applications such as Dell OpenManage IT Assistant (ITA). Figure 2 details the management application and the corresponding management tasks that can be accomplished after the initial BMC configuration using bmccfg.

Application	Type	Scope	Description
BMC setup module	Configuration	Local	Module that provides basic BMC setup during power-on self-test (POST)
DTK bmccfg tool	Configuration	Local	Powerful CLI tool to configure the BMC in pre-OS environments
Dell OpenManage Server Administrator (OMSA)	Configuration	Local and remote	Web application that performs one-to-one platform management to configure the BMC in post-OS environments; OMSA running locally can be accessed remotely from a management station using a supported browser
ipmish IPMI management utility	Management	Remote	Console application for the control and management of remote systems, which provides both LAN channel and serial channel access to the BMC
SOL Proxy IPMI management utility	Configuration and management	Remote	Utility that provides LAN-based administration of remote systems using SOL
Dell OpenManage IT Assistant	Management	Remote	Application that serves as an SNMP alert–receiving management station

Figure 1. Tools to configure or manage the BMC

BMC management application	BMC management tasks
ipmish IPMI management utility (BMC LAN channel and serial channel access)	Remote SEL access Power control System identification SOL activation
Console software (BMC serial channel access)	SEL access Power control System identification System information
SOL Proxy IPMI management utility (BMC SOL access)	Text utility console redirection Remote BIOS setup Microsoft and Red Hat Linux text console redirection
ITA (PET receiver)	SNMP traps

Figure 2. BMC management applications and corresponding management tasks

Configuring the BMC for out-of-band logon for use by ipmish

To accommodate out-of-band remote management, the BMC allows configuration of users. Each user is associated with a username, password, and privilege level, which can be different for the LAN channel and the serial channel. The BMC allows for nine users, which can be configured as described in this section.

For purposes of configuring the BMC, a user's information, name, and password are all organized according to a numeric user ID. To assign a username to a user ID, use the following command (where *x* is a number from 2 to 10):

```
bmccfg username --userid=x --name=usernamestring
```

To assign a password to user ID 3, use the following command:

```
bmccfg passwordaction --userid=3
   --action=setpassword --password=mypassword
```

To confirm a password, use the following command:

```
bmccfg passwordaction --userid=3
   --action=testpassword --password=mypasword
```

Because the password is spelled incorrectly, this command will generate a "Password test failed" error message.

BMC users can be disabled so that they do not have LAN or serial access to the BMC. To do so, use the following command:

```
bmccfg useraction --userid=x --action=disable
```

To enable access for user ID *x*, use the following command:

```
bmccfg useraction --userid=x --action=enable
```

Configuring LAN channel settings for use by an IPMI management utility

To manage the BMC over a LAN channel using a management utility such as IPMI Shell (ipmish), the configurations described in this section must be performed.[1]

Enabling IPMI over the BMC LAN channel. To enable IPMI access to the BMC over the LAN channel, use the following command:

```
bmccfg lanchannelaccess --ipmioverlan=alwaysavail
```

This command allows all inbound IPMI management traffic into the BMC. Note the following critical information:

- The BMC management traffic will not travel into the BMC if LAN on Motherboard (LOM) is used in an EtherChannel or link aggregation teaming strategy.
- The IPMI management traffic traveling outbound from the BMC is configured differently (using the `--pefalerting` option in the `lanchannelaccess` subcommand).

To disable IPMI access to the BMC over the LAN channel, use the following command:

```
bmccfg lanchannelaccess --ipmioverlan=disable
```

Setting up privilege levels for the LAN channel and LAN users. The BMC uses four levels of privilege: Administrator, Operator, User, and No Access. Each user is given a privilege level, which can be updated at any time. Each IPMI command has a required privilege level that must be met before execution is allowed. To determine which levels are required for each command, consult the IPMI specification.[2]

The LAN and serial channels can also be capped at a specific privilege level. For instance, the LAN channel could be configured to allow only User access. In this case, even if an administrator with User access connects through the LAN, the administrator will only have User access privileges. LAN access to BMC users can be restricted in the following ways:

- Restrict the LAN channel privilege limit to Administrator, Operator, User, or No Access using the following command:

```
bmccfg lanchannelaccess --channelprivlmt
```

- Restrict the user privilege limit on the LAN channel to Administrator, Operator, User, or No Access using the following command:

```
bmccfg lanuseraccess --userid=x --usrprivlmt
```

Serial channel access is configured and restricted in the same manner as that described for LAN access, except that administrators should use the `serialchannelaccess` command instead of the `lanchannelaccess` command. *Note:* Be sure to monospace the delimiters.

Configuring other parameters of the LAN channel. The other necessary configurations that must be made before using an IPMI management utility such as ipmish over the LAN channel are described in this section.

To obtain a Dynamic Host Configuration Protocol (DHCP) address for the BMC LAN, use the following command:

```
bmccfg lancfgparams --ipaddrsrc=dhcp
```

Note: If the BMC sends a request for an IP address and never receives an answer, it defaults to IP address 169.254.0.2. This IP address indicates that the BMC was unable to obtain a DHCP address.

To set up the BMC LAN to have a static IP address, assign an IP address (such as 192.168.100.10), subnet mask (such as 255.255.255.0), and gateway address (such as 192.168.100.1) by using a command similar to the following:

```
bmccfg lancfgparams --ipaddrsrc=static
   --ipaddress=192.168.100.10 --subnetmask=
   255.255.255.0 --gateway=192.168.100.1
```

Once these steps have been performed, administrators can issue `ipmish` commands. For example, to obtain the SEL, use the following command:

```
ipmish -ip bmcipaddress -u username -p password
   sel get
```

Configuring serial channel settings for use by an IPMI management utility

To manage the BMC over a serial channel using a management utility such as ipmish, the configurations described in this section must be performed.

Enabling IPMI over the BMC serial channel. To set up the BMC serial port to be used for IPMI management traffic, use the following command:

```
bioscfg --serial1=bmcserial
```

Note: The system must be rebooted for the changes to take effect.

[1] For more information about ipmish, refer to support.dell.com/support/edocs/software/smbmcmu/en/index.htm.

[2] For more information about the IPMI specification, visit developer.intel.com/design/servers/ipmi.

To enable IPMI access to the BMC serial port, use the following command:

```
bmccfg serialchannelaccess --ipmioverserial=
    alwaysavail
```

To disallow access to the BMC serial port, use the following command:

```
bmccfg serialchannelaccess --ipmioverserial=disable
```

Communication with the serial port can occur in either of two modes: basic or terminal. Basic mode requires an IPMI-aware serial application such as ipmish to communicate with the BMC, while terminal mode uses any console software application for sending and receiving raw IPMI commands and a limited set of ASCII text commands.

When using ipmish, administrators should make sure that the connection mode in the terminal configuration is set to basic mode by using the following command:

```
bmccfg serialcfgparams --connectionmode=basic
```

Configuring other parameters of the serial port. This section describes other necessary configurations that must be made before using an IPMI management utility, such as ipmish, over the serial channel.

To configure the flow control and the baud rate for the serial port, use the `msgcommflowctrl` and `msgcommbitrate` subcommands in the bmccfg utility. Once these steps have been performed, administrators can issue `ipmish` commands. For example, the following command powers down the managed node using the first serial port of the management station at 19,200 bps with hardware flow control:

```
ipmish -com 1 baud 19200 -flow cts -u username
    -p password power off
```

Configuring serial channel settings for use by console software

Console software such as Hilgraeve HyperTerminal can be used to communicate with the BMC in terminal mode. The main advantage of switching the BMC serial port to terminal mode is that terminal mode provides a printable, text-based mechanism for communicating IPMI messages between console software and the BMC. This makes it easy to develop scripted tools for generating IPMI messages to the BMC and receiving IPMI messages from the BMC. Also, the BMC provides an ASCII text–based command set to perform a subset of systems management functions.

To switch the multiplexer to route traffic to the BMC's serial port, use the following command:

```
bioscfg --serial1=bmcserial
```

To set the connection mode for `serialcfgparams` to terminal, use the following command:

```
bmccfg serialcfgparams --connectionmode=terminal
```

To make sure that the baud rate set in the console software is in sync with the rates set in the BMC serial channel, use the following command:

```
bmccfg serialcfgparams --msgcommbitrate
```

Once terminal mode has been set up, issue either raw IPMI commands such as `[18 00 22]` and then press Enter (to start a new line) to reset the BMC watchdog timer, or issue ASCII text commands such as `[SYS TMODE]` and then press Enter (to start a new line) to query whether the BMC is in terminal mode.

Refer to the IPMI 1.5 specification for more details on the format of IPMI commands and the available set of ASCII text commands.

Configuring SOL settings for use by an IPMI management utility

The BMC provides a mechanism—SOL—that allows the baseboard serial controller of a managed node to redirect its serial data over an IPMI session using IP. SOL enables remote console applications such as SOL Proxy to provide access to interfaces, BIOS, OS, and applications. SOL requires the serial multiplexer to be set to the BMC LAN as follows:

```
bioscfg --serial1=bmclan
```

This sets the multiplexer to connect the baseboard serial controller to the BMC, which forwards serial data to SOL Proxy as IP packets. Enable SOL as follows:

```
bmccfg solcfgparams --solenable
```

To set the SOL baud rate to 9,600 or 19,200 bps, use the following command:

```
bmccfg solcfgparams --solbitrate
```

Setting up the minimum privilege level required by the user—User, Operator, or Administrator—is done using the following command:

```
bmccfg solcfgparams --solprivlevel
```

Also, some advanced configurations of the SOL *character accumulate interval* and *character send threshold* can be set using bmccfg. Character accumulate interval defines the amount of time the BMC will wait before transmitting a partial SOL character data packet. Character send threshold is defined as the number of characters for which the BMC will wait before automatically transmitting the SOL packet. The preceding two parameters are set using the following commands:

```
bmccfg solcfgparams --solcharaccuminterval
bmccfg solcfgparams --solcharsendthreshold
```

Configuring PEF settings for SNMP alerting

The BMC can be configured—by matching against a set of PEFs—either to perform shutdown actions or to send event alerts. Alerts are delivered as SNMP traps in the Platform Event Trap (PET) format. The Dell BMC supports sending alerts to as many as four destinations over the LAN channel.

The BMC can also be used to monitor voltage, fan rpm, temperature, processor health, chassis intrusion status, and the watchdog timer. All of these events can be logged in the SEL and the status of the event log is also monitored. Administrators should consult system documentation to determine which areas of the system can be monitored, because this typically varies from system to system.

To enable the BMC to send PEF alerts to alert destinations, use the following command:

```
bmccfg lanchannelaccess --pefalerting=enable
```

To disable the BMC from sending PEF alerts, use the following command:

```
bmccfg lanchannelaccess --pefalerting=disable
```

To configure the IP address of the destination receiving the SNMP traps, use the following command (where *x* is a number from 1 to 4):

```
bmccfg lancfgparams --alertdest=x --destipaddr=
    IPAddressoftheconsolereceivingtraps
```

To set the community string and host name identifier of the trap, use the following command:

```
bmccfg lancfgparams --commstring
bmccfg pefcfgparams --hostname
```

All filters previously listed can be enabled to alert the destinations by entering the following:

```
bmccfg pefcfgparams --filter=filtername
    --filteralert=enable
```

Valid arguments for the `--filter` suboption include `fanfail`, `voltfail`, `discretevoltfail`, `tempwarn`, `tempfail`, `intrusion`, `redundegraded`, `redunlost`, `procwarn`, `procfail`, `powerwarn`, `powerfail`, `hardwarelogfail`, and `autorecovery`.

For example, `bmccfg pefcfgparams --filter=intrusion --filteralert=enable` will enable filter alerting for the chassis intrusion event. Organizations can use an SNMP management console, such as ITA, to receive traps from such events. To disable alerts and perform only shutdown actions, use `bmccfg pefcfgparams --filter=`*`filtername`* `--filteralert=disable`.

Configuring filters to perform local shutdown actions

BMC filters can also be used to perform local actions. To configure a filter to perform a local shutdown action, enter the following:

```
bmccfg pefcfgparams --filter=filtername
    --filteraction=action
```

Valid arguments for the `--filteraction` suboption include `powercycle`, `powerdown`, `reset`, and `none`.

PEFs can be configured both to perform a shutdown action and to send alerts. In such a case, the alert action is deferred until the shutdown action has been performed.

Managing Dell servers with the BMC

The BMC can be a very powerful and flexible management device. Access to the extensive management capabilities of the BMC can be streamlined through the Dell OpenManage Deployment Toolkit, and standard command-line tools can then be used to quickly check server status.

Anusha Ragunathan is a software engineer in the Dell Product Group and is part of a team that develops deployment tools for PowerEdge servers as part of the Dell OpenManage systems management offerings. Anusha has a master's degree in Computer Science Engineering from Arizona State University in Tempe, Arizona, and a bachelor's degree in Computer Science Engineering from Bharathiyar University in India.

Alan Brumley is a software engineer in the Dell Product Group and works on the Dell OpenManage Deployment Toolkit team. His interests include computerized numerical control (CNC) system design. Alan has a bachelor's degree in Computer Engineering from the University of South Carolina.

Ruoting Huang is a development engineer on the Dell OpenManage Deployment Toolkit team. His interests include parallel processing and internetworking. Ruoting has an M.S. in Computer Science from the Asian Institute of Technology in Thailand.

FOR MORE INFORMATION

IPMI specification:
developer.intel.com/design/servers/ipmi

Remote OS Deployment

Using Dell OpenManage Server Assistant 8 and DRAC 4

Administrators can take advantage of the Dell™ Remote Access Controller 4 (DRAC 4) and Dell OpenManage™ Server Assistant when deploying operating systems remotely on eighth-generation Dell servers. This article provides a step-by-step approach to remotely deploying and configuring a Microsoft® Windows® or Red Hat® Enterprise Linux® operating system on eighth-generation Dell PowerEdge™ servers equipped with the DRAC 4.

BY MICHAEL E. BROWN, MANOJ GUJARATHI, AND GONG WANG

Effectively managing Dell PowerEdge servers is critical to helping ensure the maximum value of an IT infrastructure. Dell OpenManage provides open, flexible systems management tools that can integrate and help standardize and automate server management processes. Dell OpenManage tools can help lower overall management costs by increasing the server-to-administrator ratio through centralized management of distributed systems and remote access to Dell PowerEdge servers at virtually any time. For organizations of every size, Dell OpenManage helps provide a comprehensive set of tools to deploy, monitor, and manage system updates and changes.

Dell OpenManage Server Assistant (DSA) 8.x is an application in the Dell OpenManage suite that facilitates installations for Microsoft Windows, Red Hat Enterprise Linux, and Novell® NetWare® operating systems on Dell PowerEdge and Dell PowerEdge SC servers.[1] This application provides basic operating system (OS) installation features, and helps ensure that Dell-tested device drivers are installed for all supported peripherals. It also helps enterprise IT organizations carry out RAID setup and configuration, network adapter teaming, OS replication, and other customizable installation options.

The Dell Remote Access Controller 4 (DRAC 4) offers remote management capabilities for eighth-generation Dell servers such as the PowerEdge 1850, PowerEdge 2800, and PowerEdge 2850. The DRAC 4 also offers support for console redirection (continuous video) and virtual media through its remote management features. Console redirection can be used to access the system console remotely when the system is in either graphical or text mode. By using virtual media, the administrator can use the CD drive or floppy drive of any system on the network as if it were a local drive on the server.

Hardware requirements and setup for remote deployment using DSA

Carrying out remote deployment of an OS on a managed server with DSA requires an eighth-generation Dell PowerEdge server equipped with a DRAC 4. Administrators also need a management station, which could be any system on the network that has network access to the

[1] For more information about DSA 8.x, see “Using Dell OpenManage Server Assistant 8.x to Optimize Installation of Dell PowerEdge Servers” by Michael E. Brown, Niroop Gonchikar, Nathan Martell, and Gong Wang in *Dell Power Solutions,* June 2004.

managed server. The DRAC 4 must be connected to the network, and the administrator must have login privileges to access and log in to the DRAC 4.

On a bare-metal server, there are several ways to configure the network settings for the DRAC 4. These settings should be configured once, when the server is first provisioned.[2] If the OS is being redeployed on a previously configured system, administrators can skip the steps discussed in the next two sections.

Setting up network connectivity using option ROM

To use option ROM to configure the DRAC 4 network settings, administrators should perform the following steps:

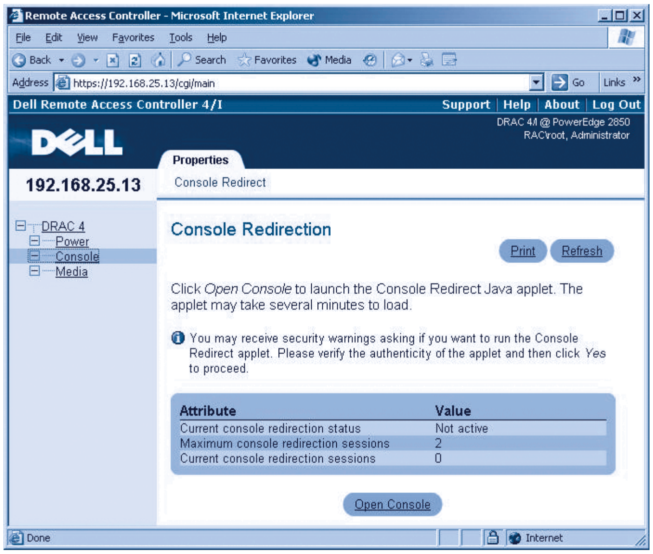
1. Reboot the system. During the power-on self-test (POST) press Ctrl + D within five seconds of the time the DRAC 4 banner is displayed.
2. When the Setup screen appears, make changes in the network settings—such as enabling or disabling Dynamic Host Configuration Protocol (DHCP)—and edit the static IP address, subnet mask, gateway, Domain Name System (DNS), and Ethernet configuration options. Be sure that the DRAC 4 network interface card (NIC) setting is enabled.
3. Press R to save the changes and reboot the DRAC 4.

Using option ROM requires administrators to perform the configuration task locally. Dell provides a mechanism known as console redirection to remotely configure the DRAC 4 as explained in the next section.

Setting up network connectivity using BIOS serial console redirection

Many organizations use special-purpose serial concentrators to provide remote access to their server systems. These concentrators generally take input from a serial port and make this port available over the network using a Telnet connection, a Web interface, or a secure protocol such as Secure Shell (SSH). The Dell PowerEdge server BIOS has a feature that allows administrators to connect the serial port to one of these concentrator devices and to access BIOS functionality through the serial port. Administrators can use this feature to remotely set up network connectivity for the DRAC 4 by performing the following steps:

1. Connect the serial port of the server to the serial concentrator and power up the system.
2. Access the serial concentrator according to the manufacturer's directions on how to access that particular model.
3. Configure the DRAC 4 network settings as described in the previous section.


Remote Access Controller - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address
Go Links
Dell Remote Access Controller 4/I
Support | Help | About | Log Out
DELL
Properties
Console Redirect
DRAC 4
Power
Console
Media
Print
Refresh
Open Console
Done
Internet

https://192.168.25.13/cgi/main

DRAC 4/I @ PowerEdge 2850
RAC\root, Administrator

192.168.25.13

Console Redirection

Click *Open Console* to launch the Console Redirect Java applet. The applet may take several minutes to load.

You may receive security warnings asking if you want to run the Console Redirect applet. Please verify the authenticity of the applet and then click *Yes* to proceed.

Attribute	Value
Current console redirection status	Not active
Maximum console redirection sessions	2
Current console redirection sessions	0

Figure 1. Console Redirection page of the DRAC 4 Web-based interface

Configuration of the DRAC 4

After the network settings have been configured on the DRAC 4, administrators must configure DRAC 4 console redirection and virtual media support. Only after these steps have been accomplished can administrators launch DSA to complete the remote OS installation. These steps must be performed each time administrators want to connect to the server.

Configuring the DRAC 4 console redirection feature

The DRAC 4 console redirection feature allows administrators to manage a system remotely in either graphical or text mode. Using the console redirection feature, the remote system can be operated by using the keyboard, video, and mouse on the local management station to control the corresponding devices on a remote system.

To use the console redirection feature, all browsers must have the supported Java Virtual Machine (JVM) plug-in (version 1.4.2 or later) installed. If using a Microsoft Windows OS–based management station, administrators should clear and disable the Java cache from the Java plug-in control panel.

To open a console redirection session, administrators should perform the following steps:

1. Connect and log in to the DRAC 4 Web-based interface in a Web browser on the management station by typing the IP address of the DRAC 4 in the browser address bar.
2. After logging in to the interface, click "Console" in the left pane to open the Console Redirection page (see Figure 1). The number of available console redirection sessions for the remote system will appear on the Console Redirection page.

[2] For additional ways to configure DRAC 4 network settings, refer to the *Dell Remote Access Controller 4 User's Guide* at support.dell.com/support/edocs/software/smdrac3/drac4/en/ug/racugc2.htm.

The DRAC 4 supports a maximum of two concurrent console redirection sessions.

3. Open a new console by clicking “Open Console” at the bottom of the Console Redirection page.
4. Click “Yes” to accept the remote access controller (RAC) Web security certificate.

Configuring the DRAC 4 virtual media feature

The DRAC 4 virtual media feature allows CD or floppy disk drives to be used on the management station as if they were connected directly to the managed server. Using this feature, administrators can remotely install new operating systems, install applications, or update drivers from virtual CD or virtual floppy disk drives.

On a Microsoft Windows management station, Microsoft Internet Explorer must be used as the browser for the virtual media feature. Browser security must be set to the medium or low setting to enable the browser to download and install signed Microsoft ActiveX® controls. Administrator rights are required to install the ActiveX control plug-in for the browser. On Red Hat Enterprise Linux management stations, write permission is required to install the plug-in for the Mozilla and Netscape browsers.

Figure 2 shows the Virtual Media page of the DRAC 4 Web-based interface. The following steps can be used to set up the virtual media feature from the management station:

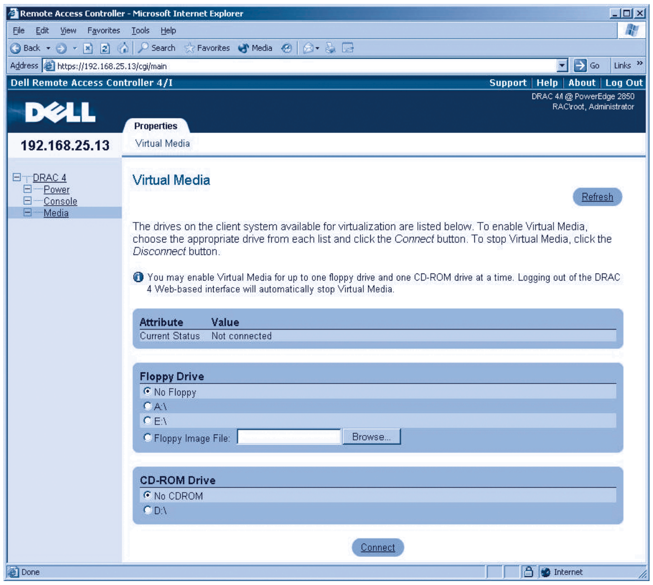
1. Connect and log in to the DRAC 4 from a Web browser on the management station. To use the virtual media feature to connect or disconnect virtual media, log in to the DRAC 4 as a user with Access Virtual Media permission.
2. After logging in to the interface, click “Media” in the left pane to open the Virtual Media page (see Figure 2). A prompt to install the virtual media plug-in will display if the virtual media feature is being run for the first time at the management station.
3. View the current status in the Attribute field. If the associated Value field displays “Connected,” deselect the relevant radio buttons to disconnect the virtual floppy disk image, virtual floppy disk drive, or virtual CD drive before reconnecting to the desired drives.
4. Select a physical CD or floppy disk drive to virtualize, and click “Connect” at the bottom of the Virtual Media page. The virtual floppy disk can be connected to a maximum of one available 1.44-inch virtual floppy disk image, 1.44-inch virtual floppy disk drive, or Universal Serial Bus (USB) key. The virtual CD can be connected to exactly one local CD drive.
5. After clicking “Connect,” access any selected drives on the managed server’s console as though they were physical drives on that server.

Remote deployment of an operating system

Once the virtual media and console redirection features are properly configured for the DRAC 4 on an eighth-generation Dell PowerEdge server, administrators can launch and optimize the OS installation using the Dell OpenManage Server Assistant CD. DSA helps to streamline OS installation and reduce the attended time needed to install supported operating systems on a server. Each Dell PowerEdge server is sold with a copy of the Dell OpenManage Server Assistant CD that will operate on that server. If attempting to use DSA on additional PowerEdge servers, carefully check the CD label for the list of supported servers.

The following steps start the installation of an OS through the DRAC 4 using the Dell OpenManage Server Assistant CD. *Note:* The list of operating systems supported for installation on an individual server is based on the model of the server. Operating systems that can be installed using virtual media include Microsoft Windows 2000 Server with Service Pack 4 (SP4), Microsoft Windows Server™ 2003, Red Hat Enterprise Linux 2.1, and Red Hat Enterprise Linux 3. Installation of Novell NetWare operating systems through virtual media is not supported on eighth-generation Dell PowerEdge servers.

1. Follow the steps in the “Configuring the DRAC 4 virtual media feature” section in this article to connect the local physical CD drive to a virtual CD drive.
2. Insert a supported version of the Dell OpenManage Server Assistant CD into the virtualized CD drive at the management station.
3. As described in the “Configuring the DRAC 4 console redirection feature” section in this article, open a console redirection session to monitor and perform the OS installation process.


Remote Access Controller - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address
Go Links
Dell Remote Access Controller 4/I
Support | Help | About | Log Out
DELL
Properties
DRAC 4
Power
Console
Media
Refresh
Connect
Done
Internet

https://192.168.25.13/cgi/main

DRAC 4/I @ PowerEdge 2850
RAC\root, Administrator

192.168.25.13

Virtual Media

Virtual Media

The drives on the client system available for virtualization are listed below. To enable Virtual Media, choose the appropriate drive from each list and click the *Connect* button. To stop Virtual Media, click the *Disconnect* button.

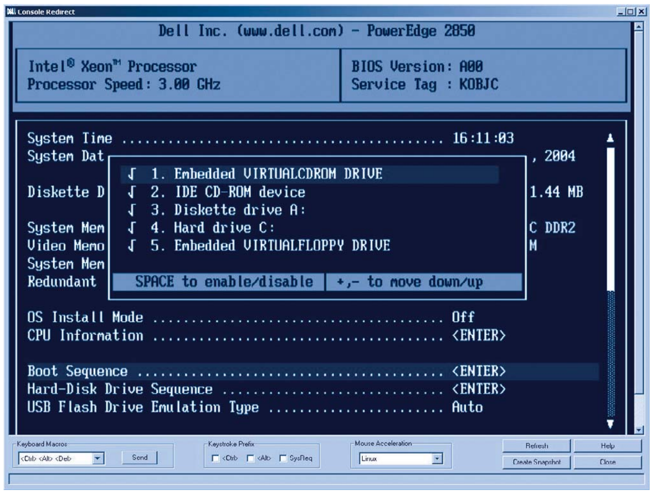
You may enable Virtual Media for up to one floppy drive and one CD-ROM drive at a time. Logging out of the DRAC 4 Web-based interface will automatically stop Virtual Media.

Attribute	Value
Current Status	Not connected

Floppy Drive
● No Floppy
○ A:\
○ E:\
○ Floppy Image File: Browse...

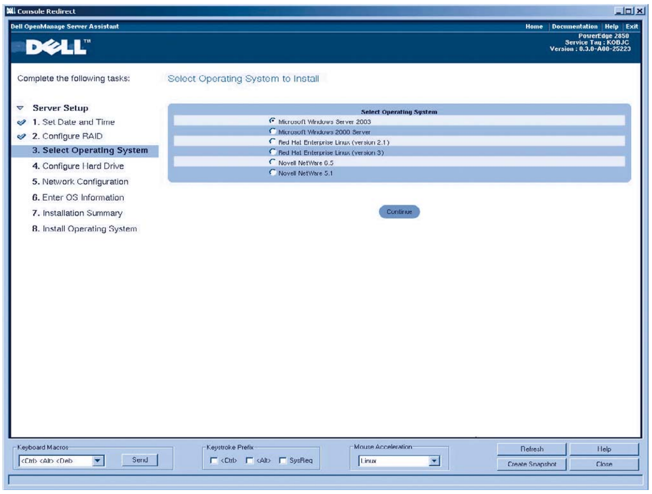
CD-ROM Drive
● No CDROM
○ D:\

Figure 2. Virtual Media page of the DRAC 4 Web-based interface



```
Console Redirect

                 Dell Inc. (www.dell.com) - PowerEdge 2850

 Intel® Xeon™ Processor             BIOS Version: A00
 Processor Speed: 3.00 GHz          Service Tag : KOBJC

 System Time ........................................ 16:11:03
 System Dat  ┌──────────────────────────────────────────┐ , 2004
             │ √ 1. Embedded VIRTUALCDROM DRIVE         │
 Diskette D  │ √ 2. IDE CD-ROM device                   │ 1.44 MB
             │ √ 3. Diskette drive A:                   │
 System Mem  │ √ 4. Hard drive C:                       │ C DDR2
 Video Memo  │ √ 5. Embedded VIRTUALFLOPPY DRIVE        │ M
 System Mem  │                                          │
 Redundant   │ SPACE to enable/disable │ +,- to move down/up │

 OS Install Mode ...................................... Off
 CPU Information ...................................... <ENTER>

 Boot Sequence ........................................ <ENTER>
 Hard-Disk Drive Sequence ............................. <ENTER>
 USB Flash Drive Emulation Type ....................... Auto
```

Figure 3. Boot Sequence menu in the BIOS settings



Console Redirect

Dell OpenManage Server Assistant — Home | Documentation | Help | Exit

PowerEdge 2850
Service Tag : KOBJC
Version : 0.3.0-A00-25223

Complete the following tasks:

Server Setup
1. Set Date and Time ✓
2. Configure RAID ✓
3. Select Operating System
4. Configure Hard Drive
5. Network Configuration
6. Enter OS Information
7. Installation Summary
8. Install Operating System

Select Operating System to Install

Select Operating System
● Microsoft Windows Server 2003
○ Microsoft Windows 2000 Server
○ Red Hat Enterprise Linux (version 2.1)
○ Red Hat Enterprise Linux (version 3)
○ Novell NetWare 6.5
○ Novell NetWare 5.1

Figure 4. DSA console redirection page for selecting the OS to install on the managed server

4. The DRAC 4 allows administrators to remotely perform several power management actions on the managed server. To start the installation process, after the console redirection session has been started, navigate back to the Server Control page of the DRAC 4 Web-based interface to reboot the system.
5. At the console redirection window, press F2 during the POST of the managed server to enter the BIOS setup.
6. Navigate down to the Boot Sequence menu. In this menu, make sure the virtual CD drive is enabled and edit the virtual drive to be the first drive in the boot sequence using the keys indicated on the screen. Most servers use the + key and the – key to move menu entries up and down (see Figure 3).
7. Save the changes and exit. The server will reboot upon exit.
8. Follow the on-screen instructions to complete the OS installation with DSA. The DSA application will request the CD for the OS being installed (see Figure 4). DSA copies this CD to the hard drive. After the copy process completes, the managed server reboots into the OS unattended installation mode.

> Dell OpenManage provides open, flexible systems management tools that can integrate and help standardize and automate server management processes.

Remote OS deployment to enhance administrator productivity

Starting with eighth-generation Dell servers, such as the PowerEdge 1850, PowerEdge 2800, and PowerEdge 2850, administrators can take advantage of the DRAC 4 and Dell OpenManage Server Assistant to deploy operating systems remotely. When deploying an OS remotely, administrators can use DSA features as effectively as they would on a local system. This approach can prove extremely useful for provisioning eighth-generation Dell PowerEdge servers in distributed data centers or other remote locations. Remote OS deployment can save time by enabling administrators to deploy operating systems without traveling to remote locations, and can also help distributed enterprises keep IT staffing requirements to a minimum.

Michael E. Brown is a software developer in the Dell Enterprise Software Development Group. He is the technical lead for Dell OpenManage Server Assistant. He attended Southwestern Michigan College and is a Red Hat Certified Engineer® (RHCE®), a Microsoft Certified Systems Engineer (MCSE), and a Certified Novell Administrator (CNA).

Manoj Gujarathi is a systems engineer in the Dell OpenManage Software Development Group. He is currently working on Dell OpenManage Server Assistant and has worked on various Dell OpenManage applications in the past. Manoj has an M.S. in Engineering from Washington State University and an M.S. in Computer Science from Texas Tech University.

Gong Wang is a software engineer in the Dell Enterprise Software Development Group. Before joining Dell, he worked as a research scientist at the Georgia Institute of Technology (Georgia Tech) and as an instructor at Wuhan University in China. Gong has an M.S. in Human-Computer Interaction and an M.S. in Experimental Psychology from Georgia Tech.

FOR MORE INFORMATION

Serial console setup:
www1.us.dell.com/content/topics/global.aspx/power/en/ps1q03_stanton?c=us&l=en&s=corp

Migrating Enterprise Databases

from Sun Servers to the Dell PowerEdge 2850 Running Microsoft Windows Server 2003

The latest generation of Dell™ servers continues the trend of increasing performance and value by using Intel® Xeon™ processors with Extended Memory 64 Technology and the latest server hardware architecture features. To demonstrate the value and benefits of migrating from Sun servers, Dell engineers moved a large enterprise database application from a Sun server to a Dell server. The Dell PowerEdge 2850 with dual Intel Xeon processors at 3.4 GHz was able to handle 77 percent more orders per minute than the Sun Fire V440 with four UltraSPARC IIIi processors at 1.28 GHz, and was 49 percent less expensive than the Sun system.

BY TODD MUIRHEAD AND DAVE JAFFE, PH.D.

Previous studies conducted in Dell labs have shown the advantages of migrating large databases from proprietary UNIX/RISC-based platforms such as the Sun Fire V440 to industry-standard servers such as the Dell PowerEdge 6650, a server configured with four Intel Xeon processors MP.[1] In the study discussed in this article, Dell engineers compared a Sun Fire V440 configured with four UltraSPARC IIIi processors at 1.28 GHz to the PowerEdge 2850 server, one of Dell's eighth-generation servers based on Intel Xeon processors with Extended Memory 64 Technology (EM64T). EM64T enables enterprises to continue using 32-bit versions of critical software—such as database management systems—and upgrade to 64-bit versions as they become available in the near future. With two Intel Xeon processors running at up to 3.6 GHz, an 800 MHz frontside bus (FSB), and up to 12 GB of double data rate 2 (DDR2) memory at 400 MHz, the PowerEdge 2850 offers excellent server performance in a 2U, or 3.5-inch, server.

To demonstrate the ease and potential benefits of migration from Sun Solaris–based servers, in November 2004 Dell engineers quickly and easily moved a large (100 GB) database running on a leading enterprise database manager from a four-processor Sun Fire V440 server to the dual-processor Dell PowerEdge 2850 server. No data was lost during the migration, and the application (an online DVD store) ran more than twice as fast on the PowerEdge 2850 server with two Intel Xeon processors at 3.4 GHz than on the Sun Fire V440 server with four UltraSPARC IIIi processors at 1.28 GHz. To ensure a fair test comparison, the test team equipped the PowerEdge 2850 server with

[1] For the previous migration studies, see "Migrating Databases from Sun Systems to Dell Servers Running Microsoft Windows Server 2003" by Todd Muirhead; Dave Jaffe, Ph.D.; and Kerstin Ackerman in *Dell Power Solutions,* October 2004; and "Migrating Databases from Sun Systems to Dell Servers Running Red Hat Enterprise Linux AS 3" by Todd Muirhead and Dave Jaffe, Ph.D., in *Dell Power Solutions,* October 2004.

	Sun Fire V440	Dell PowerEdge 2850
Operating system	Solaris 9 12/03	Microsoft Windows Server 2003, Enterprise Edition
CPU	Four UltraSPARC IIIi processors at 1.28 GHz with 1 MB of L2 cache	Two Intel Xeon processors at 3.4 GHz with 1 MB of L2 cache
Memory	8 GB	8 GB (four 2 GB dual in-line memory modules, or DIMMs)
Internal disks	Four 73 GB 10,000 rpm Ultra320 SCSI	Two 73 GB 10,000 rpm Ultra320 SCSI
Network interface cards (NICs)	Two 10/100/1000 Mbps (internal)	Two 10/100/1000* Mbps (internal)
Disk controller	On-board SCSI	PowerEdge Expandable RAID Controller, embedded internal (PERC 4/ei)
Fibre Channel HBAs	Two QLogic QLA2340	Two QLogic QLA2340
Remote management	Serial management card	Dell Remote Access Controller 4 (DRAC 4) without modem
Service	Three-year Gold Support with 24/7 on-site	Three-year Gold Support with 24/7 on-site**
Hardware price (without HBAs)	$27,339	$14,817
Price of database licenses ($15,000 per CPU)	$60,000	$30,000
Total price	$87,339***	$44,817***

*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.

**Service may be provided by a third party. Technician will be dispatched if necessary following phone-based troubleshooting. Subject to parts availability, geographical restrictions, and terms of service contract. Service timing dependent upon time of day call placed to Dell. United States only.

***Source: U.S. prices for the Sun Fire V440 server and Dell PowerEdge 2850 server are cited from the Sun and Dell online stores, respectively (www.sun.com and www.dell.com), as of November 29, 2004. Prices include server hardware and software as well as the OS.

Figure 1. Configurations of the Sun Fire V440 server and the Dell PowerEdge 2850 server

8 GB of memory and two Intel Xeon processors at 3.4 GHz, slightly slower than the fastest processors available for the system. Similarly, the team configured the Sun Fire V440 server with 8 GB of memory and four UltraSPARC IIIi processors at 1.28 GHz, which is classified on the Sun Web site as the "medium" Sun Fire V440 server configuration, with the processor speed being slightly slower than the fastest UltraSPARC IIIi processor available at the time, 1.593 GHz.[2]

The following sections describe the Sun and Dell hardware used in these tests, the database setup, how the database was migrated from the Sun server to the Dell server, the DVD store application, details of how the tests were performed, and test results.

Configuring the hardware

In this study, a large (100 GB) database was migrated from a four-processor Sun Fire V440 to a two-processor PowerEdge 2850. The processors used for both systems were not the fastest speed offered by each vendor at the time of the tests.[3] See Figure 1 for further configuration details.

Figure 1 shows list prices of the two systems. The list price of the database manager license, based on processor count, is included in the calculation of the total price. The operating system (OS) cost is included in the server hardware price. Pricing of the configurations did not include the storage area network (SAN) hardware or software. The Dell PowerEdge 2850 server ran Microsoft Windows Server 2003, Enterprise Edition, and the Sun Fire V440 server ran Solaris 9 12/03.

The Sun Fire V440 server uses a 64-bit architecture that allows a flat addressing model to directly address the entire 8 GB of memory installed in the server. The 32-bit architecture of the PowerEdge 2850 server required some additional parameters to enable access to the full 8 GB of memory configured in the system. The `/3GB` and `/PAE` parameters were added to the Windows® boot.ini file to enable the additional memory support required for the Windows OS. These additional parameters will not be needed when the 64-bit Windows Server 2003 OS and 64-bit database software are available.

Storage for both the Dell and Sun servers was provided by a SAN-attached Dell/EMC CX700 Fibre Channel storage array. Each server was attached to the SAN via two QLogic host bus adapters (HBAs). Each server also was assigned to a set of logical storage units (LUNs) that used the same number and type of disk drives. Dell engineers consulted the EMC compatibility matrix,[4] which includes both the Dell PowerEdge 2850 server and Sun Fire V440 server, to ensure that the proper versions of drivers and firmware were installed on both the Dell and Sun systems. Both the 8.2.3.21 version of the QLogic QLA2340 driver for Windows and EMC® PowerPath® 3.0.6 for Windows were installed on the Dell system. PowerPath 4.2.0 for Solaris and QLogic driver version 4.13 for Solaris were installed on the Sun Fire V440 server. PowerPath provides load balancing and failover capabilities for the dual HBAs that were present in all systems. The storage components used for both servers in the test are shown in Figure 2.

Setting up the database and performing the migration

The same leading enterprise database server was installed and configured on the Sun Fire V440 server and the Dell PowerEdge 2850 server. Installation and configuration of the database software were performed

[2] The Sun Fire V440 server offered two processor speeds at the time of testing: 1.28 GHz and 1.593 GHz. For more information about Sun Fire V440 configurations, visit store.sun.com/CMTemplate/CEServlet?process=SunStore&cmdViewProduct_CP&catid=104994.

[3] The Sun Fire V440 server offered two processor speeds at the time of testing: 1.28 GHz and 1.593 GHz. The PowerEdge 2850 server offered five processor speeds: the fastest was 3.6 GHz and the slowest was 2.8 GHz. Server testing performance varies with processor speed.

[4] The EMC compatibility matrix is available at www.emc.com/interoperability/index.jsp.

Disk enclosures	Three Dell/EMC 2 GB Disk Array Enclosures (DAE2)
Disks	Forty 73 GB 10,000 rpm
LUNs	Three 10-disk RAID-10 LUNs for data Two 2-disk RAID-1 LUNs for logs One 5-disk RAID-0 for temporary data staging (to load the database) One hot-spare disk
Software	EMC Navisphere® Manager EMC Access Logix™ EMC PowerPath

Figure 2. Dell/EMC storage configuration

according to the database provider's installation guide for the respective server platforms. The enterprise database server provided the same Java-based tool for installation on both Solaris and Windows, allowing the test team to select exactly the same options.

Dell engineers used a database creation assistant wizard to create the initial database instance on the Sun Fire V440 server. They then used scripts to finish the creation and loading of the test database. The database consisted of approximately 100 GB of data, indexes, and stored procedures.

The 100 GB database was moved from the Sun Fire V440 server to the Dell PowerEdge 2850 server using migration tools included with the database server. Using a few commands, the test team converted the Sun data files and then transported those data files from the 64-bit Solaris-based server to the 32-bit Windows-based server. The basic steps were to create copies of the data files on the Sun server, convert the copies into a Windows-readable format, move the newly created files to the Dell server, and connect the transported data files into the database instance running on the Dell server.

Database initialization parameters were obtained by allowing the database to run in an auto-tuning mode on the Sun Fire V440 server. The test team set a target of 7 GB of RAM for the database to use. After allowing the auto-tuning to occur on the Sun server, the team applied the tuned settings to the target Windows system as closely as possible. Some slight differences in the memory initialization parameters occurred because of the different memory models on the two platforms. Specifically, on the PowerEdge server, the database initialization parameters were `USE_INDIRECT_DATA_BUFFERS` and `DB_BLOCK_BUFFERS` instead of `DB_CACHE_SIZE` to specify the amount of memory to be used. A slightly larger amount of memory was available to the database on the Sun Fire V440 server because a small amount of memory on the PowerEdge 2850 server must be used to enable the very large memory support on the 32-bit platform. The Sun server's System Global Area (SGA) was 6.84 GB, compared to 6.78 GB on the Dell system. The nondefault database initialization parameters used for this study are listed in Figure 3.

Both systems used the database-provided storage manager for the database data files and log files. This storage manager reduces the overhead of a file system by allowing the database to access and manage the storage more directly.

The database tablespaces were set up exactly the same on each of the systems (see Figure 4). The three 10-disk RAID-10 LUNs were put into one disk group in the storage manager, which was used for the data, index, undo, and temporary tablespaces. Each of the two RAID-1 LUNs was assigned to its own disk group and used for redo logs.

The application

The test described in this article used the same version of the online DVD store database application (DVD Store 2, or DS2)[5] employed in a previous Red Hat® Linux® migration test. This version of the database application includes advanced database features such as transactions, triggers, and referential integrity constraints. In addition, the database application includes functionality typical of some

Parameter	Sun Fire V440	Dell PowerEdge 2850
DB_BLOCK_BUFFERS	n/a	668281
DB_BLOCK_SIZE	8192	8192
DB_CACHE_SIZE	5335154688	n/a
JAVA_POOL_SIZE	16777216	16777216
LARGE_POOL_SIZE	16777216	16777216
OPEN_CURSORS	300	300
PROCESSES	150	150
PGA_AGGREGATE_TARGET	2136997888	2136997888
SHARED_POOL_SIZE	1962934272	1762934272
SORT_AREA_SIZE	65536	65536
UNDO_RETENTION	300	300
USE_INDIRECT_DATA_BUFFERS	n/a	TRUE

Figure 3. Database initialization parameters

Tablespace	Contains	Space used/available
CUSTTBS	Customers table	34 GB/38 GB
INDXTBS	Indexes	30 GB/32 GB
ORDERTBS	Orders and orderlines tables	20 GB/24 GB
DS_MISC	Product and categories tables	0.05 GB/1 GB
UNDOTBS	Undo tablespace	1 GB/3 GB
TEMP	Temporary table	12 GB/12 GB

Figure 4. Database tablespaces used in the test environment

[5] For more information about DS2, see "Migrating Databases from Sun Systems to Dell Servers Running Red Hat Enterprise Linux AS 3" by Todd Muirhead and Dave Jaffe, Ph.D., in *Dell Power Solutions,* October 2004.

Table	Columns	Number of rows
Customers	CUSTOMERID, FIRSTNAME, LASTNAME, ADDRESS1, ADDRESS2, CITY, STATE, ZIP, COUNTRY, REGION, EMAIL, PHONE, CREDITCARD, CREDITCARDEXPIRATION, USERNAME, PASSWORD, AGE, INCOME, GENDER, PROD_ID_IDX, PROD_ID1, PROD_ID2, PROD_ID3, PROD_ID4, PROD_ID5, PROD_ID6, PROD_ID7, PROD_ID8, PROD_ID9, PROD_ID10	200 million
Orders	ORDERID, ORDERDATE, CUSTOMERID, NETAMOUNT, TAX, TOTALAMOUNT	120 million
Orderlines	ORDERLINEID, ORDERID, PROD_ID, QUANTITY, ORDERDATE	600 million
Products	PROD_ID, CATEGORY, TITLE, ACTOR, PRICE, QUAN_IN_STOCK, SPECIAL, COMMON_PROD_ID1, COMMON_RATING1, COMMON_PROD_ID2, COMMON_RATING2, COMMON_PROD_ID3, COMMON_RATING3, SALES	1 million
Reorder	PROD_ID, DATE_LOW, QUAN_LOW, DATE_REORDERED, QUAN_REORDERED, DATE_EXPECTED	Variable
Categories	CATEGORY, CATEGORYNAME	16

Figure 5. Database schema for the online DVD store

online stores, such as reporting previous purchases to the user and recommending titles enjoyed by others.

The database, which is about 100 GB and represents an online DVD store with 1 million DVD titles, was driven by a ProC-language program simulating users logging in to the online store; browsing for DVDs by title, author, and category; and then submitting orders. The driver program measured the number of orders per minute (opm) that the database could handle as well as the total response time as seen by the simulated end users.

The database schema

The DVD store comprised five main tables and one additional table (see Figure 5). The Customers table was prepopulated with 200 million customers: 100 million U.S. customers and 100 million customers from the rest of the world. The Orders table was prepopulated with 10 million orders per month for a full year. The Orderlines table was prepopulated with an average of five items per order. The Products table contained 1 million DVD titles. In addition, the Categories table listed the 16 DVD categories.

When the QUAN_IN_STOCK value for each product in the Products table fell below a specified value, the database application was triggered to write information about the product to the Reorder table. A separate process (not modeled) monitored this table to initiate reordering of needed titles.

The stored procedures

The DVD store database was managed using six stored procedures. The first two were used during the login phase. If the customer was a returning customer, Login was used to retrieve the customer's information, in particular the CUSTOMERID. If the customer was a new customer, New_customer was used to create a new row in the Customers table with the customer's data. Following the login phase, the customer might search for a DVD by category, actor, or title using Browse_by_category, Browse_by_actor, and Browse_by_title, respectively. Finally, after the customer had made a selection, the Purchase stored procedure was called to complete the transaction. Visit *Dell Power Solutions* online at www.dell.com/powersolutions to see the DS2 stored procedures build scripts.

The stored procedures of the application include features that model today's online stores. During login, for example, the user's previous order (up to 10 titles) is reported, along with titles recommended by other customers who liked those titles. Browse_by_category returns those titles in the specified categories that are currently on sale. The Purchase stored procedure checks the QUAN_IN_STOCK field from the Products table to determine whether a title is available. This is done using a database transaction; therefore, if insufficient quantity exists to fill the order, the QUAN_IN_STOCK data is not updated, nor is a new record written to the Orders table.

The OLTP driver application

A multithreaded driver program was written to model an online transaction processing (OLTP) or order-entry workload. Each thread of the OLTP driver application connected to the database and made a series of stored procedure calls that simulated customers logging in, browsing, and purchasing. Because no simulated customer

Parameter	Description	Value(s) used in test
n_threads	Number of simultaneous connections to the database	See Figure 7
warmup_time	Warmup time before statistics are kept	1 minute
run_time	Runtime during which statistics are kept	Varied
pct_returning	Percent of customers who are returning	80%
pct_new	Percent of customers who are new	20%
n_browse_category	Number of searches based on category	Range: 1–3 Average: 2
n_browse_actor	Number of searches based on actor	Range: 1–3 Average: 2
n_browse_title	Number of searches based on title	Range: 1–3 Average: 2
n_line_items	Number of items purchased	Range: 1–9 Average: 5
net_amount	Total amount of purchase	Range: $0.01–$400.00 Average: $200.00

Figure 6. OLTP driver application parameters

System	Simultaneous database connections	Orders per minute (larger is better)	Average response time (seconds)	CPU utilization	Dell PowerEdge 2850 performance advantage over Sun Fire V440	Total hardware and software price	Price/performance ($/opm—lower is better)	Dell price/ performance advantage
Four-processor Sun Fire V440 (processors: 1.28 GHz with 1 MB of L2 cache)	8	8,077	0.058	91%	n/a	$87,339	$10.81	n/a
Dual-processor Dell PowerEdge 2850 (processors: 3.4 GHz with 1 MB of L2 cache	11	14,274	0.045	90%	1.77×	$44,817	$3.14	244%

Figure 7. Sun-to-Dell migration results

think times or key times were factored in, the database connections remained full—thereby simulating what happens in a real multi-tiered application in which a few connections are pooled and shared among the Web servers that may be handling thousands of simultaneous customers. In this way, Dell engineers achieved a realistic simulation of database activity without needing to model thousands of customers.

The Dell PowerEdge 2850 configured with two Intel Xeon processors at 3.4 GHz, while priced 49 percent less than the Sun Fire V440, could handle 77 percent more orders.

Each thread of the OLTP driver application modeled a series of customers going through the entire sequence of logging in, browsing the catalog several ways, and finally purchasing the selected items. Each sequence completed by a customer counted as a single order. The OLTP driver application measured order rates and the average response time to complete each order. Several tunable parameters were used to control the application. These parameters are shown in Figure 6.

Examining the test results

The DVD store database running on the Sun Fire V440 server was tested using the OLTP driver application described in the preceding section. The database was then moved to the Dell PowerEdge 2850 server, where it was tested with the same OLTP driver application. The tests measured how many orders per minute each database server could handle, while keeping each server's CPU utilization around 90 percent—a typical system target to allow additional capacity for order spikes. CPU utilization was measured using the vmstat program for the Solaris OS and the Windows Performance Monitor program for the Windows Server 2003 OS.

As shown in Figure 7, the Dell PowerEdge 2850 configured with two Intel Xeon processors at 3.4 GHz, while priced 49 percent less than the Sun Fire V440, could handle 77 percent more orders. In terms of price/performance, the Sun Fire V440 server cost $10.81 per opm whereas the Dell Power Edge 2850 server cost $3.14 per opm. In other words, the Sun Fire V440 cost more than three times, or 244 percent, more than the Power Edge 2850 server per unit of work that it could handle.

Benefiting from industry-standard servers

By easily and quickly migrating a large database built with a leading enterprise database server from a Sun Fire V440 server configured with 8 GB of memory and four UltraSPARC IIIi processors at 1.28 GHz to a dual-processor Dell PowerEdge 2850 server configured with 8 GB of memory and two Intel Xeon processors at 3.4 GHz, Dell engineers demonstrated that the PowerEdge 2850 server was 77 percent faster than the Sun Fire V440 server. Based on prices cited from the Sun and Dell online stores as of November 29, 2004, the PowerEdge 2850 was also 49 percent less expensive than the Sun Fire V440, a price/performance advantage of 244 percent. While the Sun UltraSPARC architecture currently supports 64-bit operating systems and applications, the Intel Xeon processors with EM64T technology shipping in eighth-generation Dell servers such as the PowerEdge 2850 server will support 64-bit Windows Server 2003 and 64-bit applications when that software ships in the coming year, while currently supporting today's 32-bit software.

Todd Muirhead is an engineering consultant on the Dell Technology Showcase team. He specializes in SANs and database systems. Todd has a B.A. in Computer Science from the University of North Texas and is Microsoft Certified Systems Engineer + Internet (MCSE+I) certified.

Dave Jaffe, Ph.D., is a senior consultant on the Dell Technology Showcase team who specializes in cross-platform solutions. Previously, he worked in the Dell Server Performance Lab, where he led the team responsible for Transaction Processing Performance Council (TPC) benchmarks. Before working at Dell, Dave spent 14 years at IBM in semiconductor processing, modeling, and testing, and in server and workstation performance. He has a Ph.D. in Chemistry from the University of California, San Diego, and a B.S. in Chemistry from Yale University.

Managing
Dell PowerEdge Server Alerts
Using Dell OpenManage Server Administrator

Eighth-generation Dell™ PowerEdge™ servers are enabled by Dell OpenManage™ systems management tools to provide a set of alerting mechanisms that proactively notify system administrators of abnormalities before failures occur. This article introduces and explains the three different alerting mechanisms provided by Dell OpenManage Server Administrator.

BY HAIHONG ZHUO, MICHAEL O'HARA, AND JIANWEN YIN, PH.D.

Enterprise servers that run mission-critical applications should be designed to run error-free to help maximize performance and minimize downtime. It is critical that system administrators be notified of any system abnormalities before hardware failures occur if feasible—otherwise, as soon as possible. Eighth-generation Dell servers such as the PowerEdge 1850, PowerEdge 2800, and PowerEdge 2850 help achieve this goal by providing a set of three different alerting mechanisms that can be viewed or configured through the Dell OpenManage systems management suite.

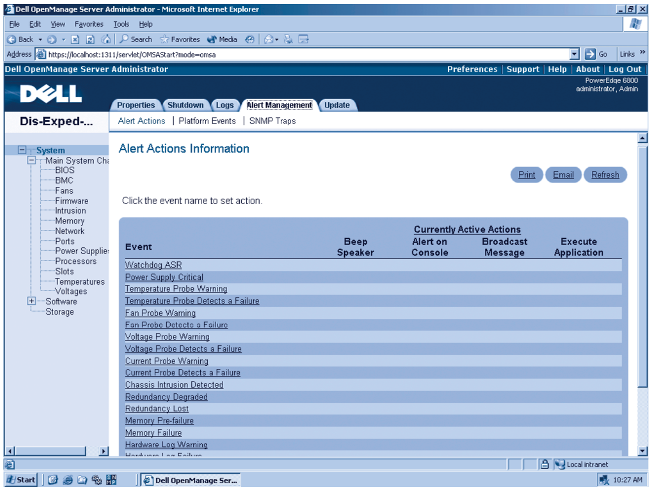
Dell PowerEdge servers managed through Dell OpenManage tools can be configured to generate and send Simple Network Management Protocol (SNMP) traps, trigger local alert actions (such as beeping the speakers on a system that needs attention), or generate and send Platform Event Traps (PETs) when they detect system error events or status changes. Configuration of these alerting mechanisms is done using Dell OpenManage Server Administrator (OMSA), a software tool that allows administrators to manage individual servers locally or remotely using a graphical user interface (GUI), or locally using the OMSA command-line interface (CLI).[1]

Local alert actions

Dell PowerEdge servers that are managed by OMSA can be configured to trigger certain local actions on the occurrence of specified events—for instance, when a system's temperature probe or voltage probe detects a warning or failure. OMSA alerts the administrator by beeping the speaker on the affected system, popping up an alert message on the system, invoking an application on the system, or broadcasting an alert message through a messenger service to other systems that are on the same network and have drives mapped to the affected system.

The OMSA Web-based GUI lists all system events for which local alert actions can be configured (see Figure 1). Administrators can also view a list of events

[1] For online documentation that provides detailed instructions on how to use Dell OpenManage Server Administrator, visit docs.us.dell.com/docs/software/svradmin/index.htm.

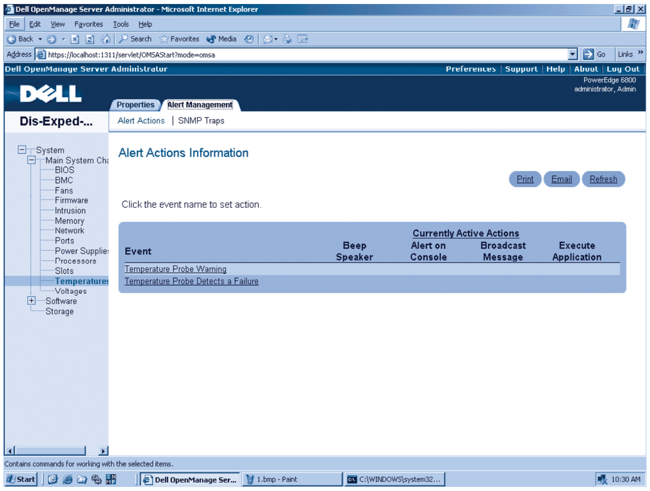

Dell OpenManage Server Administrator - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address https://localhost:1311/servlet/OMSAStart?mode=omsa
Go Links
Dell OpenManage Server Administrator
Preferences | Support | Help | About | Log Out
PowerEdge 6800
administrator, Admin
Properties Shutdown Logs Alert Management Update
Dis-Exped-...
Alert Actions | Platform Events | SNMP Traps
System
Main System Cha
BIOS
BMC
Fans
Firmware
Intrusion
Memory
Network
Ports
Power Supplie
Processors
Slots
Temperatures
Voltages
Software
Storage
Print Email Refresh
Local intranet
Start
Dell OpenManage Ser...
10:27 AM

Alert Actions Information

Click the event name to set action.

	Currently Active Actions			
Event	Beep Speaker	Alert on Console	Broadcast Message	Execute Application
Watchdog ASR				
Power Supply Critical				
Temperature Probe Warning				
Temperature Probe Detects a Failure				
Fan Probe Warning				
Fan Probe Detects a Failure				
Voltage Probe Warning				
Voltage Probe Detects a Failure				
Current Probe Warning				
Current Probe Detects a Failure				
Chassis Intrusion Detected				
Redundancy Degraded				
Redundancy Lost				
Memory Pre-failure				
Memory Failure				
Hardware Log Warning				
Hardware Log Failure				

Figure 1. Available system alert actions


Dell OpenManage Server Administrator - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address https://localhost:1311/servlet/OMSAStart?mode=omsa
Go Links
Dell OpenManage Server Administrator
Preferences | Support | Help | About | Log Out
PowerEdge 6800
administrator, Admin
Properties Alert Management
Dis-Exped-...
Alert Actions | SNMP Traps
System
Main System Cha
BIOS
BMC
Fans
Firmware
Intrusion
Memory
Network
Ports
Power Supplie
Processors
Slots
Temperatures
Voltages
Software
Storage
Print Email Refresh
Contains commands for working with the selected items.
Start
Dell OpenManage Ser...
1.bmp - Paint
C:\WINDOWS\system32...
10:30 AM

Alert Actions Information

Click the event name to set action.

	Currently Active Actions			
Event	Beep Speaker	Alert on Console	Broadcast Message	Execute Application
Temperature Probe Warning				
Temperature Probe Detects a Failure				

Figure 2. Alert actions for temperature sensors

related to a category of monitored components—for example, temperature sensors—by selecting a component category from the left navigation bar of the GUI (see Figure 2).

Local alert action configurations can also be viewed and managed using the OMSA CLI. Administrators can display a list of events for which local alert actions can be configured using the command `omreport system alertaction`. The CLI provides an online help system, which can be displayed by typing `-?` after any command.

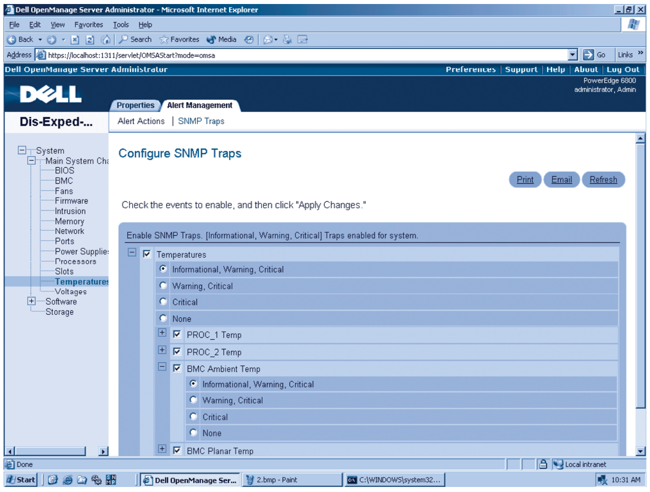
In today's enterprises, IT environments are becoming ever more complicated. As the number of computing systems that administrators must manage continues to grow, remote management becomes more popular. Locally managed alert actions may be insufficient in circumstances where administrators are not working in close physical proximity to the systems under management—for example, they cannot hear a system's speakers beep or see an alert message on a system's monitor. Remote alerting mechanisms are therefore an indispensable feature of a total systems management software package. OMSA provides two remote alerting methods to meet this need: SNMP traps and PETs.

Remote alerts using SNMP traps

When properly configured, Dell PowerEdge servers equipped with OMSA can generate and send SNMP traps for system events such as component failure or status change. While alerting using local alert actions is configured strictly by the category of the monitored components, alerting using SNMP traps in OMSA provides more flexibility and granularity.

Systems managed using OMSA can be configured to generate SNMP traps for a status change of the system, for a category of monitored components such as temperature sensors, or even for an individual component such as the temperature sensor for a specific CPU. Additionally, systems can be configured to generate SNMP traps for different severity levels of a system event: informational, warning, or critical.

By default, the generation of all SNMP traps is enabled. SNMP traps can be enabled or disabled by system, by category of component, by individual component, and by severity level. In the OMSA GUI, administrators can indicate the severity level for which they would like to enable alerts for the system. Administrators can select from a list of all categories of monitored components for a system, and then also indicate the severity level for which they would like to enable alerts for each category. Additionally, after enabling a category of components, the administrator can select components to enable from a list of individual components within that category if they do not want to enable every component in the category (see Figure 3).


Dell OpenManage Server Administrator - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address https://localhost:1311/servlet/OMSAStart?mode=omsa
Go Links
Dell OpenManage Server Administrator
Preferences | Support | Help | About | Log Out
PowerEdge 6800
administrator, Admin
Properties Alert Management
Dis-Exped-...
Alert Actions | SNMP Traps
System
Main System Cha
BIOS
BMC
Fans
Firmware
Intrusion
Memory
Network
Ports
Power Supplie
Processors
Slots
Temperatures
Voltages
Software
Storage
Print Email Refresh
Done
Local intranet
Start
Dell OpenManage Ser...
2.bmp - Paint
C:\WINDOWS\system32...
10:31 AM

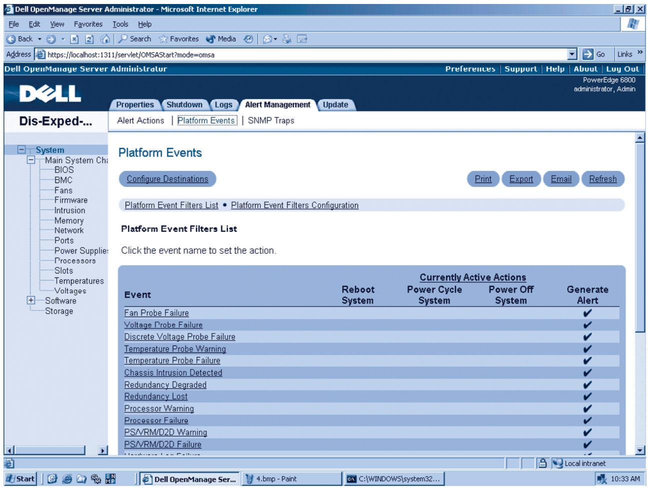
Configure SNMP Traps

Check the events to enable, and then click "Apply Changes."

Enable SNMP Traps. [Informational, Warning, Critical] Traps enabled for system.

- ☑ Temperatures
  - ● Informational, Warning, Critical
  - ○ Warning, Critical
  - ○ Critical
  - ○ None
  - ☑ PROC_1 Temp
  - ☑ PROC_2 Temp
  - ☑ BMC Ambient Temp
    - ● Informational, Warning, Critical
    - ○ Warning, Critical
    - ○ Critical
    - ○ None
  - ☑ BMC Planar Temp

Figure 3. SNMP traps enabled for temperature sensors


Dell OpenManage Server Administrator - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address https://localhost:1311/servlet/OMSAStart?mode=omsa
Go Links
Dell OpenManage Server Administrator
Preferences | Support | Help | About | Log Out
PowerEdge 6800
administrator, Admin
DELL
Properties
Shutdown
Logs
Alert Management
Update
Dis-Exped-...
Alert Actions | Platform Events | SNMP Traps
System
Main System Cha
BIOS
BMC
Fans
Firmware
Intrusion
Memory
Network
Ports
Power Supplie
Processors
Slots
Temperatures
Voltages
Software
Storage
Configure Destinations
Print
Export
Email
Refresh
Platform Event Filters List • Platform Event Filters Configuration
Local intranet
Start
Dell OpenManage Ser...
4.bmp - Paint
C:\WINDOWS\system32...
10:33 AM

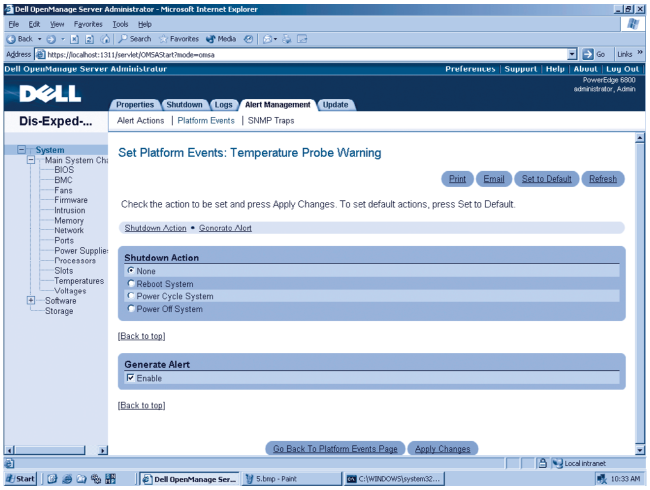
Platform Events

Platform Event Filters List

Click the event name to set the action.

Event	Currently Active Actions Reboot System	Power Cycle System	Power Off System	Generate Alert
Fan Probe Failure				✔
Voltage Probe Failure				✔
Discrete Voltage Probe Failure				✔
Temperature Probe Warning				✔
Temperature Probe Failure				✔
Chassis Intrusion Detected				✔
Redundancy Degraded				✔
Redundancy Lost				✔
Processor Warning				✔
Processor Failure				✔
PS/VRM/D2D Warning				✔
PS/VRM/D2D Failure				✔

Figure 4. Platform Event Filters



Set Platform Events: Temperature Probe Warning

Check the action to be set and press Apply Changes. To set default actions, press Set to Default.

Shutdown Action

● None
○ Reboot System
○ Power Cycle System
○ Power Off System

Generate Alert

☑ Enable

Figure 5. Actions for an Event Filter

SNMP trap destinations and the SNMP community to which the traps should be sent are configured in the server's operating system (OS). SNMP trap configuration can also be viewed and managed using the CLI. For example, executing the command `omconfig system events type=fans source=snmptraps severity=warning` tells the server not to generate SNMP traps for events with an informational severity level for fans.

Remote alerts using PETs

In addition to local alert actions and SNMP traps, eighth-generation Dell PowerEdge servers support generating and sending alerts at the hardware and firmware levels. These servers are equipped with a microcontroller called the baseboard management controller (BMC) that is compliant with the Intelligent Platform Management Interface (IPMI) 1.5 specification. The BMC provides the intelligence behind the autonomous monitoring and recovery features at the hardware and firmware levels.

When an event occurs on the platform, an event message is generated and logged in the BMC hardware system event log (SEL). The BMC checks whether the event meets the Platform Event Filter (PEF) criteria configured by the administrator through OMSA. If so, the BMC generates an SNMP PET and sends the PET to the destination or destinations designated by the administrator. Once established, this process does not require the OS or OMSA systems management software and allows alerts to be sent even when the system is powered down or unable to boot to the OS.

For platform events to be filtered properly and PETs to be generated and sent as desired, administrators must first perform certain configuration steps. OMSA provides an interface that allows administrators to configure when PET alerts should be generated and sent and where they should be sent.

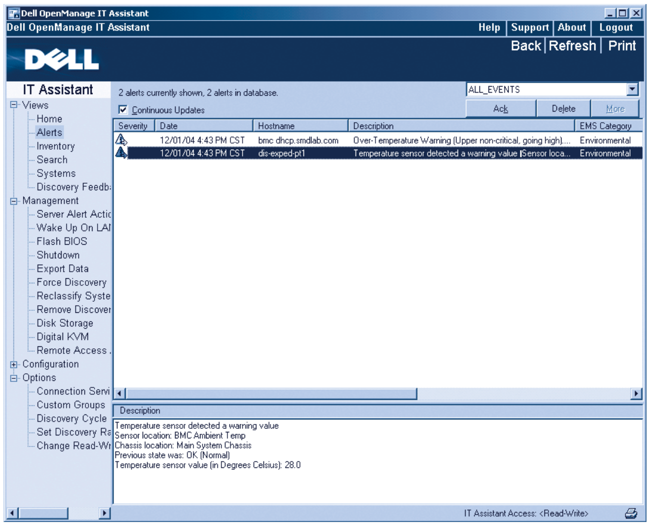
Administrators must first configure PEFs on the BMC to indicate which actions the BMC should take when a filter is triggered. Figure 4 shows a list of such event filters, and Figure 5 illustrates how actions for each event filter can be configured. The Reboot Action for platform events should be configured with caution because the platform events can cause the server to reboot if the specified trigger event occurs. Reboots initiated by platform events occur without the knowledge of the OS—hence, the OS does not have an opportunity to perform a graceful shutdown first. *Note:* For PET alerts to be generated and sent, the Enable box in the Generate Alert section of the GUI should be selected, as shown in Figure 5.

When properly configured, Dell PowerEdge servers equipped with OMSA can generate and send SNMP traps for system events such as component failure or status change.

The PET alerting feature of the BMC can be globally enabled or disabled by selecting and deselecting the Enable Platform Event Filter Alerts box in the Platform Event Filters Configuration section of the GUI. This setting does not affect any of the platform event–triggered shutdown actions.

For PETs to be transmitted as desired, SNMP trap destinations and the SNMP community must also be configured. Administrators should click "Configure Destinations" on the Platform Events screen, shown in Figure 4, to view the Platform Event Alert Destinations screen and configure the SNMP community. Clicking on a destination number in the destination list displays the Set Platform Events Alert Destination screen. On this screen, administrators can

SO... WHO'S GONNA TAKE THE LEAD ON SECURING THE NETWORK?
PICK ME. (I'M A SURE THING.)



2 alerts currently shown, 2 alerts in database.

ALL_EVENTS

☑ Continuous Updates

Severity	Date	Hostname	Description	EMS Category
⚠	12/01/04 4:43 PM CST	bmc dhcp.smdlab.com	Over-Temperature Warning (Upper non-critical, going high)....	Environmental
⚠	12/01/04 4:43 PM CST	dis-exped-pt1	Temperature sensor detected a warning value [Sensor loca...	Environmental

Description

Temperature sensor detected a warning value
Sensor location: BMC Ambient Temp
Chassis location: Main System Chassis
Previous state was: OK (Normal)
Temperature sensor value (in Degrees Celsius): 28.0

IT Assistant Access: <Read-Write>

Figure 6. ITA GUI showing SNMP trap highlighted

check the Enable Destination box and enter the IP address of the destination system to which the PET alerts should be sent.

Finally, the network interface card (NIC) of the BMC should be enabled and correctly configured so that PET alerts can be sent. Administrators should select System > Main System Chassis > BMC on the left navigation bar of the OMSA GUI, and select Configuration > LAN. Administrators should make sure that the Enable IPMI Over LAN box in the NIC Configuration section of the GUI is selected, and that the NIC has the correct IP address assigned to it—either by using Dynamic Host Configuration Protocol (DHCP) or a static IP address.

PET configuration can also be viewed and managed using CLI commands. Detailed instructions can be found in the OMSA online documentation.

SNMP traps versus PET alerts

The destination system for SNMP traps and PET alerts must have an SNMP manager program installed, such as Dell OpenManage IT Assistant (ITA), which receives, recognizes, filters, and acts upon SNMP traps and PET alerts. PET alerts are, in fact, SNMP traps with specified trap fields defined in the IPMI Platform Event Trap format.

If both SNMP traps and PET alerts are configured for the same type of component and event (for example, when a temperature sensor detects a warning state), an SNMP trap is generated and sent by OMSA when the event occurs, while a PET alert is generated and sent by the affected system's BMC. If destinations for both of the alerts are set to the same management station, which has ITA installed, both the SNMP trap and the PET alert will reach the management station and be recognized, as shown in Figure 6. The SNMP trap (the first line item in Figure 6) and the PET alert (the second line item in Figure 6) have different values for the host name. The SNMP trap uses the OS host name as its host name because the trap is generated and sent by OMSA and the OS, while the PET alert uses the BMC host name because the PET is generated and sent by the BMC. These two alerts also display different values in the description.

Maximized uptime and performance through alerting

Alerting mechanisms provided by OMSA allow administrators to be notified of system events, or status changes of system components or of the system itself. Different alerting mechanisms—at the local, remote, or hardware and firmware levels—provide administrators with options and flexibility regarding how they can be warned of system abnormalities. The sooner and more accurately administrators are notified of potential and actual system failures, the more time they have to discover the root cause of a given problem and take appropriate action to help maintain continuous system uptime.

SNMP traps can be enabled or disabled by system, by category of component, by individual component, and by severity level.

Haihong Zhuo is a software engineer consultant working in the System Management Instrumentation team in the Dell Enterprise Software Development Group. She participated in the development of ITA 6.x. Haihong has an M.S. in Computer Engineering from The University of Texas at Austin and a B.S. in Electrical Engineering from Tsinghua University in China.

Michael O'Hara is a software engineer senior consultant on the Managed Node Framework Development team in the Dell Product Group. He focuses on OMSA user interfaces, with particular emphasis on the CLI. Michael has a BSc. in Mathematics from the University of Warwick in England.

Jianwen Yin, Ph.D., is a software engineer senior consultant on the Managed Node Framework Development team in the Dell Product Group. His major focus is on OMSA and systems management. Jianwen has a Ph.D. in Computer Science from Texas A&M University.

FOR MORE INFORMATION

IPMI 1.5 overview and specification:
www.intel.com/design/servers/ipmi/index.htm

IPMI PET 1.0 specification:
"IPMI Platform Event Trap Format Specification v1.0" by Intel, Hewlett-Packard, NEC, and Dell, ftp://download.intel.com/design/servers/ipmi/pet100.pdf

OMSA online documentation:
docs.us.dell.com/docs/software/svradmin/index.htm

Deploying Dell Update Packages Using Microsoft Systems Management Server 2003

Administrators can gain a powerful set of systems management tools by integrating Dell™ Update Packages with Microsoft® Systems Management Server (SMS) 2003. This article provides an overview of how administrators can use Microsoft SMS 2003 to build a software distribution solution that helps manage Dell PowerEdge™ servers in a heterogeneous environment.

BY SANDEEP KARANDIKAR AND MANOJ GUJARATHI

Enterprise IT departments typically manage a heterogeneous mix of decentralized hardware and software applications. Successful administration of complex environments can be challenging, and IT professionals rely on robust systems management offerings to help achieve their goals.

One of the thorniest enterprise management issues is patch management, which includes keeping application software, operating systems, and system software up-to-date. The basic tasks of patch management include inventorying an organization's hardware for existing software version levels; comparing the inventory results to a list of current or desired software versions; and allowing users to apply one or more updates.

To address enterprise needs for enhanced patch management, Dell is expanding Dell OpenManage™ system software update technologies and products. This article discusses how Dell Update Packages can be integrated with Microsoft Systems Management Server (SMS) 2003 to manage Dell PowerEdge servers in a heterogeneous enterprise environment.

Dell Update Packages are server BIOS, firmware, and driver system software updates delivered in a consistent, self-extracting zip file format. They are available on the Dell support site (support.dell.com) and offer ease and flexibility for updating system software on Dell PowerEdge servers in an operating system (OS)–present environment. The current suites of packages allow administrators to update the following system software components:[1]

- System BIOS
- System firmware, also known as the Embedded Server Management (ESM) firmware
- Remote access controller (RAC) firmware
- PowerEdge Expandable RAID Controller (PERC) firmware and device drivers
- Network interface card (NIC) drivers

[1] For the current list of supported devices, visit the Dell support site at support.dell.com.

Dell Update Packages offer key software distribution functionality[2] such as:

- Capability for updating the BIOS, firmware, and drivers supplied and supported by Dell
- A command-line interface (CLI) for batch access to the update management functions
- Installed-server software inventory using command-line switches
- Self-contained, built-in authentication and runtime dependency checks
- A cumulative logging facility to track server software update activity

Integrating Dell Update Packages with SMS 2003

Microsoft SMS 2003 provides systems management capabilities that address the areas of security patch management, application deployment, and asset management using Windows® management services integration. Using Dell Update Packages, administrators can perform the following functions:

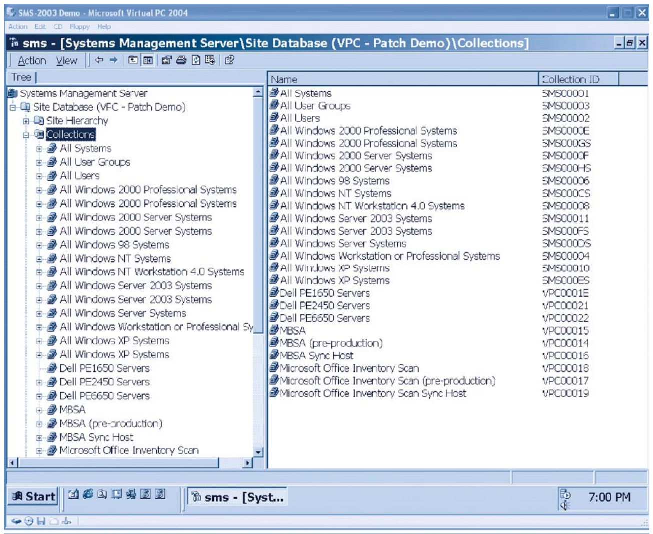
- Apply an individual update to a system through CLI execution using update package distribution parameters
- Control the update distribution sequence to ensure that updates occur in an appropriate order
- Batch together multiple updates for deployment across a manageable collection
- Use Microsoft resource kit tools to interpret return codes provided by Dell Update Packages

Integrating Dell Update Packages with SMS 2003 helps provide the following broad areas of functionality:

- Creating manageable collections
- Selecting appropriate packages for distribution
- Creating advertisements for update packages
- Authorizing updates
- Verifying updates

Creating manageable collections

In heterogeneous environments, IT administrators must identify Dell PowerEdge servers to target these servers for system software updates. SMS 2003 provides built-in capabilities to identify a group of servers on the network by OS type—for example, all systems running Microsoft Windows 2000 Advanced Server.


SMS-2003 Demo - Microsoft Virtual PC 2004
Action Edit CD Floppy Help
sms - [Systems Management Server\Site Database (VPC - Patch Demo)\Collections]
Action View
Tree
Systems Management Server
Site Database (VPC - Patch Demo)
Site Hierarchy
Collections
All Systems
All User Groups
All Users
All Windows 2000 Professional Systems
All Windows 2000 Professional Systems
All Windows 2000 Server Systems
All Windows 2000 Server Systems
All Windows 98 Systems
All Windows NT Systems
All Windows NT Workstation 4.0 Systems
All Windows Server 2003 Systems
All Windows Server 2003 Systems
All Windows Server Systems
All Windows Workstation or Professional Sy
All Windows XP Systems
All Windows XP Systems
Dell PE1650 Servers
Dell PE2450 Servers
Dell PE6650 Servers
MBSA
MBSA (pre-production)
MBSA Sync Host
Microsoft Office Inventory Scan
Start
sms - [Syst...
7:00 PM

Name	Collection ID
All Systems	SMS00001
All User Groups	SMS00003
All Users	SMS00002
All Windows 2000 Professional Systems	SMS0000E
All Windows 2000 Professional Systems	SMS000GS
All Windows 2000 Server Systems	SMS0000F
All Windows 2000 Server Systems	SMS000HS
All Windows 98 Systems	SMS00006
All Windows NT Systems	SMS000CS
All Windows NT Workstation 4.0 Systems	SMS00008
All Windows Server 2003 Systems	SMS00011
All Windows Server 2003 Systems	SMS000FS
All Windows Server Systems	SMS000DS
All Windows Workstation or Professional Systems	SMS00004
All Windows XP Systems	SMS00010
All Windows XP Systems	SMS000ES
Dell PE1650 Servers	VPC0001E
Dell PE2450 Servers	VPC00021
Dell PE6650 Servers	VPC00022
MBSA	VPC00015
MBSA (pre-production)	VPC00014
MBSA Sync Host	VPC00016
Microsoft Office Inventory Scan	VPC00018
Microsoft Office Inventory Scan (pre-production)	VPC00017
Microsoft Office Inventory Scan Sync Host	VPC00019

Figure 1. Creating collections for Dell PowerEdge servers

IT administrators create collections by installing the Microsoft Baseline Security Analyzer (MBSA) or Microsoft Office Inventory Scan tools to distribute security patches or Office update packages. A collection represents a single view of the various assets in the network based on certain predefined grouping criteria. IT administrators can use Microsoft scanning tools to create hardware-focused grouping criteria that is based on the existing network asset information in the SMS repository database. This information is gathered by the SMS client on managed systems through Windows Management Instrumentation (WMI) services.

Larger enterprises typically have four to five models of Dell PowerEdge server platforms spanning two to three generations. Administrators can use server model numbers to group similar servers in a collection. For example, in a corporate data center hosted on 200 Dell PowerEdge 1650 servers and 200 Dell PowerEdge 6650 servers, administrators could create two collections entitled "Dell PowerEdge 1650 servers" and "Dell PowerEdge 6650 servers," as shown in Figure 1.

Once administrators create collections in the SMS site console, the collections can be populated. Administrators define membership rules using the data collected by the SMS client from various nodes. By using these built-in SMS mechanisms, administrators can group various Dell servers and create unique collections based on the type of servers, or make these collections even more granular based on the type of Windows OS or the role of the server, for example.

[2] For more information about Dell Updates Packages, refer to the documentation at support.dell.com or see "An Introduction to Dell Update Packages" by Karl Friedrich and Sandeep Karandikar in *Dell Power Solutions,* August 2003; and "Scripting Dell Update Packages on Windows and Linux" by Manoj Gujarathi, Pritesh Prabhu, and Subbu Ganesan in *Dell Power Solutions,* October 2004.

DELL™

This hardware-centric view of the network enables the following advantages from a systems management perspective:

- Brings similar Dell PowerEdge servers to the same level by deploying a relevant collection of Dell Update Packages
- Reduces network traffic by targeting only systems that need updates
- Simplifies deployment by phasing updates across different server collections
- Isolates and manages mission-critical systems using the same software distribution method

Network administrators should set up test beds comprising different models of Dell PowerEdge servers to evaluate system software update packages before deploying them in the production environment.

Selecting appropriate packages for distribution

After creating and populating the collections, administrators must determine which packages need to be distributed, and then make the SMS site console aware of those packages. Once all packages that need to be deployed on Dell PowerEdge servers have been identified, administrators can apply those packages using SMS wizards.

Dell Update Packages are released periodically on support.dell.com and quarterly through subscription CD releases provided by the Dell Custom Solutions Group. Each subscription CD contains a validated set of packages, which are also available through the support Web site. Organizations can determine the applicability of these update packages to the systems in their collections by using one of the following mechanisms:

- Reviewing the release information available at support.dell.com accompanying each system software update released
- Using the informational spreadsheet provided with each subscription CD released from the Dell Custom Solutions Group
- Reviewing the directory structure from each subscription CD to determine applicability of patches for a collection
- Double-clicking the page executable to open the package and display the following release information contained within each package: release date, system software versioning information, description of the update package, hardware devices that the package will update, server and OS type that the package supports, and prerequisites that must be satisfied on the target system before applying updates

Reviewing the package information using any of the preceding four methods allows IT administrators to determine which collections are eligible for the updates. For example, BIOS and ESM firmware updates are released per server model and thus can be distributed only to those specific server models. Device drivers, on the other hand, are hardware-device specific and not platform or server specific. The execution logic necessary to confirm the appropriateness of the update for the target system is bundled with the update, along with verification of software compatibility. In addition, administrators can use other package applicability constraints such as supported OS to restrict the execution of packages on supported OS configurations through the SMS site console.

IT administrators can also use the inventory option provided by each Dell Update Package to determine applicability of the package for the target system. *Note:* This operation should be performed on a test system to validate the packages before they are distributed in a production environment. Refer to the "Sequencing updates" section in this article to determine the order in which the updates should be applied to Dell PowerEdge servers.

SMS 2003 provides a wizard-driven interface for package distribution. Administrators can use a shared SMS distribution point option instead of distributing update packages to individual systems. Dell Update Packages require administrative privileges for execution on the managed system. After applying packages, administrators should note the return code from execution of these packages. Administrators can choose to create their own mapping to interpret the error codes using the Management Information Format (MIF) file specification option in the package properties.[3]

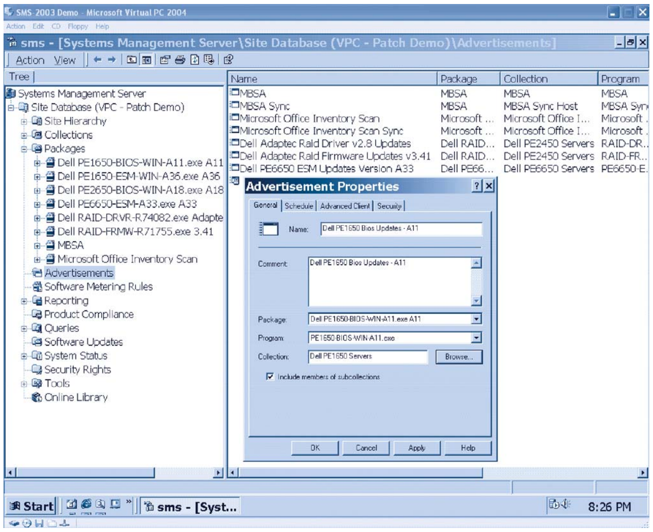
Creating advertisements for update packages

Once the collection criteria have been defined and applicable packages selected, the next step in the software update distribution process is to create an advertisement for the update packages. The advertisement creation process involves directing the packages defined in the SMS site console to the appropriate collection by creating an advertisement for each update package. Figure 2 shows an example of advertisement creation.

In Figure 2, the Dell PE1650-BIOS-WIN-A11.exe update package advertisement for the Dell PowerEdge 1650 servers is specified in the "Dell PE1650 Servers" collection. Without a manageable collection defined, the package would essentially be distributed to every Dell and non-Dell server in a heterogeneous environment. Creating an advertisement allows administrators to further refine their targets—in this example, the update is distributed only to those PowerEdge 1650 servers on which the BIOS update is applicable.

To create an advertisement, administrators must specify the various command-line options with which the Dell Update Packages need

[3] For more details about this option, refer to *Systems Management Server (SMS) 2003 Concepts, Planning, and Deployment Guide* by Microsoft Corporation, www.microsoft.com/resources/documentation/sms/2003/all/cpdg/en-us/default.mspx.


SMS-2003 Demo - Microsoft Virtual PC 2004
Action Edit CD Floppy Help
sms - [Systems Management Server\Site Database (VPC - Patch Demo)\Advertisements]
Action View
Tree
Systems Management Server
Site Database (VPC - Patch Demo)
Site Hierarchy
Collections
Packages
Dell PE1650-BIOS-WIN-A11.exe A11
Dell PE1650-ESM-WIN-A36.exe A36
Dell PE2650-BIOS-WIN-A18.exe A18
Dell PE6650-ESM-A33.exe A33
Dell RAID-DRVR-R74082.exe Adapte
Dell RAID-FRMW-R71755.exe 3.41
MBSA
Microsoft Office Inventory Scan
Advertisements
Software Metering Rules
Reporting
Product Compliance
Queries
Software Updates
System Status
Security Rights
Tools
Online Library
Advertisement Properties
General | Schedule | Advanced Client | Security
Browse...
OK Cancel Apply Help
Start
sms - [Syst...
8:26 PM

Name	Package	Collection	Program
MBSA	MBSA	MBSA	MBSA
MBSA Sync	MBSA	MBSA Sync Host	MBSA Syn
Microsoft Office Inventory Scan	Microsoft ...	Microsoft Office I...	Microsoft .
Microsoft Office Inventory Scan Sync	Microsoft ...	Microsoft Office I...	Microsoft .
Dell Adaptec Raid Driver v2.8 Updates	Dell RAID...	Dell PE2450 Servers	RAID-DR..
Dell Adaptec Raid Firmware Updates v3.41	Dell RAID...	Dell PE2450 Servers	RAID-FR..
Dell PE6650 ESM Updates Version A33	Dell PE66...	Dell PE6650 Servers	PE6650-E.

Name: Dell PE1650 Bios Updates - A11

Comment: Dell PE1650 Bios Updates - A11

Package: Dell PE1650-BIOS-WIN-A11.exe A11

Program: PE1650-BIOS-WIN-A11.exe

Collection: Dell PE1650 Servers

☑ Include members of subcollections

Figure 2. Advertisement creation for a Dell Update Package

to be executed. Figure 3 lists available command-line options.

Each update package is designed to update specific device software on a given type of server. Use of a software distribution service such as SMS 2003 means that update packages can be run in silent mode, without user intervention, as defined by the `/s` option. Some of the updates require a system reboot to become active. Administrators can choose to run multiple updates sequentially and schedule a single reboot once the final update is completed. Doing so means using the `/r` option only once, when the final update package is pushed to managed systems. Alternatively, administrators can have the system restart capability of SMS 2003 perform a similar task.

If the target systems have Dell OpenManage Server Administrator installed, administrators can use the Dell OpenManage Server Administrator CLI to inventory the target systems. If Dell OpenManage Server Administrator is not installed, administrators can use the update package's CLI to inventory the target systems. Update packages are self-contained and do not require additional components to be installed on managed systems. Each update package performs three functions during execution:

- Inventory the system to determine the version of software installed
- Compare the installed software version with the one available in the update package
- Execute an upgrade, downgrade, or reinstallation based on the command-line option selected

Execution is halted whenever the prerequisites for the application update are not met. Figure 4 shows commonly performed tasks and the command-line options to invoke them.

Authorizing updates

The update authorization process involves selecting the runtime environment and scheduling the updates. This process occurs during the creation of an advertisement, and administrators should note the following when scheduling updates:

- To guarantee that the update sequencing is honored, administrators should not create advertisements with conflicting schedules. Out-of-sequence updates create the potential for failure when update prerequisites are not met.
- To optimize network bandwidth, SMS uses distribution points to share packages across different systems. In such scenarios, administrators should be careful when specifying log locations through command-line parameters for different updates. These logs must be local to each target system.
- The update authorization process needs to take into account the planned maintenance window as well as the system downtime incurred to complete the updates.
- No tape backup or resource-intensive system activity should be in progress when applying updates.
- By using collections, administrators can phase out deployment of system software updates across the organizations.
- Since Dell Update Packages update firmware software on the embedded system devices, or on drivers that talk to the attached peripherals, no rollback capabilities are provided. If a rollback is desired, administrators must download an earlier version of the firmware software and use SMS to distribute the update. System software deployment requires advance planning so administrators can ensure that all updates are applied for a given collection in the allotted maintenance window.

SMS provides the capability to force the target systems to be updated after the expiration of the advertisement by making the system software updates mandatory.

Command-line option	Description
`/?` or `/h`	Displays command-line usage help
`/e=`*path*	Extracts files to a specified path*
`/s`	Executes the update package in silent mode
`/f`	Forces a downgrade or update to the same or an older version*
`/c`	Determines whether the update can be applied to the target system*
`/r`	Reboots if necessary after the update*
`/l=`*file*	Appends log messages to a specified ASCII file*
`/u=`*file*	Appends log messages to a specified Unicode file*

*Must be used with `/s`.

Figure 3. Dell Update Package command-line options

CLI task	Syntax
Apply the update in unattended, or *silent,* mode; force a downgrade or update to the same version.	*pkgname*.exe /s /f
Apply the update in silent mode; log the execution results to a file named *pkgname*.log.	*pkgname*.exe /s /l=%*systemdrive*%\Dell\*pkgname*.log *
Determine whether the update can be applied to the target system.	*pkgname*.exe /c
Apply the update in silent mode; reboot the system if required.	*pkgname*.exe /s /r

*SMS 2003 handles software distribution by using network shares or distribution points. Specifying %*systemdrive*% in the syntax ensures that logs are created locally on the managed system or in the context of package execution.

Figure 4. Common CLI options for Dell Update Packages

Verifying updates

Dell Update Packages return well-defined error codes in addition to the entries that are generated in the Windows system event log (SEL). Microsoft resource kit tools provide utilities to remap these error codes into codes that the SMS site console can understand. Figure 5 describes error codes that are returned from the Dell Update Packages.

Administrators can review additional details about each of these errors by retrieving the logs from the target systems. The entry in the Windows SEL specifies the result of the execution and the location of the log files. The results from the software updates are archived locally on the managed system. By default, these logs are available at *SystemDrive*\Dell\UpdatePackage\log\ on target systems.

Return code	Message	Description
0	SUCCESS	The update operation was successful and a reboot is not necessary.
1	UNSUCCESSFUL	The update failed because an error occurred during operation.
2	REBOOT_REQUIRED	The update was successful and a reboot is required to make the update active.
3	SOFT_DEP_ERROR	The update was not successful because the target system already had the updated version or a lower version of the BIOS, driver, or firmware provided by the update package. To avoid receiving this error, administrators can provide the /f option. However, using /f allows systems to potentially be downgraded.
4	HARD_DEP_ERROR	The update was unsuccessful because the server did not meet BIOS, driver, or firmware prerequisites necessary for the update to be applied, or no supported device was found on the target system.

Figure 5. Dell Update Package return codes

Log file names, which typically use the format *pkgname*.log, are overwritten each time the package executes. The /l or /u command-line switch in the package creation wizard allows administrators to specify alternate log path and file names. The logs are appended each time the package executes. When the /l option is used, these logs provide an audit trail of various updates applied to the system. The logs from all managed systems can be archived in a central repository to track system software upgrades.

Sequencing updates

Dell Update Packages may have dependencies on each other, and administrators must deploy these packages through SMS 2003 in the correct order.[4] Administrators must also understand Dell Update Package prerequisites to facilitate streamlined deployment. Failure to understand the correct update sequence and update dependencies on other system components can cause unnecessary reboots or even render target systems unstable.

Deploying updates to Dell PowerEdge servers

Dell Update Packages are designed to provide administrators with a powerful set of system software management tools. Integrating Dell Update Packages with Microsoft SMS 2003 greatly empowers administrators to remotely deploy Dell Update Packages for various system components on a set of Dell PowerEdge servers. This approach not only provides flexibility in deploying updates to a few Dell servers—or a few hundred—but it also helps ease systems maintenance by bringing server components to a known, appropriate level.

Sandeep Karandikar works as a systems consultant in the Dell Product Group, Engineering. He has an M.S. in Computer Science from the University of Colorado, Boulder, and a B.S. in Telecommunications from the University of Bombay in India.

Manoj Gujarathi is a systems engineer in the Dell OpenManage Software Development Group. He has worked on various Dell OpenManage applications including Dell Update Packages. Manoj has an M.S. in Engineering from Washington State University and an M.S. in Computer Science from Texas Tech University.

[4] For more details about dependencies and scripting techniques for Dell Update Packages, refer to "Scripting Dell Update Packages on Windows and Linux" by Manoj Gujarathi, Pritesh Prabhu, and Subbu Ganesan in *Dell Power Solutions,* October 2004.

SAN Connectivity Kit

for Dell/EMC AX100 and EMC CLARiiON AX100

Switches, HBAs, and software all in one box. Just add storage!

Everything you need to build and manage a small SAN … all in one SKU. EMC lab tested for the AX100 storage platform.

QLogic SAN Connectivity Kit™ for AX100

(Order#SAN-C3050-E)

- 1 SANbox 3050-E Fibre Channel Switch
- 2 SANblade QLA200-E Host Bus Adapters (HBAs)
- SANsurfer® Management Software
- 8 SFPs

All for a Great Price!

Add-on Switches and HBAs

SANbox® 3050 Fibre Channel Switch for AX100

(Order#SB3050-08A-E)

- Wizard-based installation takes just minutes
- 8 – 2Gb ports
- SANsurfer® Management Software... Included

SANblade™ QLA200 HBA for AX100

(Order#QLA200-E-SP)

- Wizard-based installation takes just minutes
- Multi-OS support
- SANsurfer® Management Software... Included

For more information, visit **www.Dell.com** and search for **A0434012**.

SIMPLE | LOW COST | SANS

POWERED BY QLOGIC™

Managing Dell Client Systems with

Enterprise Management Solutions

By using enterprise management solutions, administrators can readily deploy additional desktop systems—changing the BIOS, setting security options, and installing and deploying security software. This article provides examples and step-by-step procedures to help administrators create packages with the Dell™ Client Configuration Utility and to distribute packages using the Dell OpenManage™ Client Administrator.

BY JIM LATHAN

Managing a large number of client systems is a challenge for IT staff within any organization. It is common for organizations to have hundreds or thousands of desktop and mobile PCs, often spread across multiple locations. Managing these systems requires an effective systems management strategy as well as a suitable open-standard, multivendor set of systems management tools. An important component of Dell's client systems management strategy is provided in the Dell Client Configuration Utility (DCCU). This powerful yet simple tool allows configuration management of multiple systems, asset and inventory reporting, and hardware security settings. In addition, DCCU enables interoperability with existing local area network (LAN)/wide area network (WAN) infrastructures and systems management solutions such as Dell OpenManage Client Administrator (OMCA), Altiris Client Management Suite, and Microsoft® Systems Management Server (SMS).

Using the DCCU and OMCA to upgrade client systems

The DCCU can display inventory and system parameters of managed Dell PCs such as Dell Precision™, OptiPlex™, and Latitude™ systems—including the identification of operating system (OS) version, processor speed, and total available memory. This information can be critical when determining how to upgrade a system that may not be in the same location as the IT staff.

Administrators also need the ability to manage and configure the settings on their PCs, and the DCCU can provide this capability through a user-friendly, interactive graphical user interface (GUI) with nearly 50 manageable parameters in the System Management BIOS (SMBIOS), CMOS (complementary metal-oxide semiconductor), and OS. Each parameter is listed with all of its allowable values, where applicable. Descriptions of the parameters are included and can be accessed by simply moving the cursor over the name of each parameter in the GUI application. Enabling Wake-on-LAN (WOL) and Preboot Execution Environment (PXE) on every client PC is an example of how the DCCU allows system administrators to accomplish several tasks with little effort.

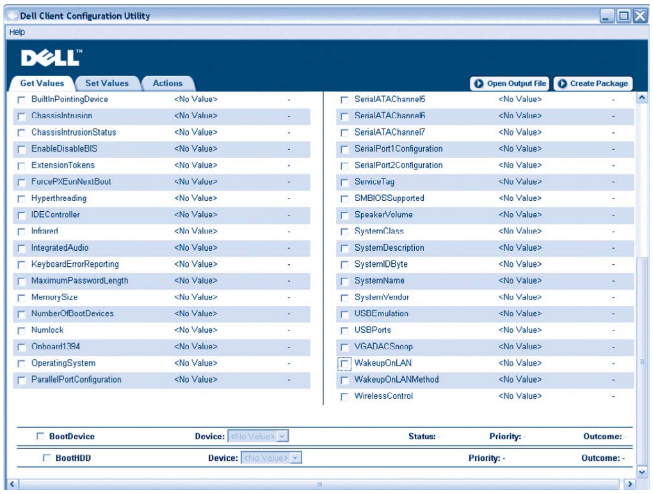
During a PC's three- to four-year life cycle, administrators sometimes must apply an update to the firmware or BIOS of the PC. Even for a highly stable PC platform like the Dell OptiPlex desktop, at least one BIOS update is recommended. With the DCCU, Dell offers IT managers a simple yet powerful tool for managing PCs both locally and remotely. The DCCU enables administrators to update the BIOS of remote systems by creating a small executable package on the local administrator machine. The process is simple:

1. Use the `-nopause` command option when executing the BIOS file for a system BIOS that supports silent installations; for older systems or a BIOS that does not support silent installations, download the updated BIOS from the Dell support Web site (support.dell.com).
2. Convert the unpackaged system BIOS file to a BIOS .hdr file using the `writehdrfile` command option. For example:

   ```
   Sx270A04.exe -writehdrfile
   ```

3. Select the flash function from within the DCCU and browse to the downloaded BIOS file.
4. Create the executable package and distribute it to the target PCs for execution. After delivery, the PCs will automatically reboot and perform the update—preserving the integrity of the BIOS settings—and then automatically remove the update package from each PC, leaving behind only a tiny XML results file.

Dell OMCA has an extensive reporting feature that is designed to provide details of all instrumented PCs and servers in an organization. This powerful reporting feature allows administrators to compile reports describing the make and model of each instrumented PC and even the BIOS revision. In OMCA, these custom reports are known as collections. Using the OMCA Deployment Server Console, administrators can create a package that contains the newly created DCCU executable and simply drag and drop that package to a system or a group of systems for scheduled deployment. For example, administrators can disable the floppy drive in all public-access computers or change the automatic boot order on all PCs in their organization.



□ BuiltInPointingDevice	<No Value>	-
□ ChassisIntrusion	<No Value>	-
□ ChassisIntrusionStatus	<No Value>	-
□ EnableDisableBIS	<No Value>	-
□ ExtensionTokens	<No Value>	-
□ ForcePXEonNextBoot	<No Value>	-
□ Hyperthreading	<No Value>	-
□ IDEController	<No Value>	-
□ Infrared	<No Value>	-
□ IntegratedAudio	<No Value>	-
□ KeyboardErrorReporting	<No Value>	-
□ MaximumPasswordLength	<No Value>	-
□ MemorySize	<No Value>	-
□ NumberOfBootDevices	<No Value>	-
□ Numlock	<No Value>	-
□ Onboard1394	<No Value>	-
□ OperatingSystem	<No Value>	-
□ ParallelPortConfiguration	<No Value>	-
□ SerialATAChannel5	<No Value>	-
□ SerialATAChannel6	<No Value>	-
□ SerialATAChannel7	<No Value>	-
□ SerialPort1Configuration	<No Value>	-
□ SerialPort2Configuration	<No Value>	-
□ ServiceTag	<No Value>	-
□ SMBIOSSupported	<No Value>	-
□ SpeakerVolume	<No Value>	-
□ SystemClass	<No Value>	-
□ SystemDescription	<No Value>	-
□ SystemIDByte	<No Value>	-
□ SystemName	<No Value>	-
□ SystemVendor	<No Value>	-
□ USBEmulation	<No Value>	-
□ USBPorts	<No Value>	-
□ VGADACSnoop	<No Value>	-
□ WakeupOnLAN	<No Value>	-
□ WakeupOnLANMethod	<No Value>	-
□ WirelessControl	<No Value>	-

□ BootDevice Device: <No Value> Status: - Priority: - Outcome: -

□ BootHDD Device: <No Value> Priority: - Outcome: -

Figure 1. Dell Client Configuration Utility: Get Values screen

Creating update packages with the DCCU

The DCCU requires the Microsoft .NET Framework to be installed on the system that will be used to create the DCCU packages (usually the administrator's system), but it does not require the Microsoft .NET Framework to be installed on the end-user systems on which the DCCU will perform tasks. Visit www.dell.com/openmanage and click the Client Systems tab to download the DCCU.

Creating a package with the DCCU is very simple; the GUI allows administrators to select which BIOS fields to audit or change. Consider an example scenario in which an administrator creates a package to enable WOL. First, the administrator launches the DCCU from the Start > Programs > Dell Applications menu. Figure 1 illustrates the default startup of the DCCU.

The default display of the DCCU allows the administrator to create a GET VALUES package, which is used to retrieve the current BIOS and OS settings on the end-user systems. To change the boot order, a GET package must be created with the BootDevice and BootHDD options selected. The GET package creates a results file that is retrieved and imported into the DCCU console, providing the administrator with the current Hard Drive and Boot Order settings and options for the specific system. A GET package should be performed on the specific system for which the boot order must be modified. This will ensure that the SET package works correctly when deployed by OMCA.

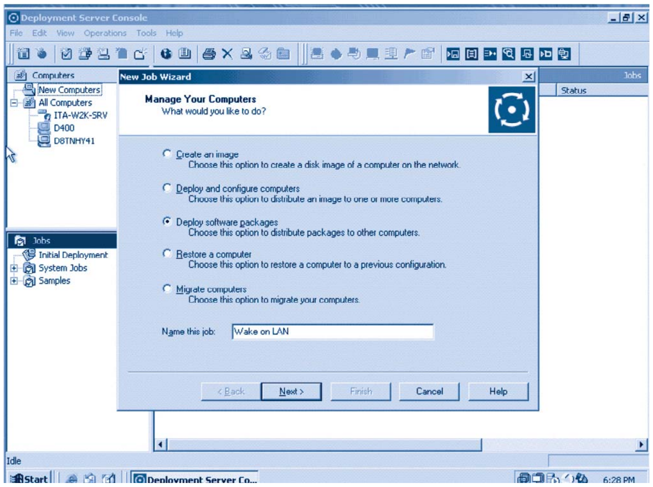
A SET VALUES package is created by selecting the SET VALUES tab at the top of the DCCU GUI. The WOL feature is set by scrolling down until the WakeupOnLAN option appears.

The administrator selects the WakeupOnLAN check box, which enables the value selection drop-down menu. System administrators can select the option that best fits their needs, but "enabled for all NICs" typically is the optimal choice to enable WOL.

The administrator finishes the package creation by clicking the Create Package button at the top right of the DCCU interface. The DCCU will prompt for a location and file name to assign the new package. The default location where the DCCU saves the created packages is C:\Program Files\Dell\Dell Client Configuration Utility\Client\packages, and *PackageName*.exe is the default file-naming convention for DCCU. In this example scenario, the DCCU package to be created is called WOL.exe and is located in the default DCCU directory location. Next, DCCU will compile the WOL package and confirm its success by displaying a message on the console. The WOL package is now ready for delivery to the client systems.

Distributing update packages with OMCA

System administrators can deliver update packages, such as the WOL package described in the preceding scenario, to Dell client systems with various enterprise systems management application



Manage Your Computers
What would you like to do?

○ Create an image
Choose this option to create a disk image of a computer on the network.

○ Deploy and configure computers
Choose this option to distribute an image to one or more computers.

● Deploy software packages
Choose this option to distribute packages to other computers.

○ Restore a computer
Choose this option to restore a computer to a previous configuration.

○ Migrate computers
Choose this option to migrate your computers.

Name this job: Wake on LAN

Figure 2. OMCA Deployment Server Console: Using the New Job Wizard to deploy software packages

suites, such as Microsoft SMS or Altiris Deployment Solution—or even distribute the packages as part of an NT logon script. Dell has partnered with Altiris to create OMCA, an affordable yet powerful suite of applications that are designed to help maximize total cost of ownership (TCO) and minimize the amount of resources needed to manage enterprise-wide systems. The OMCA suite is upgradeable to the full suite of Altiris products for specific license fees.

The DCCU uses the OMCA Deployment Server Console to display all systems that have been discovered and are managed by OMCA. By using the Deployment Server Console, an administrator in the example scenario can distribute the WOL package to the desired systems. Several methods can be used to create the DCCU job, but the example scenario uses the New Job Wizard method. With the console open, the administrator right-clicks on the Jobs section of the console and selects "New Job Wizard" to start the job creation process. From the wizard, the administrator selects "Deploy Software Packages" and enters a name for the job so that it is easily identified (see Figure 2).

The inventory report feature in OMCA is easy to access. By clicking the Next button, administrators can use the New Job Wizard to access the OMCA inventory and use the inventory items as conditions for executing jobs. Here, the administrator can specify which systems receive the WOL package. To create a package that uses no conditions for execution, the administrator can select "Not using condition for the job" and click the Next button.

The New Job Wizard needs to know where the WOL package resides, so the administrator clicks the Add button on the next page of the wizard—the RapidInstall and PC Transplant Packages screen—to open the navigation window and provide the location of the WOL package.

The administrator selects the WOL.exe file and clicks the Open button; a warning message appears stating that the package selected is not a valid RapidInstall or PC Transplant package and asks the administrator whether to continue with the selection. The administrator selects "Yes" and the wizard updates the New Job Wizard window. More than one package can be added in this window.

The administrator clicks the Next button to display the Select Computers screen; here, the administrator can either assign the job to a specific computer or group of computers or not apply the job to any computer at this time. Not assigning the job allows the administrator to drag and drop the job to computers in the Systems list on the main console. The administrator closes the New Job Wizard by clicking the Next button and then the Finish button.

Completing the New Job Wizard places the newly defined job in the Jobs menu on the main OMCA Deployment Server Console. After selecting the WOL job and dragging the job to a computer for deployment, the administrator is prompted to schedule the job or execute it immediately. The default setting, when using the drag-and-drop method, is to execute the package immediately. In the example scenario, the WOL package will then be deployed by the OMCA console; after OMCA verifies that the WOL package is intact, OMCA will execute WOL.exe, leaving only the results file documenting either the success or failure of setting the BIOS feature.

Managing Dell client systems effectively

The example scenario described in this article helps illustrate how easily and effectively system administrators can manage just one or thousands of Dell client systems—whether Dell Precision workstations, OptiPlex desktop PCs, Latitude notebooks, or a combination thereof—without having to physically interact with the end-user computer. By using the Dell Client Configuration Utility with Dell OpenManage Client Administrator, Altiris Client Management Suite, or Microsoft SMS, administrators can deploy additional systems, change BIOS settings, configure security options, install and deploy software, and apply updates or patches in a fraction of the time that would be required without using these tools. This reduction in systems management time can lead to an overall reduction in TCO, enabling organizations to reallocate valuable IT staff resources to other tasks.

Jim Lathan is a Dell client technologist based in Elgin, South Carolina. Specializing in desktop, notebook, and workstation pre-sales deployment, Jim has Microsoft Certified Systems Engineer (MCSE), Dell Certified Systems Expert (DCSE), and Cisco Certified Network Associate (CCNA) certifications as well as 20 years of experience in IT management and support.

FOR MORE INFORMATION

Dell OpenManage:
www.dell.com/openmanage

Agentless Monitoring of Dell PowerEdge Servers Using Mercury SiteScope

Mercury SiteScope® software is an agentless, end-to-end infrastructure performance and availability monitoring tool. It can help organizations to keep their IT infrastructure—including Dell™ PowerEdge™ servers—up and running. This article provides an overview of Mercury SiteScope, including information about the solution's conformance to the Intelligent Platform Management Interface (IPMI) specification, and explains how SiteScope can be used to manage Dell servers.

BY BILL FITZGERALD AND BOB URE

In today's organizations, IT systems play a mission-critical role. Operations groups must be able to ensure that the enterprise's core business applications are constantly available to users worldwide. To accomplish this goal, IT teams require cost-effective, rapidly deployable monitoring solutions that can help ensure uptime and maximize the performance of a company's IT infrastructure.

The IT industry's earliest systems monitoring solutions were agent-based. With this type of architecture, an additional piece of software, or *agent,* is installed and run on the physical production system that is being monitored. Once this agent is installed and configured properly, it monitors only the system on which it resides for uptime and performance, and then sends information on the system's performance metrics to an alerting console or central monitoring server.

Though agent-based monitoring solutions may be effective for some implementations including mainframe environments, they add a level of complexity, cost, and overhead that is not suitable for many environments—especially large, complex, distributed applications.

The evolution to agentless monitoring

IT management technologies have steadily evolved since the earliest agent-based solutions. Today's IT systems are typically based on multitiered, distributed, and clustered infrastructures, which often consist of hundreds if not thousands of physical systems. These complex environments are not adequately served by agent-based monitoring systems, primarily because of an agent's intrusiveness and its cost to install and maintain. An effective monitoring solution must be easy to deploy and provide low total cost of ownership (TCO) as well as high return on investment (ROI).

Agentless monitoring enables operations groups to monitor complex, distributed systems without installing agents or software on the production systems. Agentless monitoring solutions are designed to nonintrusively monitor all parts of the IT system remotely from a single host system, creating a wide variety of benefits that help lead to higher ROI, including:

- **Enabling rapid deployment:** Whereas agent-based monitoring systems can be time-consuming to install and configure, often requiring the assistance of engineers and experts, agentless monitoring solutions can be deployed at a much more accelerated pace. To monitor multiple systems with an agent-based solution, operations groups have to install and configure agents on every system. By contrast, an agentless solution is installed only once on a host system, and then configured to connect remotely to monitor each system in the IT environment.
- **Easing maintenance and upgrades:** With an agent-based solution, it takes a team of experts to maintain and reconfigure all of the agents on every system as monitoring requirements evolve and change. When upgrades to a newer version of the agent are required, the maintenance team has to visit every production server and upgrade each agent individually. This amount of effort dramatically increases the TCO of monitoring software. In contrast, if a monitor needs to be reconfigured or the software needs to be upgraded for an agentless solution, one IT staff member can accomplish the task with minimal effort because all monitoring resides on a single, central server.

SiteScope can provide Dell customers with an end-to-end solution that is designed to consolidate Dell OpenManage, IPMI, application, transaction, and core infrastructure health and performance information into one easily managed console.

- **Reducing the risk of affecting production systems:** Installing an agent on a production system means adding a new variable to the risk matrix by increasing the complexity and functional interoperability among all installed components. Often, the agent's intrusiveness causes the very performance degradation that it was intended to detect and alert upon.

  Consider a database server that is being monitored by an agent. The agent could consume CPU and memory resources while filling disk space through log-file data accumulation. As a result, the agent could be the direct cause of a performance slowdown or a server crash. Agentless solutions do not require any installation of software on production systems. Because they are nonintrusive, agentless monitoring solutions help mitigate the risk of being the actual cause of performance degradation or system failure.
- **Allowing for system infrastructure and expansion:** Early mainframe or custom client/server environments were fairly static by nature. Once a mainframe was deployed or an application developed in-house was online, the infrastructure remained unchanged for long stretches of time. In contrast, a huge benefit of distributed systems running Web, enterprise resource planning (ERP), customer relationship management (CRM), or e-mail applications is that they can be easily expanded to accommodate growth and scale accordingly.

  Many successful online retail applications experience the addition of tens of servers every week and the changing of application code daily. Agent-based monitoring solutions often require experts to install and configure an agent on each additional system or application component brought online. However, with an agentless solution, the IT staff can simply point to the additional servers, deploy the preconfigured monitoring template, and begin monitoring immediately.

As a result of the preceding features and benefits, agentless monitoring can lead to increased ROI over agent-based approaches by significantly reducing TCO. In this way, agentless monitoring enables IT and operations groups to spend more time building and improving production systems and less time monitoring them.

Mercury SiteScope for agentless monitoring

Mercury SiteScope is an agentless monitoring tool designed to ensure the availability and performance of distributed IT infrastructures, including servers, operating systems, network devices, network services, applications, and application components. This proactive, Web-based infrastructure monitoring solution is lightweight and highly customizable—and it does not require high-overhead agents on production systems.

Understanding how it works

Mercury SiteScope's agentless architecture is designed to give administrators a centralized view of infrastructure monitoring without the need to install agents or software on production systems. SiteScope

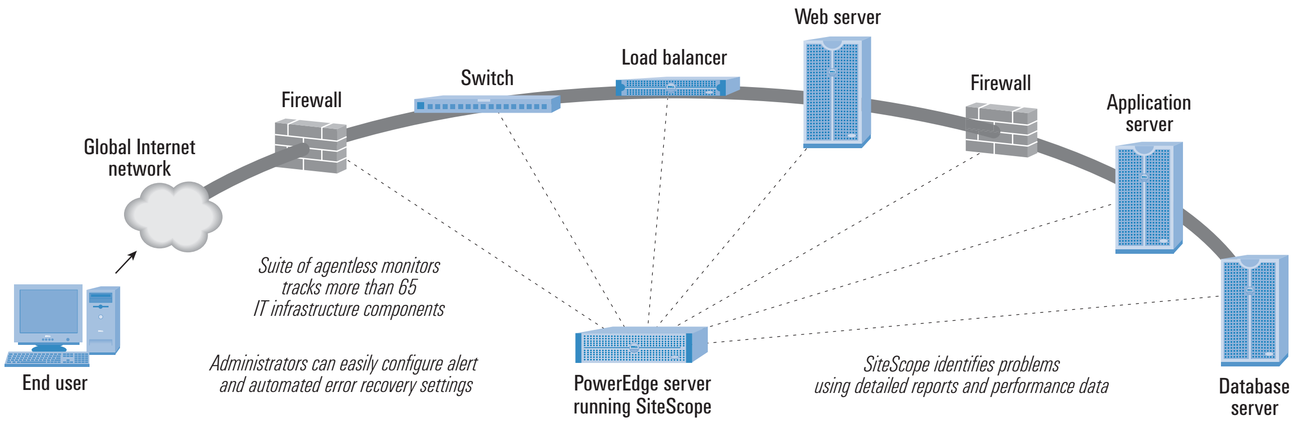

Web server
Load balancer
Switch
Firewall
Firewall
Application server
Global Internet network
Suite of agentless monitors tracks more than 65 IT infrastructure components
Administrators can easily configure alert and automated error recovery settings
End user
PowerEdge server running SiteScope
SiteScope identifies problems using detailed reports and performance data
Database server

Figure 1. A Dell PowerEdge server loaded with Mercury SiteScope monitors the IT infrastructure

is implemented as a Java server application. It runs on the server as a daemon process, monitors system parameters, sends alerts, and generates summary reports. Administrators connect to SiteScope using a Web browser, and can view status information and make configuration changes from a centralized location, as shown in the example scenario in Figure 1.

SiteScope also allows system administrators to remotely monitor multiple servers from a central installation without the need for agents on the remotely monitored systems. SiteScope accomplishes remote monitoring by logging into systems as a user from its central server, which can run on Microsoft® Windows®, UNIX®, and Linux® operating system platforms. Several login formats are supported, including Telnet, remote login (rlogin), HTTP, Secure Shell (SSH), and Network BIOS (NetBIOS).

Incorporating the IPMI specification

The Intelligent Platform Management Interface (IPMI) 1.5 specification facilitates the management of rack-mounted Internet servers and systems in remote environments over serial, modem, and local area network (LAN) connections. IPMI 1.5 also includes the ability to automatically alert IT managers of system errors and enable immediate system recovery, regardless of power state or supported communications media.

SiteScope is currently the industry's only agentless end-to-end infrastructure monitoring solution that supports IPMI. Using the IPMI 1.5 specification for hardware monitoring, SiteScope is designed to collect data from any server that is IPMI enabled, such as eighth-generation Dell PowerEdge servers. SiteScope's IPMI monitor takes advantage of the LAN interface and collects metrics remotely. This means that SiteScope can support IPMI-enabled systems without requiring agents to run on the server that is being monitored.

Agentless monitoring enables IT and operations groups to spend more time building and improving production systems and less time monitoring them.

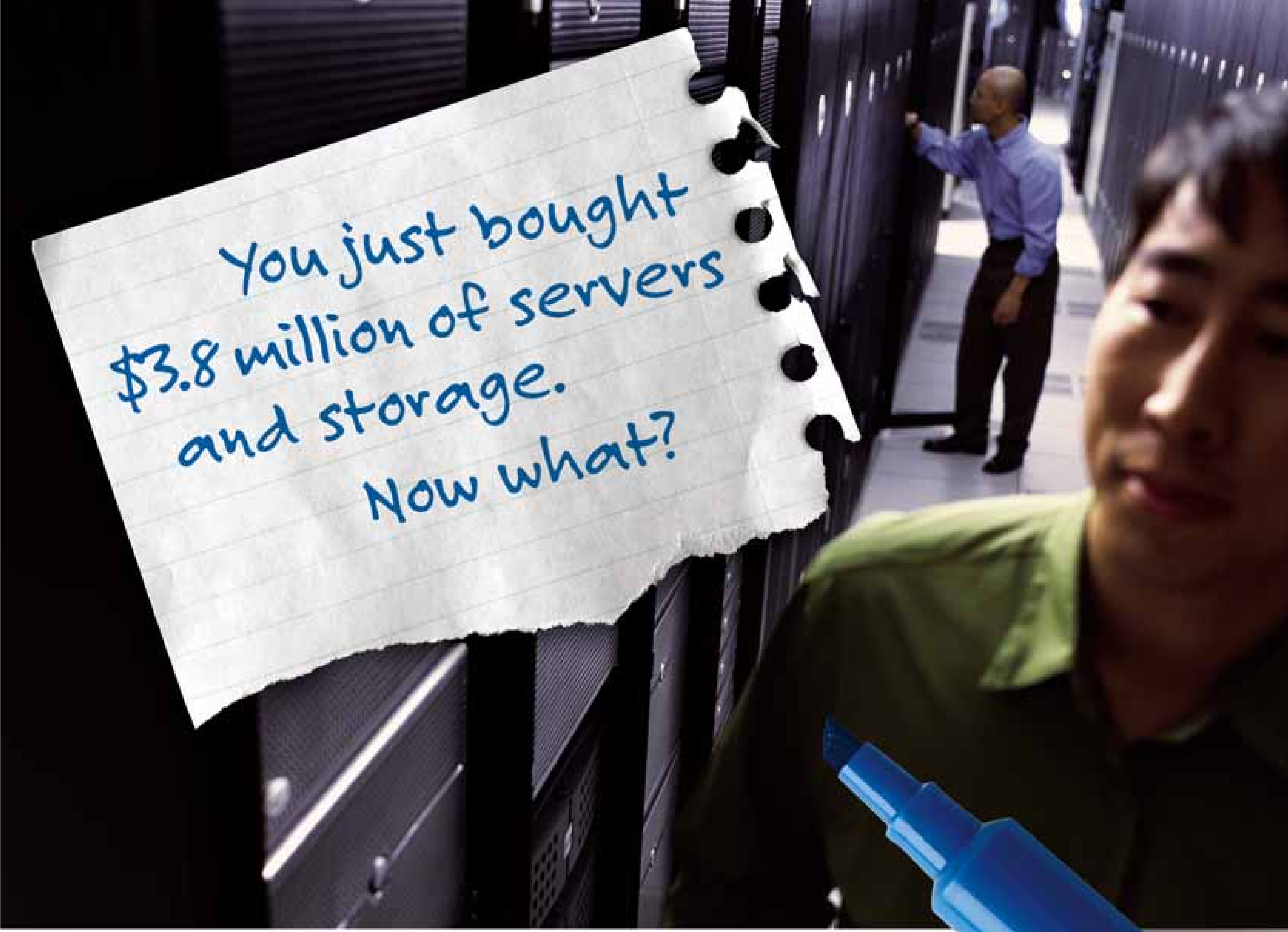
One huge benefit of the IPMI specification is that it provides a facility for monitoring critical metrics—even if the server is down or hung. This is essential for enterprise-wide monitoring because it enables SiteScope to determine whether the underlying hardware is functioning properly. In contrast, agent-based systems can become inoperable when the host server crashes.

Temperature is a critical metric for determining server health. Many IT organizations have multiple racks of servers operating in small spaces. Such dense environments can create excessive temperatures that lead to server failure. Because it supports IPMI, SiteScope has the capability to monitor temperature changes in real time and identify negative trends. This helps IT staff to understand temperature patterns and develop profiles of normal fluctuations to head off potentially devastating overheating situations.

The benefits of monitoring Dell servers with Mercury SiteScope

Mercury SiteScope monitors and helps ensure the availability and performance of a wide variety of distributed IT infrastructures, including Dell PowerEdge servers, PowerVault™ storage systems, PowerConnect™ devices, applications, network devices, infrastructure servers, and system components. SiteScope continuously monitors the performance of these components and sends alerts when potential problems are detected.

Mercury SiteScope software is available directly from Dell; small to medium-sized businesses and large enterprises can purchase SiteScope installed on a Dell PowerEdge 1850 server. Alternatively,


You just bought
$3.8 million of servers
and storage.
Now what?

organizations can buy SiteScope separately and install it on an existing server to monitor the end-to-end IT infrastructure without installing agents.

Dell PowerEdge servers configured with SiteScope can help reduce TCO by:

- Capturing an accurate and timely picture of performance data for infrastructure components
- Eliminating the need for extra memory or CPU power on production systems to run an agent
- Reducing the time and cost of maintenance by consolidating all maintenance onto one central server
- Removing the requirement to take a production system offline to update its agent
- Eliminating the time-consuming task of tuning monitor agents to coexist with other agents
- Reducing installation time by removing the need to physically visit production servers or wait for software distribution operations
- Abolishing the possibility that an unstable agent can cause system downtime and subsequent loss of business
- Providing built-in best practices that specify the most important metrics to monitor
- Integrating with other Mercury software products to enable real-time views and reports that correlate infrastructure performance across all business processes

Using Mercury SiteScope and Dell OpenManage

In addition to implementing Mercury SiteScope, for example, many enterprises have incorporated hardware monitoring solutions such as Dell OpenManage™ infrastructure. For this reason, an effective infrastructure management solution must have the capability to work cooperatively within an organization's existing environment to protect established IT investments.

Mercury SiteScope integrates with Dell OpenManage by collecting data from the Simple Network Management Protocol (SNMP). SiteScope's SNMP by MIB (management information base) monitor is designed to capture all Dell OpenManage data, including power state, device and subsystem status, and thermal information. By using other monitors, SiteScope provides the capability to collect data from the applications that are running on Dell servers and related IT infrastructure. In so doing, SiteScope can provide Dell customers with a holistic, end-to-end solution that is designed to consolidate Dell OpenManage, IPMI, application, transaction, and core infrastructure health and performance information into one easily managed console.

Agentless architectures in distributed IT environments

As enterprises move from monolithic mainframe or traditional client/server architectures to distributed environments, traditional agent-based monitoring solutions can become too inefficient and expensive to maintain. Agentless architectures can help provide a powerful and scalable solution for monitoring distributed applications and hardware.

Agentless monitoring solutions are designed to nonintrusively monitor all parts of the IT system remotely from a single host system, creating a wide variety of benefits that help lead to higher ROI.

SiteScope's incorporation of the IPMI specification provides an ideal way to monitor Dell systems including eighth-generation Dell PowerEdge servers. Whether by utilizing SiteScope alone, or by integrating it with an existing Dell OpenManage or third-party monitoring solution, SiteScope enables IT professionals to effectively monitor their systems, manage their customers' experiences, and report to clients that they are meeting promised service levels. With SiteScope, enterprises can effectively monitor, identify, and resolve problems within their IT infrastructure before their customers and critical business applications are affected.

Bill Fitzgerald is a director of business development in Mercury's Application Management Business Unit. Bill is responsible for driving the Dell and Mercury distribution partnership. He has more than 12 years of international and domestic business development and sales experience in the enterprise management, software, and services industries. Bill has a degree in international studies from Eastern Illinois University.

Bob Ure is a senior product marketing manager in Mercury's Application Management Business Unit. He is responsible for shaping SiteScope to meet customers' infrastructure performance and availability monitoring needs. Bob has more than 15 years of product management and marketing experience in the enterprise management market and has a degree in international business and an M.B.A., both from Brigham Young University.

FOR MORE INFORMATION

Agentless monitoring of Dell servers using Mercury SiteScope:
www.mercury.com/dell

Mercury SiteScope:
www.mercury.com/dell

Exploring Enhanced Technologies in the VERITAS Backup Exec Suite

As productivity and resources continue to be a major focus for organizations of all sizes, the VERITAS Backup Exec™ suite provides administrators with a single point of control for many common tasks. The release of VERITAS Backup Exec 10 *for Windows Servers* enables administrators to take advantage of technological enhancements that ease backup and recovery tasks across Microsoft® Windows® operating system environments.

BY MIKE GARCIA AND MICHAEL PARKER

Today's companies strive for excellence in performance, results, service, and efficiency. And they require this goal not only of themselves but also of the partners with whom they do business and the solutions on which they choose to run their businesses. Data protection is no exception. Given the explosion of data, organizations must protect and manage their business-critical information using a backup system they trust—one that is designed to meet their current and future needs and can deliver high performance, ease of use, and value.

For more than 10 years, VERITAS Backup Exec *for Windows Servers* has been a leading data protection solution for Microsoft Windows operating system environments, and it has been used in various organizations—from small, growing businesses to large, global enterprises. VERITAS Backup Exec 10 *for Windows Servers* is designed to meet the evolving needs of today's businesses. The Backup Exec suite (www.backupexec.com) helps deliver continuous, fast, disk-based data protection that is simple to grow and simple to manage—key attributes for the rapidly changing world of Windows environments.

Introducing VERITAS Backup Exec 10 *for Windows Servers*

Backup Exec 10 is a cost-effective backup solution that can scale from a single server to a multi-server storage area network (SAN). At the same time, it is simple enough for a novice user to install and administer, and flexible enough to protect large Windows environments.

For streamlined protection and management of Windows environments, Backup Exec 10 delivers significant features and enhancements including:

- **Backup Exec SmartLink technology:** Integrates VERITAS® Replication Exec and VERITAS Storage Exec functionality for centralized administration.
- **In-depth support for Microsoft SharePoint® Portal Server 2003:** Safeguards a SharePoint Portal Server 2001 and SharePoint Portal Server 2003 corporate knowledge base with online, fast, and reliable data protection and recovery—including protection and recovery of server farms.
- **Backup Exec Remote Agent *for Linux/UNIX Servers:*** Provides high-performance, network-wide data protection for 32-bit and 64-bit remote servers running Linux® and UNIX® operating systems.
- **Backup Exec Central Admin Server Option:** Delivers simple, centralized management of multiple Backup Exec media servers.
- **Backup Exec Advanced Disk-Based Backup Option:** Enables near-zero-impact backups and fast restores through advanced disk-based backup and recovery techniques, including synthetic and off-host backups.

- **Multi-staged disk backups:** Enforces defined retention periods before moving disk-based backups onto tape, helping support disk-to-disk-to-tape strategies.
- **Enhanced application platform support:** Integrates fast, reliable data protection support for various application platforms.

Integrating VERITAS tools with Backup Exec SmartLink technology

Backup Exec SmartLink technology enables an additional management dimension within the VERITAS Backup Exec suite protection environment. Leveraging integration with Backup Exec, VERITAS Storage Exec and VERITAS Replication Exec use Backup Exec SmartLink technology to help organizations simplify management of their Windows environments. Administrators can immediately back up or archive data from Storage Exec reports using Backup Exec, and they can view Replication Exec job status and alerts in the Backup Exec console.

VERITAS Storage Exec 5.3. Many organizations struggle to accommodate their ever-increasing demands for storage. The typical response is to increase capacity. However, this reactive approach is only a temporary solution that does not address the real issue—valuable storage resources wasted on inappropriate, nonbusiness files. VERITAS Storage Exec can help organizations proactively combat storage growth with automated data and storage management that can help reclaim large amounts of wasted storage space and control storage growth. Storage Exec generates 27 reports that provide organizations with visibility into their storage utilization, to help develop effective storage management practices for long-term cost savings and efficiency.

With VERITAS Storage Exec, administrators can establish real-time quotas on shared Windows storage; block inappropriate file types such as MP3s, games, scripts, and viruses; and identify duplicate files. Storage Exec also minimizes legal exposure by helping ensure that illegal, copyrighted, nonbusiness information does not end up on business servers. In addition, Storage Exec automatically controls Windows storage allocation, helping minimize backup windows and enhance service levels.

The VERITAS Storage Exec QuickStart edition, with support for a single server, is included in the Backup Exec suites available from Dell (www.dell.com/veritas) to enhance managing data for backup.

VERITAS Replication Exec 3.1. While most organizations invest heavily in protecting data at the main office, remote offices often receive little attention and sometimes no protection at all. This situation is often due to a lack of budget and on-site resources. Some organizations have taken steps to protect remote data but do so through ineffective and costly means, often relying on remote IT staff or non-IT personnel to perform backups at each office. Not only is this approach costly, requiring recurring investments in staff and media, but it also exposes the organization to data loss when backups are performed by untrained personnel.

VERITAS Replication Exec helps provide continuous data protection of remote offices while minimizing costs and IT workload. Replication Exec copies data from multiple remote offices over an IP connection to a central location at the main office for consolidated backups (see Figure 1). By centralizing backups and eliminating the need for backup hardware, media, and administration at remote offices, Replication Exec can help organizations achieve significant cost savings. Replication Exec is an automated technology that is flexible, easy-to-use, and capable of running entirely on its own with minimal IT effort. It integrates with Backup Exec—using Backup Exec SmartLink technology—for streamlined management, enabling administrators to monitor company-wide data protection from one console.

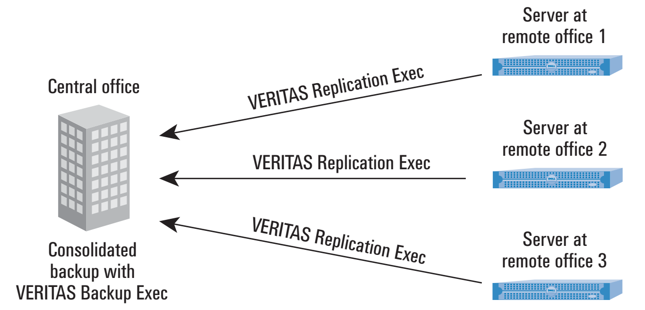

Central office
Consolidated backup with VERITAS Backup Exec
VERITAS Replication Exec
VERITAS Replication Exec
VERITAS Replication Exec
Server at remote office 1
Server at remote office 2
Server at remote office 3

Figure 1. Using VERITAS Replication Exec to back up remote offices

Centralizing management with the Central Admin Server Option

The VERITAS Backup Exec Central Admin Server Option delivers simple, centralized management of multiple Backup Exec media servers. A robust and scalable backup solution, the Central Admin Server Option lets administrators centrally manage backup and restore operations, load balancing, fault tolerance, monitoring, and reporting for several Backup Exec media servers—whether the servers are in a data center, distributed throughout the network, or at remote offices. Key benefits of the Central Admin Server Option include:

- Transformation of individually managed Backup Exec server environments into one centrally managed storage resource, which can help lower total cost of ownership for Windows platforms
- Single point of control across multiple managed media servers, which can result in minimal time and effort required to make changes across multiple servers simultaneously
- Delegation of jobs to available devices for efficient utilization of storage resources
- Centralized alerting for fast reaction times and quick problem resolution
- Automated job distribution to enable monitoring of all job activity on managed media servers

Enhancing backup efficiency with the Advanced Disk-Based Backup Option

As data amounts continue to increase while backup and restore windows grow smaller, Backup Exec 10 delivers the VERITAS Backup

Reprinted from *Dell Power Solutions,* February 2005. Copyright © 2005 Dell Inc. All rights reserved.

Exec Advanced Disk-Based Backup Option. This option offers fast, minimal-impact backups and restores through synthetic and off-host backups. Synthetic backups reduce backup windows and network bandwidth requirements by merging the initial full backup with recurring incremental backups of just the new or changed files. Additionally, synthetic backups allow for quick client restore from a single backup image—making restores of full and incremental backups from multiple tapes unnecessary. The off-host backup feature enables high backup performance and frees the remote system by processing the backup operation of the remote system on a Backup Exec media server instead of on the remote system or on a host system.

Benefits of synthetic full backups

A synthetic full backup is an operation that combines data from prior full and incremental backups to produce a data set that is indistinguishable from a full backup created at the time of the last incremental backup. Key benefits of synthetic full backups include:

- Minimized backup times by moving the creation of synthetic backups outside the time-critical backup window
- Protection of additional business-critical resources within the same backup window
- Low network bandwidth consumption during the backup window
- Fast restores because data is restored from one data set, not from full and multiple incremental backups—eliminating the recovery of redundant copies of files and the multiple steps within a nonsynthetic recovery operation

Benefits of off-host backups

An off-host backup is an operation that requires a SAN to leverage transportable snapshots to produce a near-zero-impact backup. Key benefits of off-host backups include:

- Backup load moved from the protected resource to the backup server
- Backup window effectively eliminated
- Data integrity of applications ensured prior to backup
- Minimum impact on the application server when performing off-host backups with the VERITAS Backup Exec Agent for Exchange Server or VERITAS Backup Exec Agent for Microsoft SQL Server

Exploring additional features in Backup Exec 10

VERITAS Backup Exec 10 *for Windows Servers* offers several additional features that enhance its capabilities, including Microsoft SharePoint Portal Server integration, a remote Linux/UNIX agent, and multi-staged disk backups.

Microsoft SharePoint Portal Server integration. The enhanced VERITAS Backup Exec Agent for Microsoft SharePoint Portal Server provides granularity that enables administrators to perform backups and restores of components of the portal server configuration rather than the entire SharePoint environment. Distributed SharePoint Portal Server 2003 server farm configurations are now fully supported, and restores can be made to the original SharePoint Portal Server information store or redirected to another SharePoint Portal Server information store without affecting other workspaces.

Remote Linux/UNIX agent. The VERITAS Backup Exec Remote Agent *for Linux/UNIX Servers* offers high-performance backups and restores of 32-bit and 64-bit systems, using source compression and packetizing of data to minimize network bandwidth requirements. This approach enables network administrators to perform rapid backup and restore operations on Linux and UNIX servers connected to their Windows-based Backup Exec media server over the network.

Multi-staged disk backups. Backup Exec 10 delivers a multi-staged disk backup capability. This disk staging enables administrators to automatically store single or multiple sets of data to disk for a predefined retention period and then archive to tape—to help provide off-site disaster recovery or meet data compliance requirements.

Enhanced application support. VERITAS Backup Exec 10 integrates fast, reliable data protection support for application platforms such as Microsoft Exchange and SQL Server, Lotus Domino, and Oracle® Database 10*g*. Microsoft Virtual Server 2005 and VMware® GSX Server™ 3.1 environments are supported by installing the Backup Exec media server on the host virtual server and protecting the entire virtual server. Also, users of Microsoft Operations Manager can proactively monitor and manage Backup Exec servers through the Microsoft Operations Manager Management Pack, which is currently available for download at no cost from the VERITAS Web site (www.veritas.com/Products/van?c=option&refId=321).

Simplifying backup management

Simplicity and ease of use are requirements for any organization. With the Central Admin Server Option, VERITAS Backup Exec 10 *for Windows Servers* allows administrators to manage and monitor multiple Backup Exec servers from a single console. This simple, centralized management approach helps keep administration levels and costs low, and scales easily as needs grow over time.

Backup management can be further enhanced by the Backup Exec SmartLink technology, another feature of Backup Exec 10 *for Windows Servers.* By integrating VERITAS Replication Exec and VERITAS Storage Exec with the Backup Exec suite, SmartLink technology provides a single point of administration for many tasks—allowing administrators to manage resources efficiently. These features of Backup Exec 10, along with many other enhancements, can help organizations of all sizes to meet the growing demands that data backup and recovery place on an IT infrastructure.

Mike Garcia is senior staff product manager within the Small-to-Medium Enterprise Data Protection Group at VERITAS Software.

Michael Parker is a product marketing manager in the VERITAS Data Management Group for Windows data protection solutions. He has a degree in Economics from Northwestern University.

Enhancing Backup and Recovery

Using Dell PowerVault Storage Powered by CommVault Galaxy

Simplicity is the name of the game with software designed for data protection processes. This article explores the functionality of CommVault Galaxy® Dell™ Edition backup software and describes how its features can help enhance an IT organization's backup and recovery strategies.

BY JOE POLLOCK AND CASEY BURNS

Data management is not only a buzzword for IT managers—it is becoming a serious concern. No longer can companies rely on ad hoc storage solutions and piecemeal software. IT organizations are striving to build infrastructures that use consistent hardware and software.

CommVault Systems recently teamed with Dell to develop CommVault Galaxy and GalaxyExpress, two suites of data protection software that provide easy-to-use basic network backup and restore capabilities. GalaxyExpress can help address the needs of small to medium businesses and branch offices. It lets organizations perform backup and restore functions using features normally associated with higher-priced storage software products.

Galaxy is a fully featured software package that offers advanced features and client software aimed at driving down the backup costs of medium to large clusters. Dell commissioned CommVault to tune GalaxyExpress for the Dell PowerVault™ network attached storage (NAS) system and higher-end Dell/EMC storage area network (SAN) equipment. The combined offering from Dell and CommVault delivers a cost-effective and powerful IP and Fibre Channel storage system for small to medium businesses. This article focuses on the NAS solutions.

The Galaxy products simplify management of data in storage networks by offering a single, unified view based on either a logical application or physical system perspective. Managing critical data using this unified view enables the Galaxy solution to take traditional backup and recovery to a new level regarding ease of use, configuration, reliability, scalability, and flexible deployment.

In conjunction with additional product offerings including data migration, compliance archiving, and snapshot management systems, CommVault provides a complete set of data and storage management tools specifically designed to help provide total data protection. The following sections explain the Galaxy approach and briefly describe some of the core benefits that Galaxy software can offer. Both Galaxy products were designed to provide ease of use, ease of management, and advanced technology to help organizations manage their data protection requirements.

Although data is generally backed up for the purpose of being recoverable later, some backup software products focus only on the backup process itself instead of actual data recovery. To simplify the recovery process, Dell and CommVault teamed up to create Dell editions of four-time industry award–winner CommVault Galaxy backup and recovery software. The result is an easy-to-use data protection solution that is designed to help organizations restore data and resume operations quickly and efficiently.

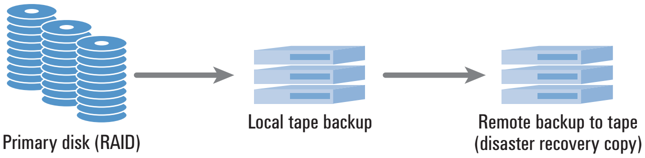


Figure 1. Traditional backup to tape

Limitations of traditional backup methods

Traditionally, data has been stored on disks and backed up to magnetic tape cartridges (see Figure 1). This practice of backing up data to tape for both local and remote disaster recovery storage has been implemented and optimized during the last 15 years.

As data growth has increased through the years, tape manufacturers have responded by providing enhanced cartridge capacity and faster throughput. However, in the last few years, data volume has grown exponentially and tapes have been unable to provide enough capacity and speed to enable organizations to protect their data within shrinking backup windows.

Advantages of NAS for LAN-based backups and restores

Because the file-serving and storage requirements on the local area network (LAN) are already directed to the NAS server, bundling backup software into NAS is a simple and highly effective step that helps enable centralized backups and restores. Using NAS as a backup server provides the following additional advantages:

- **Consolidation of backup equipment:** When storage is attached to the network instead of directly attached to the server, it is no longer necessary to attach a tape device to each server to enable backups. Tape equipment can be consolidated directly onto the NAS server. This allows businesses to invest in a limited number of high-quality tape devices and to maintain these tape devices in temperature- and humidity-controlled environments.
- **Streamlined backup management:** Using the NAS server to control the backup and restore processes helps simplify management by centralizing backup operations. System administrators are no longer required to go to individual servers to execute backups. Using a Web-based interface, administrators can schedule the NAS device to back up all servers to tape so that no write conflicts arise between servers.
- **Client system backups:** Notebooks are not typically configured for tape backups, and rarely do users consistently back up their desktop PCs or notebooks. NAS backup servers provide a means to automate the backup process for these client systems.
- **Disk-to-disk backups:** Disk-to-disk backups are enabled by backing up client data to the NAS server (see Figure 2). Disk-to-disk backups are generally faster than disk-to-tape backups, thus helping reduce the backup window. When point-in-time software capabilities are used, backups can be accomplished in seconds. For backed-up data that must be frequently accessed, disk-to-disk restore operations can be much less time-consuming than restoration from tape.
- **Effective management of backup windows:** All backups on the LAN can be scheduled and controlled through the NAS server. Data can be backed up from the NAS disk to tape during times when network traffic is minimal.

Advantages of disk-to-disk backups

The advent of low-cost disk solutions has created a category of secondary storage known as disk-to-disk storage. Using disk versus tape as secondary storage can offer the following advantages:

- Data can be restored quickly.
- Disks are generally faster than tapes, especially when mount and seek times for tapes are factored in.
- Tape failure issues can be eliminated.
- Disks are random access and are optimized for lookups, while tapes are sequential and can be much slower for random file recalls.
- Multiple hosts can access a disk simultaneously because disks have multiple read/write heads.
- Using disks can eliminate human errors in tape handling.

Disk-to-disk backups and disk-to-disk appliances

Organizations are faced with a problem. Not backing up data—or backing up only as much as possible in shrinking backup windows—is not a viable option, especially considering the ever-increasing government regulations regarding maintaining copies of data for future compliance requirements. To meet the need for faster and higher-capacity backups, Dell offers the PowerVault 745N system. This NAS server can offer fast access and throughput as compared to tape. When used in conjunction with the Dell PowerVault 122T, PowerVault 132T, and PowerVault 136T tape libraries, the PowerVault 745N can provide an

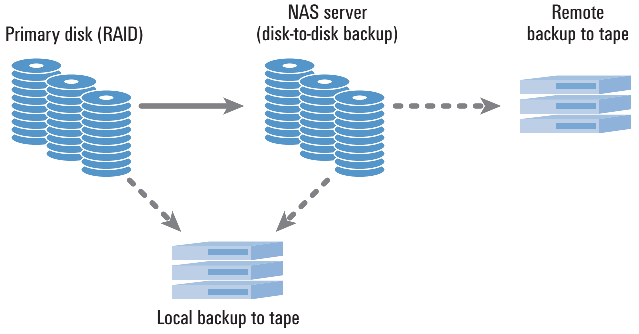


Figure 2. Disk-to-disk solutions

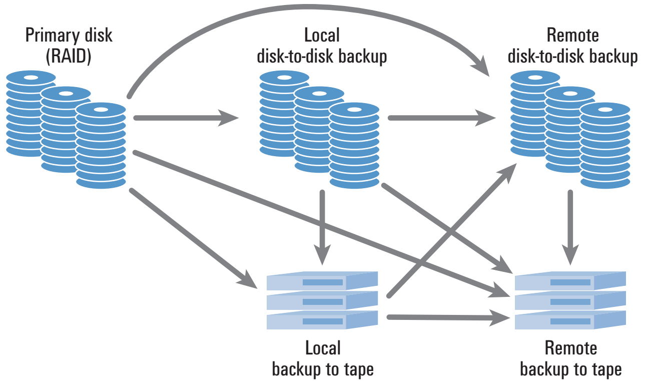


Figure 3. Policy-based management of backups

excellent means of centralizing backups to simplify operations and enhance utilization.

CommVault works with Dell to deliver a simple yet powerful software solution that has the capability to make copies of data from the disk-to-disk device to a local or remote tape device. The Galaxy software offers comprehensive media management, retention and tracking, and scalable data indexing that supports single-file, single-step recovery from disk or tape copies with one click.

Unlike some backup and recovery software packages, Galaxy software takes full advantage of the properties of disk. Galaxy has the capability to write to disk as a random access medium rather than write to disk as if it were a tape (that is, in a single-threaded, sequential data stream). Unlike tape, disk has many read/write heads, which allows multiple jobs and multiple read/write commands to be executed simultaneously. Using Galaxy software, organizations can copy data from one networked storage device to any other networked storage device at any location automatically using policy management (see Figure 3).

A unified approach to storage

One of the chief goals of the CommVault and Dell relationship is to simplify data management. Utilizing a unified, centralized approach to storage versus an ad hoc solution can help free valuable IT management resources. CommVault Galaxy and GalaxyExpress software in combination with a Dell PowerVault NAS system can help streamline backup and recovery implementation and operation. In this way, Dell helps provide enterprises with cost-effective and easy-to-use hardware and software solutions for their backup and data management needs.

Joe Pollock is a software marketing manager in the Dell Storage Product Group. He has a B.S. in Electrical Engineering from the University of Florida and has more than 11 years of technical marketing and sales experience in the storage industry.

Casey Burns is a program manager for CommVault Systems.

Branch Office Data Consolidation:

Helping to Reduce Costs and Increase Data Protection

Protecting enterprise information in geographically distributed branch locations can be challenging for IT organizations. This article describes how backup and replication tools from Dell and LEGATO can be used effectively to consolidate data in a central location—thereby helping to reduce costs, increase data protection, and improve operational efficiencies.

BY JOE POLLOCK

The traditional approach to protecting Microsoft® Exchange data and file data in geographically distributed branch locations is to deploy backup software and tape hardware in each location. This requires organizations either to hire the necessary technical staff or assign inexperienced employees to administer daily backup operations in remote locations. In the latter case, backup and recovery operations are decentralized, operating with little or no assistance from trained technical administrators in the regional data center.

This approach to data protection can be both risky and costly. It requires an ongoing investment in tape devices and other hardware, backup software, and support services for each remote office. Additional IT staff may be required to help support those daily backup and recovery operations, but there is no assurance that recovery will be possible because an untrained staff often performs backups inadequately or skips backups entirely—and all too often crucial backup tapes are not handled and tracked carefully. Unfortunately, these critical issues are typically discovered at the worst possible time—during an attempted recovery. In some instances, recovery may require flying an IT expert to the remote location, which adds time and expense to the recovery process.

A centralized approach to branch office backups

A more reliable way to protect Exchange data and other data in remote locations can be to administer backups centrally from a regional data center rather than delegating the task to inexperienced branch office staff. Backup software can enable organizations to implement remote backups from the branch office to the regional data center. However, this method can incur significant performance penalties when backing up data from the remote location to the central location because of distance, handshaking, and so forth.

An optimal approach to centralized backup of remote branch office data is to replicate data from remote sites to the regional data center. For example, LEGATO® RepliStor® software is designed to provide real-time, asynchronous movement of data, including Exchange data, which can be stored on Just a Bunch of Disks (JBOD) or on Microsoft Windows® Storage Server 2003–based network attached storage (NAS) such as Dell™ PowerVault™ devices at the branch office, at the regional data center, or at both locations.

Another step in the centralized backup and business continuance process is the regular backup of branch office data at the regional data center from a Windows Storage Server 2003 server to tape devices using backup software such as LEGATO NetWorker® software. Replicating data from the remote office to the regional data center for

Reprinted from *Dell Power Solutions,* February 2005. Copyright © 2005 Dell Inc. All rights reserved.

centralized backup can offer significant benefits. Because data replication allows organizations to perform the backup to tape only in the regional data center, it can eliminate the need to deploy tape hardware and backup software in the remote locations that are being backed up—thereby helping to reduce the costs associated with implementing a remote backup solution. Thus, data replication enables organizations to better utilize tape hardware, consolidating that investment in the data center. This approach enables tape backups to be performed from a replicated copy in the regional data center, and branch office backups can be part of the data center backup schedule. As a result, the backup window can be much larger and more flexible than it is when using backup software to copy branch office data directly to tape.

This centralized approach can help reduce or eliminate the need for trained backup administrators located at or supporting each remote office. Moreover, it helps keep branch office personnel focused on business—not backups. Because backups are conducted by trained administrators at the data center who adhere to correct data protection and tape-handling methods that follow set IT policies, the risk of data loss is low. In addition, when copies of the branch office Exchange data are stored off-site at the data center, this data can be protected from disasters such as fire, flood, or theft that may occur at the remote location.

Using LEGATO RepliStor for data replication can provide additional benefits. Because it transfers only the file operations for changes made to specific files, RepliStor helps minimize network traffic. Data replication enables data transfers to be easily and automatically restarted if a network failure occurs, so data transfers can be restarted from the point of the failure rather than from the beginning of a data set—a logical starting point for backup software. RepliStor is designed to automatically and continuously replicate data as the data is created, sending that data from the remote location to one or more secondary disk storage systems, local or remote. A second copy of production data can be created at a remote location, and this copy can always be up-to-date with the data in the regional site. This approach helps eliminate the backup window for the remote location and can avoid incurring backup-related CPU drain on the Exchange servers at the branch office.

> An optimal approach to centralized backup of remote branch office data is to replicate data from remote sites to the regional data center.

Besides replicating data, RepliStor is designed to constantly monitor Windows file-and-print services. Should a failure occur, RepliStor is designed to automatically recover those services on a second Windows server—enabling business users to continue working while administrators find and fix the problem. As soon as the failure occurs, RepliStor sends an alert notifying administrators of the problem. However, because RepliStor is designed to provide automated recovery, the on-call administrator does not have to immediately address the problem. Administrators can continue with their current tasks, and when time permits, they can find and fix the problem and test whether the system is working properly—all while business continues uninterrupted.

Replication with LEGATO RepliStor

LEGATO RepliStor is designed to provide real-time, asynchronous data replication, which helps enable high availability and high reliability of Windows servers and Windows-based NAS without the use of proprietary or specialized hardware. LEGATO RepliStor is designed to work with several popular data storage devices, including Dell/EMC CX series storage arrays as well as Windows Storage Server 2003–based NAS, including Dell PowerVault NAS servers.

RepliStor provides one-to-one, one-to-many, and many-to-one bidirectional replication of data from one server or NAS device to another, over virtually any distance. When one server or NAS device becomes unavailable, RepliStor is designed to automatically switch file system services to the second NAS, preserving the end users' capability to access and use the data. LEGATO RepliStor can be deployed in mixed environments, replicating data among Windows Storage Server 2003 devices and server systems running Windows 2000 Server and Windows Server™ 2003. This replication can be implemented over both local area network (LAN) and wide area network (WAN) connections, and offers both remote installation and administration.

At its core, LEGATO RepliStor provides a replication solution that is designed to copy files, directories, volumes, shares, and registry keys from a source system (where data needs to be protected) to a target system or systems (where the data is then replicated). After the initial synchronization of the data, RepliStor mirrors the file operations only for changes made to specific files, minimizing network traffic. It then replays those file operations on the target system to help ensure that both systems are in sync. Compression technology plays an important role in the efficiency of the replication as well, through reduced network overhead as compared to uncompressed backup software transfers. Because RepliStor data updates are designed to be continuously and automatically replicated as a background process between source and target systems, organizations can maintain multiple online copies of data without incurring additional administrative overhead. RepliStor deployment and administration is facilitated by a powerful graphical user interface (GUI) that simplifies setup and implementation.

Advanced features of granular, rules-based replication—such as filtering and scheduling—allow flexibility in the selection of what information RepliStor replicates and when the replication process

Server Virtualization

Let Dell show you the way.

Dell™ hardware and VMware® Virtual Infrastructure™ software can help your enterprise scale out cost-effectively.

No matter how big your business or what direction it may take, modular Dell servers and shared storage—together with VMware virtualization software—are designed to provide a flexible, pay-as-you-grow approach to enterprise virtualization.

Fast business response.

A virtualized IT infrastructure on multiple Dell PowerEdge™ servers running VMware ESX Server™ software can enable enterprises to migrate workloads across multiple physical servers on the fly—helping to improve operational flexibility while lessening the potential impact of hardware failures.*

DELL®

vmware®

*For more information, see "Introducing VMware ESX Server, VirtualCenter, and VMotion on Dell PowerEdge Servers" by Dave Jaffe, Ph.D.; Todd Muirhead; and Felipe Payet in *Dell Power Solutions,* March 2004.

occurs. Replication can be set to run automatically or replication and synchronization can be scheduled for a particular time or day. Administrators can also dedicate a network interface card (NIC) to RepliStor traffic, making it easy to isolate replication from the rest of the network. This approach may also help administrators to better manage the impact of replication on network bandwidth.

Backup and recovery with LEGATO NetWorker

LEGATO NetWorker protects critical business data—simplifying, centralizing, and automating backup and recovery operations across UNIX®, Windows, Linux®, and Novell® NetWare® platforms. Built on an open, highly scalable architecture, NetWorker is designed to enable organizations to standardize on one application to provide complete, fast, and reliable data protection in both large data centers and small branch offices. Reliable backup and restore processes can help minimize downtime costs and management overhead, thereby contributing to lower total cost of ownership (TCO) for storage resources.

LEGATO NetWorker helps deliver data protection capabilities and management consistency for enterprises with heterogeneous direct attach storage (DAS), NAS, and storage area network (SAN) environments. Advanced indexing, high-speed parallelism, automated media management, LAN-free and serverless backup, cluster awareness, and dynamic tape drive sharing are among the valuable features that NetWorker offers to enable administrators to effectively protect storage assets and minimize downtime.

A NetWorker data protection solution can also include application modules, which are designed to deliver online hot backups and granular recovery for Oracle®, Microsoft SQL Server, Microsoft Exchange, and SAP® applications, among others. NetWorker SnapImage™ and PowerSnap™ modules, which can provide image and instant snapshot backups, can further free applications from the impact of data protection operations and help accelerate restores. These modules are designed to be fully integrated with vendor application programming interfaces (APIs) to help achieve 24/7 availability levels and ensure complete, reliable protection. Tape autoloader and library modules can enable hands-free data protection for a wide variety of tape and robotic devices. In addition, the NetWorker DiskBackup Option (DBO) allows data to be backed up to disk, staged on disk, and automatically moved to tape, or cloned from disk to tape or disk to disk—all with single-step (one-click) recovery. A 1 TB license for DBO is included with the Dell edition of LEGATO NetWorker.

NetWorker is designed to enable organizations to standardize on one application to provide complete, fast, and reliable data protection.

NetWorker includes features that are designed to take advantage of Windows Storage Server 2003, Windows 2000 Server, and Windows Server 2003 functionality, helping to provide backup and recovery scenarios that are robust, scalable, and flexible. Centralized management, structured metadata, and open tape format (OTF) compatibility enable NetWorker to specifically address and facilitate fast data backup and recovery.

NetWorker is designed to support the Removable Storage feature of Windows Storage Server 2003, Windows 2000 Server, and Windows Server 2003 without requiring additional Microsoft or LEGATO software. The Removable Storage service acts as an agent to either the NetWorker server or storage node software to perform the actual mounting, labeling, and tracking of media. The Removable Storage service controls storage hardware's doors, drives, and robotics, and performs uniform drive-cleaning operations. NetWorker performs volume and file management, which is not provided by the Removable Storage service. The use of Removable Storage is optional with NetWorker, allowing multiple data management applications to share access to the same storage hardware.

NetWorker is designed to use the Windows Server 2003 Volume Shadow Copy Service (VSS) to provide backup functionality for active applications and services. In this process, all applications and services that are running during the backup operation are frozen for a brief time to provide a point-in-time snapshot, or *shadow copy,* of a data volume. NetWorker backs up the files on the shadow volume, and then alerts VSS that the snapshot can be released.

Backup and replication: A powerful combination

The combination of Microsoft Exchange 2003, Windows Storage Server 2003, and LEGATO RepliStor and NetWorker software can provide an optimal Exchange consolidation platform and branch office data protection strategy to help create business efficiencies and reduce costs for both branch offices and the data center. When Exchange data from a remote office is consolidated and replicated to a central location, it can be backed up by trained personnel while leveraging equipment and software that have already been established in the data center.

In addition, implementing data protection in a branch office enables information to be replicated as it is created. The approach described in this article enables a second copy of the information to be readily available in the data center for disaster recovery. As a result, trained data center administrators can help provide rapid business resumption if a remote location goes down, because data can be immediately available. In this way, an Exchange consolidation platform and data protection strategy can help ensure reliable branch office backups, quick recovery of data, business continuity, and low equipment and administration costs in branch offices while helping to increase branch office and data center productivity.

Joe Pollock is a software marketing manager in the Dell Storage Product Group. He has a B.S. in Electrical Engineering from the University of Florida and has more than 11 years of technical marketing and sales experience in the storage industry.

Efficient Bare-Metal Disaster Recovery from Yosemite Technologies

Traditional backup and recovery software is often the last line of defense for business-critical information when an organization experiences a significant data loss event. By deploying Dell™ PowerEdge™ servers or Dell PowerVault™ tape libraries equipped with Yosemite Technologies® TapeWare® 7.0 software and the Bare Metal Disaster Recovery Agent before a disaster occurs, organizations can streamline the recovery process and get back to business sooner.

BY ERIC HARLESS

When IT organizations consider data recovery, the subject may encompass file recovery, database recovery, or server recovery. While restoring or recovering an individual file can be a relatively simple task, the complete recovery of a server and its applications and data can be much more challenging. Administrators must determine a course of action in case a server's disk volumes, operating system (OS), or both are unavailable. That is precisely the type of scenario that differentiates the bare-metal disaster recovery approach from standard data recovery methods, which assume the disk volumes and the OS are still operational.

While natural disasters invoke visions of fires, floods, and hurricanes, data disasters are more likely to be caused by hardware failure—or by system-wide data corruption that started with a virus or user error. In such situations, administrators often need to start from scratch.

To a corporate executive, database administrator, or IT manager, starting from scratch rarely sounds like a good idea, especially when doing so involves business-critical data from Web, database, or application servers. In response to such concerns, Yosemite Technologies developed the Yosemite TapeWare Bare Metal Disaster Recovery Agent, which is designed to simplify and automate the bare-metal recovery process of a corrupt or failed server. Using this optional component of the TapeWare application, an administrator can quickly reboot a failed system, restore the OS as it was previously configured, and restore the server's applications and data.

Reasons to consider disaster recovery solutions

Cables become unplugged, electronic devices fail, disks stop spinning, batteries run down, viruses propagate, and—regardless of defined policies or procedures—data records are sometimes discarded or overwritten. Even with the introduction of specialized hardware and fault-tolerant solutions for clustering and replication, information can and will continue to be lost.

The ongoing viability of an organization that has suffered from a significant system or data loss is not determined strictly by that organization's capability to replace hardware or rebuild infrastructure. In many cases, continued success relies heavily on the capability to quickly and successfully recover business-critical data. Given its importance for organizational survival, disaster recovery is a data protection choice that should be given great consideration up front—before disasters strike.

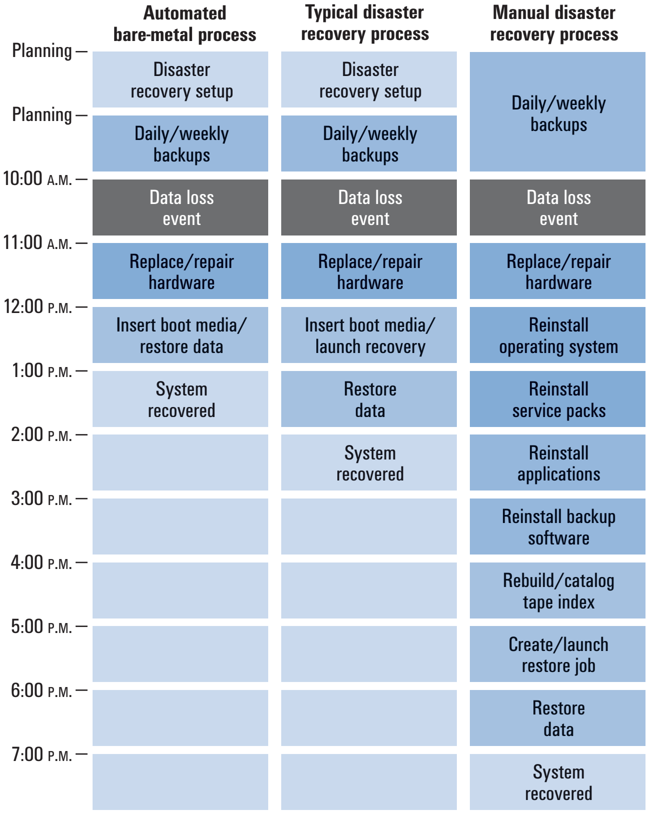
Backup concepts, while firmly rooted in the IT mind-set, continue to evolve as technology advances. In the process, backups take on new meaning and organizations develop innovative approaches toward implementing such procedures. Although sometimes delivered as a stand-alone solution, disaster recovery is more often considered to be a component of a backup and recovery solution. Put simply, disaster recovery can be defined as the capability to quickly and gracefully recover from total data loss. However, this definition has become blurred in the midst of the many cluster, replication, and other availability products that are being developed and promoted as disaster recovery solutions.

Manual recovery processes: Time-consuming and inefficient

When compared to the reality of data loss, even a manual recovery process is better than no recovery at all. According to recent polls conducted by Yosemite,[1] up to one-quarter of the midsize businesses polled plan to use manual recovery as their approach for disaster recovery. However, manually recovering a failed system can prove rather cumbersome and time-consuming for even the most seasoned IT professionals.

Whether the system failure and resulting data loss is on an e-commerce server, an accounting system devoted to payroll, or a file server full of architectural drawings, the first task after a server failure is to isolate the problem that caused the failure and take steps to correct it. Doing so can be as simple as identifying and replacing a defective SCSI controller and hard drive, or as difficult as finding a replacement for an obsolete motherboard. Administrators must then configure hard drive partitions and any special RAID sets that are needed, and once the basic hardware problems are resolved, they must locate the relevant installation CDs and license activation keys. Next, administrators must reinstall the OS. Previous system information such as network addresses, directory structures, volume sizes, and cluster data may be needed to complete the installation. In a best-case scenario, the server that is being recovered will have Internet access, because the OS and any special hardware configuration may require device drivers, patches, and service packs to restore access to the system's peripherals.

Once the base OS is up and running, administrators need to locate, install, and configure the server's applications and backup software. Typically, after hours of manual processes, administrators can finally insert a tape and start rebuilding the catalog so that they can begin selecting files to restore. Depending on the server's configuration, this catalog rebuild and restore process can consume even more time than the hours already spent on recovery. See Figure 1 for an example of efficient recovery using an automated bare-metal process versus a typical disaster recovery process.



	Automated bare-metal process	Typical disaster recovery process	Manual disaster recovery process
Planning	Disaster recovery setup	Disaster recovery setup	Daily/weekly backups
Planning	Daily/weekly backups	Daily/weekly backups	
10:00 A.M.	Data loss event	Data loss event	Data loss event
11:00 A.M.	Replace/repair hardware	Replace/repair hardware	Replace/repair hardware
12:00 P.M.	Insert boot media/ restore data	Insert boot media/ launch recovery	Reinstall operating system
1:00 P.M.	System recovered	Restore data	Reinstall service packs
2:00 P.M.		System recovered	Reinstall applications
3:00 P.M.			Reinstall backup software
4:00 P.M.			Rebuild/catalog tape index
5:00 P.M.			Create/launch restore job
6:00 P.M.			Restore data
7:00 P.M.			System recovered

Note: Recovery time examples represent system tests performed in Microsoft® Windows Server™ 2003 environments. Actual recovery time will vary based on hardware configuration and data volume.

Figure 1. Recovery using an automated bare-metal process versus a typical disaster recovery process

Automated bare-metal disaster recovery

The benefit of bare-metal disaster recovery derives from the capability to automatically re-create hard drive partitions and perform a full system recovery of the OS, applications, and data. This capability alone has the potential to shorten a typical manual recovery process by several hours. Two steps are required to prepare for the disaster recovery process:

- Make a full backup of the system exactly as it should be restored in the event of a disaster.
- Create the appropriate boot media.

The full backup is used with appropriate boot media—either bootable floppy disks, a bootable CD image, or a bootable tape device—to perform a complete restoration. Disaster recovery products are designed to be as automated as possible during both preparation and recovery phases, so that once installed, a disaster recovery product should be able to perform its tasks with minimal hands-on intervention.

[1] In September to December 2004, Yosemite conducted polls of midsize companies that were not current Yosemite customers.

For optimal protection, administrators should perform full backups either as part of a regularly scheduled backup plan or as a snapshot that is performed off-schedule. Additionally, administrators should perform a full backup each time a significant change in the system data occurs. Administrators should also create a new bootable disk set or CD any time they change a system's hardware or OS.

Of course, disaster recovery solutions are only as effective as the media rotation schedule that is put in place. If tapes are not rotated regularly and stored in secure locations, then critical data is still at risk, and in such a circumstance no disaster recovery solution will be effective.

Examining product differences

Some disaster recovery products have highly specialized functionality, and for performance or security reasons, they may require that each protected system have an attached tape device. When this is the case, organizations may need to purchase one disaster recovery product license per protected system. Alternatively, products may allow network-based recovery of a remote system using backup data archived on disk instead of on tape. However, complexities may exist when enabling network connectivity on a bare-metal system. For that reason, disk-based network disaster recovery solutions do not typically offer the same level of reliability, portability, or scalability that tape-based disaster recovery solutions can provide.

Because all operating systems do not fully support plug-and-play technology, disaster recovery operations should always be performed on the same computer system after replacing the faulty hardware that caused the system failure. Most disaster recovery solutions assume that no major changes to the hardware have occurred; the hardware to which data is restored must be virtually identical to the source system.

Administrators should be cautious of so-called disaster recovery solutions that do not fully restore the base OS. These products may attempt a scripted reinstallation of the OS and then restore just the critical data. Such methods usually incur slow restore times, may require manual intervention, and have a tendency to break down on systems using advanced hardware that requires additional drivers, service packs, or hardware not originally supported out of the box by the OS.

Cloning solutions, another option for disaster recovery, are targeted toward the desktop OS market. These products effectively allow the creation of a point-in-time image, or *snapshot,* of a system that can be stored on a hard drive or network volume. These products are traditionally used to clone an OS as a method of deploying a standard desktop image onto multiple systems. While this technique allows for quick recovery of a stock system, it is not feasible for daily data protection tasks or large application servers.

Selecting a robust product

Because hardware and device support vary by platform and OS, a viable disaster recovery product must be robust enough to offer multiple recovery methods that may include bootable floppy disks, a bootable CD image, or a bootable tape device. A robust product should also provide support for all leading tape device manufacturers to offer maximum flexibility during recovery.

In addition, disaster recovery should not be limited to server platforms; a disaster recovery product must also protect desktop PCs and workgroup environments. Usually, critical data does not reside entirely on enterprise file and application servers. Rather, data is distributed across the hard drives of desktop and notebook computers used daily by employees and executives. While most server-based backup products can back up desktop clients remotely, few offer the combination of affordable disaster recovery, local tape device support, common user interface, intelligent wizards, and robust features needed to fully protect desktops and workstations.

Preparation for disaster recovery

When a computer fails, recovery time is crucial. Loading and configuring an OS and reinstalling software can be very time-consuming for IT staff, and lack of access to important data curtails productivity for other employees. When disaster recovery is properly executed, organizations can quickly and easily achieve full restoration of a system's OS, hard drive partitions, applications, and data.

The key to painless disaster recovery is having a reliable backup in place and a disaster recovery product that enables an organization to recover even from a large-scale loss of data. The Yosemite TapeWare Bare Metal Disaster Recovery Agent takes data restoration to a high level by providing a comprehensive, easy-to-use solution that works across multiple platforms and operating systems. This approach is designed to automate and streamline both the restore aspect of disaster recovery and the creation of disaster recovery media—thereby helping ensure that these critical tasks are not overlooked. By saving IT administrators the hassle and complexities of learning a different recovery strategy for each platform and OS deployed throughout the network, Yosemite helps administrators to be more productive and better focused on data management.

Eric Harless is a product line manager at Yosemite Technologies.

FOR MORE INFORMATION

Yosemite backup and disaster recovery:
www.yosemitetech.com

File Systems for the Scalable Enterprise

Cluster file systems have the potential to improve cluster performance and expandability dramatically, offering innovative approaches for increasing bandwidth to accommodate large amounts of data, balancing file server loads, and simplifying data access for client systems. This article discusses a scalable approach to file system architecture and provides criteria that can be used to evaluate file system capabilities.

BY DAVID WEBER, ROGER GOFF, AND SURI BRAHMAROUTU

A key challenge facing architects of high-performance computing (HPC) systems is how to transfer large amounts of data in and out of clusters. Furthermore, the amount of data that needs to be processed is growing, and traditional file servers are not always capable of scaling effectively to meet that demand. Enterprise applications, databases, and Web hosting services all face similar data movement challenges. Cluster file systems have the potential to address performance and expandability challenges. Emerging file system architectures can help deliver scalable throughput by aggregating the output bandwidth of multiple servers and data arrays.

For this discussion, it is important to clarify the distinction between a file server and a file system. A *file server* is a hardware device that shares data over a network with several client systems. A file server can comprise both network attached storage (NAS) devices and general-purpose servers that host data. In contrast, a *file system* is software that provides a means to store and access data. A file system can be local to a computer, like the NT file system (NTFS) or the Linux® ext3 file system, or it can be available over a network, like Common Internet File System (CIFS) or Network File System (NFS). Accessing a networked file system is accomplished by connecting to the networked data using a program running on every client system; the program communicates with the networked file system server through a defined protocol.

File-server performance is affected by the number and type of I/O ports that the server hardware supports as well as CPU, memory, and number of storage spindles. While the number of I/O ports available in some servers is large, each server is designed with an inherent physical limit that defines the maximum throughput. Several problems can occur as a result of file-server throughput limitations:

- **Throughput requirements of the client systems can exceed the maximum throughput capability of the file server:** Using traditional file-server technology, the I/O demands of the client systems can be met only by adding file servers and dividing the data among multiple file servers. As the number of clients grows, the number of file servers also grows. Client systems end up needing multiple mount points to access the required data.
- **Client system demand for access to files in a large directory of hundreds or thousands of files can exceed the bandwidth of a file server:** In such a scenario, administrators may be able to add file servers and replicate or distribute the data among those

file servers to solve the bandwidth problem. However, doing so can lead to file lock and mount-point management issues. In addition, storage is not used efficiently when the data is replicated. Replicating also introduces complexities associated with keeping the replicated data coherent.

- **Client system demand to access a particular file, such as a large DNA database, can exceed the bandwidth of a file server:** Replicating the file across multiple file servers increases aggregate bandwidth to client systems, but this can create synchronization or locking problems when clients need to perform writes to the files in addition to reads.

An additional limitation of traditional file servers is their inability to load balance file-serving tasks across multiple file servers. Uneven load distribution is not ideal and can lead to hot spots where one or more file servers operates at maximum capacity while others remain virtually idle. Administrators are challenged to optimize resource utilization by evenly distributing the load of the file servers.

Ideally, the preceding problems could be resolved simply by adding the necessary hardware—either file servers or data arrays—to meet client system demands without rebooting or making any modifications to the client systems' mount points. In environments where administrators are limited to traditional capabilities of file servers and file systems, an increase in data throughput often cannot be achieved in a manner transparent to client systems. However, scalable file systems are now emerging—creating software-based technologies capable of growing to meet client system demands, and helping deliver required data performance without necessitating changes to client systems.

At minimum, data environments need to be reliable, accessible, and supportable. In addition, the ideal data environment should have the capability to scale in size and performance to meet the needs of thousands of client systems, without being limited by the performance of any single file server or data array. The ideal environment should also expand as the number of clients and the size of data grows, without disrupting ongoing computation on existing client systems.

How scalable file systems work

File systems can be scaled in a number of ways. These methods can be broadly classified into two categories: NAS aggregators and cluster file systems.

NAS aggregators. This approach unites multiple file servers or NAS appliances so they act as a single file-serving device. The aggregator can be a front-end platform or a distributed file system that runs on a group of appliances. Individual appliances in the group may include their own disks, or they may be NAS heads or gateways that access block storage on a storage area network (SAN).

NAS aggregators can be further classified into two groups based on the approaches they take in accessing data and metadata. In-band NAS aggregators intercept every I/O request from each client, repackage the requests if necessary, and invoke one or more back-end NAS appliances. The responses from the various appliances are then assembled if necessary and sent to the client.

In contrast, out-of-band NAS aggregators do not intercept requests in the I/O path. As a result, clients need to consult an out-of-band NAS aggregator to obtain metadata information first. Administrators must install on every client a local agent that can interpret the metadata information returned by the NAS aggregator. Clients, with the help of the local agent, then generate I/O requests directly to the appropriate NAS appliance to obtain the data.

NAS aggregators offer a cost-effective way to consolidate departmental file servers, providing a single global namespace, which helps enable centralized management. However, the NAS aggregator approach can introduce latencies caused by additional Remote Procedure Calls (RPCs).

Cluster file systems. In this approach, multiple file servers in a cluster share access to the same data through the framework of a cluster file system. The goal is to allow multiple servers to act as a single, dynamic, highly available computing entity, and the cluster file system is an essential technology to enable this method. Cluster file systems generally scale well up to a certain size and are typically limited by heavy interprocess communication (IPC).

Cluster file systems can be further classified into four groups based on the approaches they take in accessing the data and metadata. In-band SAN-based cluster file systems interpose a cluster of servers between clients that want to access the data and the SAN-attached storage. The cluster of servers manages all I/O traffic including file system metadata to locate and create files, traverse directories, and so on. The file data is cached on the cluster to improve both read and write operations. The cluster file system running on the cluster typically supports full cache coherency among all the cluster nodes.

Alternatively, out-of-band SAN-based cluster file systems access files directly over a SAN. File information such as locking and location resides in a dedicated server or a cluster of servers connected to the network. Typically, a client would first request metadata using the local area network, and then access the file contents directly over the SAN. A SAN-based cluster file system generally provides high I/O throughput, making this approach attractive for I/O-intensive applications such as genomics and video processing.

In-band and out-of-band direct attach storage (DAS)–based cluster file systems are two other classifications being offered by various vendors today. A cluster file system is implemented across a series of file servers, each with its own private storage. The DAS class of cluster file system offers the simplicity of DAS deployment but an inferior management paradigm when compared with SAN-based methods. Also, data mirroring is essential in a DAS-based scheme, to avoid loss of data access caused by failures in servers. The DAS-based

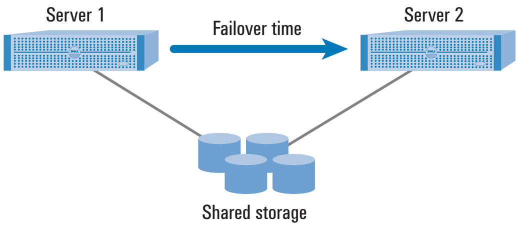


Figure 1. File system access in traditional high-availability cluster

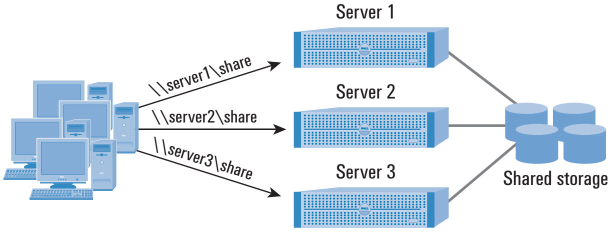


Figure 2. Traditional, non-virtualized file-server access

cluster file system architecture serves well in computing environments that can intelligently partition the data, process the data in separate partitions simultaneously, and follow up with a coalescing process. This approach enables applications to realize very high I/O throughput because multiple DAS servers are simultaneously engaged in the I/O activity.

How to choose scalable file systems

When deciding whether a scalable file system can help meet the challenges in a given IT environment, administrators should evaluate selection criteria in four basic categories: availability, performance, scalability, and manageability.

Availability. In the most basic high-availability architecture—a failover cluster—availability can be affected by factors such as I/O latency between cluster nodes and application startup time. If an application runs on only one node, and that node fails, the failover node requires some time to take ownership of the shared storage. If that failover time exceeds the window of acceptable downtime, then this architecture does not meet the need for high availability. Figure 1 shows the challenge associated with a traditional high-availability cluster.

To provide high availability in a multinode environment, cluster file systems can allow applications to concurrently access the same disks. This method eliminates the need for an ownership transfer between nodes in the event of a node failure.

Architecturally, the primary goal of high-availability file systems is to provide redundancy for all components: Clients must be able to access data across the network, even if a file-serving node fails. Virtualization can aid availability by allowing flexibility in design as well as scalability, as follows:

- **Storage virtualization:** Virtualization of storage is the abstraction of multiple physical storage devices into a single logical pool of data. This approach enables multiple data-serving devices—either file servers or NAS heads—to concurrently access the same pool of storage. If any one data-serving device fails, clients can still access the data through the remaining active devices.
- **File-server virtualization:** Virtualization of file servers is the abstraction of multiple physical file servers or NAS heads into a single logical file server. When client systems access data from a virtual file server, it is transparent to the client systems which, or how many, physical file servers deliver the requested data. Aggregating file servers so that all clients access the same network resource allows for the transparent redirection of client requests.

In Figure 2, clients access the data on server 1, server 2, and server 3 using a different path to access each server (\\server1\share, \\server2\share, \\server3\share). In Figure 3, the clients access the same three servers, but do so through a virtual network share (\\cluster\share).

Performance. Measurement of performance begins with a baseline. Benchmarks such as SPECsfs[1] can help determine the maximum possible performance of a system, but cannot answer the question of how a system performs in a specific environment unless other factors are taken into account.

Baseline performance data should describe not only the I/O capacity of the current file server or NAS head, but also the data consumption characteristics of the clients. Using this data, administrators can determine which approach meets their current needs and to what extent their architecture might scale in the future.

If a file server is capable of delivering *x* MB/sec of throughput, and the data consumption of a given client is *y* MB/sec, then the maximum number of clients that server can support is *x/y*. If a given environment needs to support 2*y* or 10*y* clients, administrators may

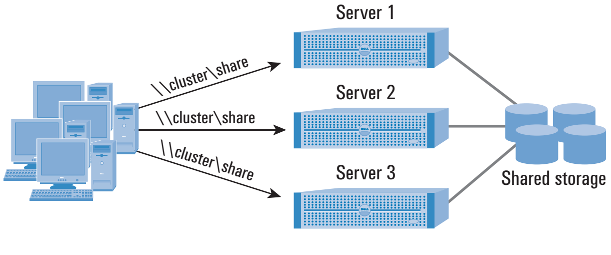


Figure 3. Virtualized file-server access

[1] For more information about SPECsfs, visit www.specbench.org/sfs97r1.

be able to add more file servers, but adding more file servers will not increase the bandwidth unless the data is distributed across the servers. If all the clients need to access the data concurrently, this solution is not optimal.

Scalability. If a file server or NAS head is nearing its I/O capacity and clients still need to access their data from a single network share, administrators may be able to replace the device with a larger, more capable model. However, at some point the cost of a new system and the disruption caused by the replacement can no longer be borne by the budget or the clients.

How data is distributed over multiple arrays is a key differentiator between SAN-based cluster file systems and DAS-based cluster file systems. At the point when data throughput needs exceed the I/O capacity of a single back-end SAN, administrators must be able to aggregate the performance of multiple SANs. Administrators must consider whether their SAN-based cluster file system can span multiple arrays and, if so, whether they can meet the challenge of hot-spot elimination and dynamic tuning, as follows:

- **Performance aggregation:** Storage virtualization can expand a DAS-based cluster file system to incorporate multiple storage arrays or SANs serviced by multiple file servers or NAS heads.
- **Data hot spots:** Blocks of data that are concurrently accessed by multiple clients—hot spots—cause I/O bottlenecks when multiple file-serving heads retrieve the same data from the storage array.
- **Dynamic tuning:** By incorporating sophisticated performance measurements into the file system, dynamic tuning enables the file system to reallocate resources in response to peak loads. Whether those loads come from network requests or device contention on the storage array, dynamic tuning can enable a file system to marshal resources and respond to load spikes without changes to the IT architecture and without impact on client systems.

Manageability. Management of high-availability storage environments can be as simple as fault alerting and automated recovery in the event of a failure. Performance management involves the collection of metrics and, eventually, the capability to make informed decisions to optimize performance. To manage scalability, however, administrators must be able to incorporate alerting and recovery information with performance metrics—not only to increase the capabilities of the scalable file system, but also to grow the scalable file system organically.

Over time, manageability has the potential to affect the total cost of an environment more than any single technology. People, hardware, and time are resources that must be managed carefully. Faults that require an administrator's presence—for example, to replace a failed component—cost money, as do architectural changes in the environment. When considering ways to enhance data throughput without replacing the existing infrastructure, administrators should evaluate two key aspects of manageability:

- **Support and interoperability:** This can be the most important factor in deciding whether an environment is manageable. For example, does the hardware provider partner with the software provider to offer recommendations for configuration and tuning? Will they work with each other to provide resolution in the event of a fault?
- **Backups:** Data backups are a critical but sometimes unconsidered component in a scalable file system environment. The technologies employed to distribute data blocks across storage arrays can pose challenges to traditional backup products.

Evaluation of current and future needs

Purpose-built storage systems are designed to deliver high performance and availability, but lack manageability and practicality if they must be redesigned or completely replaced to provide data access as the enterprise environment grows. Cluster file systems have the potential to provide excellent performance and expandability that traditional file-server architectures do not necessarily offer.

Even when current availability and data throughput needs can be met by a commercial file server or NAS cluster, organizations must consider future needs. Understanding the existing data service environment and clearly identifying objectives for tomorrow's file-service requirements are critical to choosing the most appropriate approach for increasing data movement capacity. By applying the criteria presented in this article against both traditional file service and scalable file system solutions, administrators can help enhance the likelihood of choosing the technology that best addresses their present and future data movement needs.

David Weber is an enterprise technologist in the Advanced Systems Group at Dell. He works with Dell solution consultants and customers to communicate Dell's capabilities and strategies for Linux. Before coming to Dell, David was a senior network security engineer responsible for penetration assessments at Security Design International.

Roger Goff is an enterprise technologist in the Advanced Systems Group at Dell. His current interests include Linux and Microsoft® Windows® HPC clusters and cluster file systems. Roger is a Red Hat Certified Engineer® (RHCE®) and has an M.S. and a B.S. in Computer Science from Virginia Polytechnic Institute and State University.

Suri Brahmaroutu is a software architect and strategist with the office of the CTO at Dell. Suri has been associated with designing one of the first distributed computing environments in the industry and designing several server I/O technologies, including InfiniBand. He has a bachelor's degree in Electronics and Communications Engineering and a master's degree in Computer Science from the University of Hyderabad in India.

The Promise of Unified I/O Fabrics

Two trends are challenging the conventional practice of using multiple, specialized I/O fabrics in the data center: server form factors are shrinking and enterprise applications are requiring more connections per server. However, the current practice of using multiple Ethernet and Fibre Channel connections on each server to support fully redundant, cluster-enabled computing environments inhibits scalability. Unified, high-performance I/O fabrics can enhance scalability by providing a single fault-tolerant connection. This approach allows legacy communication technologies to use one I/O "superfabric," and the reduction in physical connections can help achieve better performance, greater flexibility, and lower total cost of ownership—the primary benefits of the scalable enterprise.

BY J. CRAIG LOWERY, PH.D., AND DAVID SCHMIDT

The modern data center is a collection of server and storage components partitioned into various cooperating groups that communicate over specialized networks. In most cases, the technologies behind these networks were conceived decades ago to address particular kinds of traffic, such as user access, file transfer, and high-speed peripheral connections. Over time, data centers incrementally evolved to meet the increasing requirements of their environments, often retaining vestigial characteristics of repeatedly revamped technologies for backward compatibility and interoperability. Although the standardization and stability enabled by backward compatibility and interoperability have paved the way for the proliferation of computer systems, it is becoming increasingly difficult to extend these legacy technologies to meet the fundamentally different requirements imposed by the scalable enterprise.

For example, Ethernet—the de facto standard for local area network (LAN) communication—began as a rather cumbersome bus architecture with performance limitations imposed by the shared nature of its medium access control protocol. Today, Ethernet has become a much faster switched communication standard, evolving from 10 Mbps to 100 Mbps to 1 Gbps. Yet, the remnants of its past—the bus-based architecture and, in particular, the Carrier Sense Multiple Access with Collision Detection (CSMA/CD) protocol—introduce unnecessary overhead for the sake of compatibility, making Ethernet less attractive for protocols such as Remote Direct Memory Access (RDMA) than newer interconnects without the same historical baggage, such as InfiniBand.

Other interconnect technologies, such as SCSI, Peripheral Component Interconnect (PCI), and Fibre Channel, have followed a similar trajectory. In each case, the technology was created to solve a particular problem and, over time, has been extended to increase both performance and scope of application.

An unfortunate side effect of the proliferation of multiple interconnect technologies is the requirement that they coexist in smaller and smaller spaces. High-density servers—such as rack-dense, 1U servers and blade servers—are required to provide user experiences equivalent to larger systems such as the traditional tower. At the same time, emerging enterprise applications for these high-density systems require an increasing number of connections, or large fan-outs. Some clustering systems, such as high-availability clusters and clustered databases, require multiple LAN connections and two Fibre Channel connections for a fault-tolerant storage area network (SAN). Fitting four or more of these connections into a blade server's form factor can be challenging.

Now that the barriers to mass adoption have been addressed, unified I/O fabrics are set to revolutionize computing infrastructures through their flexible, extensible architectures.

Another drawback of legacy interconnects is that they do not inherently encompass the fabric concept. A fabric functions much like a public utility: it is a multipurpose interconnect that is accessible from virtually anywhere. The vision of the scalable enterprise depends largely on fabric *semantics*—the model of communication that determines how enterprise applications "speak" within the data center that employs the fabric—because next-generation data centers will likely be built using standard, disposable components that plug in to the infrastructure as capacity is needed. Fabrics are the key to this plug-and-play data center. Although some technologies such as Ethernet come closer than others such as SCSI to delivering a fabric-like usage semantic, they still fall short in key areas, primarily by requiring additional unnecessary overhead to support their legacy aspects. For example, using TCP/IP over Ethernet to perform RDMA significantly wastes bandwidth and is unnecessarily slow for a high-performance computing cluster rack interconnect, because TCP's sliding window protocol was designed for the unreliable Internet—not a single, well-controlled rack with a high-speed communication link.

Heterogeneous legacy interconnects are also hindered by the support structure required to maintain data centers that incorporate them. Today, IT support teams must staff skills in each interconnect technology. This redundancy is inefficient when compared to the single-culture support required to maintain a unified fabric. A unified fabric subsumes all communication functions through one fabric connection or—for redundancy—two fabric connections. The fan-out problem can be resolved at the software level by multiplexing multiple virtual interfaces over the single physical interface of a unified fabric. Some of these virtual interfaces may be designed to appear to higher layers of software as legacy technology interfaces to help provide transparency and backward compatibility.

As the deficiencies of heterogeneous interconnects in the scalable enterprise intensify and the need for fabric semantics mounts, a clear gap arises that cannot be adequately filled by additional iterations to refine older technologies or make them suitable and relevant going forward. It is this need that the unified I/O fabric is designed to address.

Understanding the requirements of unified I/O fabrics

Any technology candidate that puts itself forward as being a unified I/O fabric technology must meet the following suitability requirements:

- **Transparent coexistence:** The fabric must be able to coexist and interoperate with legacy interconnect technologies without placing additional requirements on end users and applications.
- **High performance:** The fabric must be able to accommodate the aggregate I/O that would otherwise have been distributed across legacy interconnects. Nonblocking connectivity, throughput, and latency should be optimized so that the performance of the unified fabric is the same as or better than the performance of multiple legacy networks.
- **Fault tolerance:** The fabric must respond gracefully to component failures at both the legacy interconnect layer and its own unified layer. Furthermore, to meet the requirement of transparent coexistence, the fabric should support legacy fault-tolerance technologies.
- **Standardization:** The fabric must conform to industry standards to ensure competitive pricing, multiple sources for components, longevity of the technology, and the creation of an attendant ecosystem. *Ecosystem* refers to all the companies, services, and ancillary products that must come into existence to make the technology viable and deployable on a large scale.
- **Value pricing:** The fabric must be less expensive to procure and maintain than an equivalent combination of legacy interconnects.

Unified I/O interconnects are not a new idea—proprietary solutions have been developed and deployed with some success in targeted, custom environments. However, most efforts to date have not met all of the preceding requirements, usually failing on transparency, standardization, and pricing. Recently,

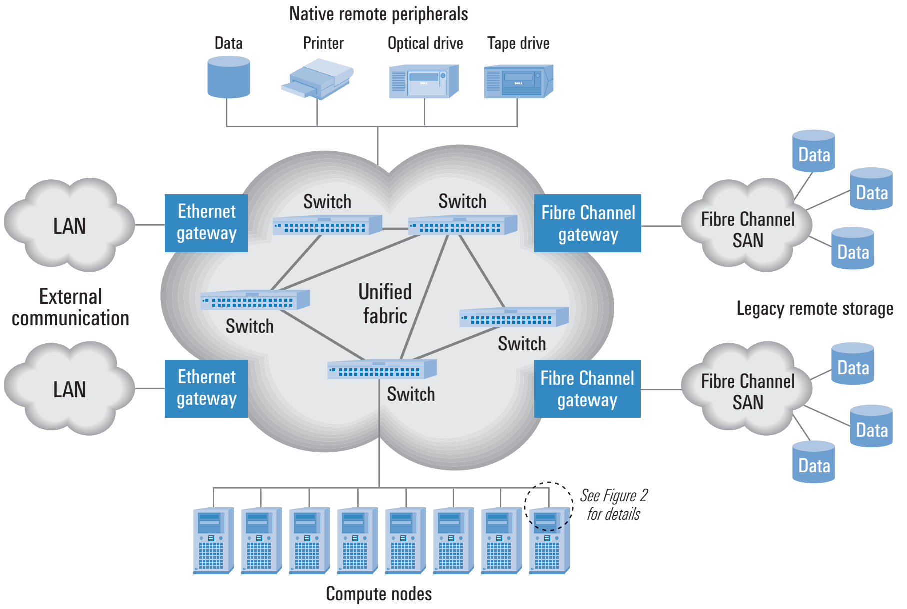


Figure 1. Unified fabric architecture

technologies like InfiniBand have been specifically designed to meet all of these requirements. Now that the barriers to mass adoption have been addressed, unified I/O fabrics are set to revolutionize computing infrastructures through their flexible, extensible architectures.

Examining the unified I/O fabric architecture

Figure 1 shows an overview of a unified I/O fabric architecture. At the center of the figure is the unified fabric, comprising one or more switches. The specific technology used in the switch is not of particular importance to the concept, although InfiniBand is one currently available candidate. Ethernet gateways allow for IP traffic between devices connected to the fabric and external networks. Fibre Channel gateways provide similar connectivity to SANs. Various remote peripherals that have native fabric interfaces are shown at the top of the figure. Such devices can communicate directly on the fabric and do not require a gateway.

The lower portion of Figure 1 shows an array of compute nodes. Although these are depicted as identical in the figure (as would be the case with blade servers), the nodes in the array can consist of various form factors and system models. Each compute node must have at least one fabric interface with which to connect, and each node must host an operating system with a software stack that consists of a native fabric driver for the physical fabric interface, as shown in Figure 2. In addition to transmitting and receiving data, this driver may incorporate or be bundled with additional software to aid in the distributed management of the fabric.

Operating systems on the compute node can communicate over the fabric either by using the native fabric protocols or by using mapped or tunneled legacy protocols, as shown in Figure 3. A mapped protocol is one that can be translated to and from the fabric protocol and requires that the fabric protocol directly support similar functionality. When no direct mapping exists, a protocol must be tunneled through the fabric, meaning that the legacy protocol's messages are embedded, or *wrapped,* in the fabric protocol for transport across the fabric. Mapping is usually more efficient because the fabric comprehends the mapped protocol and can be optimized for it. Both mapped and tunneled protocols require a gateway to connect the fabric to the legacy networks and perform the mapping and tunneling functions.

For example, the InfiniBand specification incorporates IP over InfiniBand (IPoIB), which allows IP datagrams to be mapped directly to InfiniBand packets for transport and routing over the fabric. InfiniBand also is designed to provide a standard mechanism for mapping SCSI device commands to the fabric, either directly to a SCSI device attached on the fabric or to a Fibre Channel gateway.

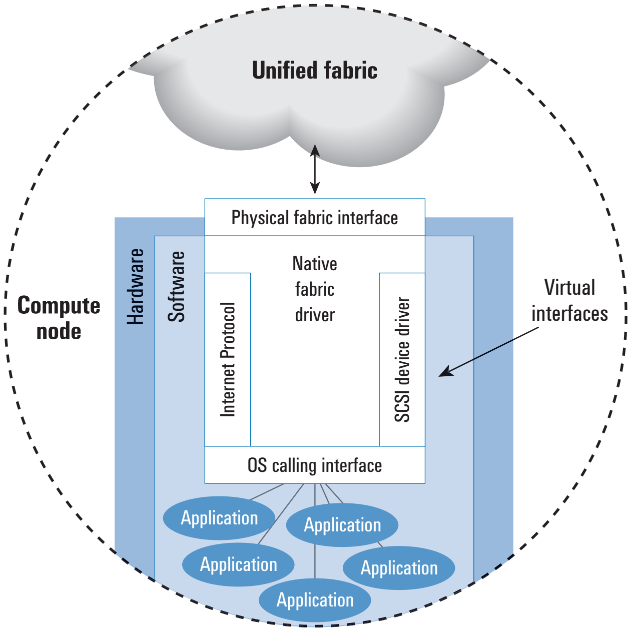
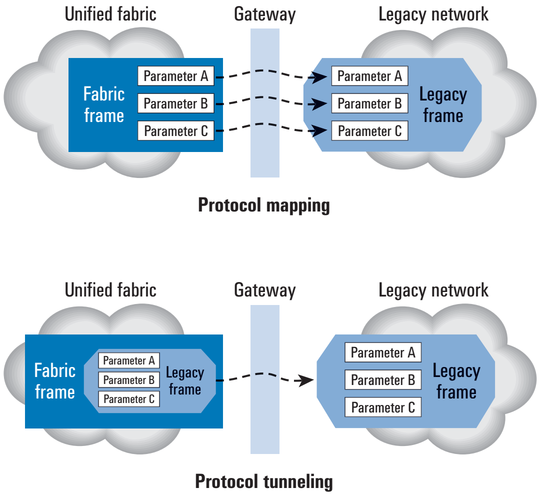


Figure 2. Compute node communication stack



Protocol mapping

Protocol tunneling

Figure 3. Protocol mapping and tunneling

Considering InfiniBand for the I/O fabric

InfiniBand, designed as a next-generation I/O interconnect for internal server devices, is one viable option for unified I/O fabrics. This switched, point-to-point architecture specifies several types of fabric connections, allowing compute nodes, I/O controllers, peripherals, and traffic management elements to communicate over a single high-speed fabric. Because it can accommodate several types of communication models, like RDMA and channel I/O, InfiniBand is a compelling candidate for unifying legacy communication technologies. InfiniBand meets most of the suitability requirements for a unified I/O fabric primarily because it was conceived and designed to function as a single, ubiquitous interconnect fabric.[1]

Transparent coexistence between legacy technologies and unified fabrics is a crucial requirement for fabric technology. InfiniBand is designed to support connections to legacy networks via target channel adapters (TCAs). TCAs enable legacy I/O protocols to use InfiniBand networks and vice versa. They support both mapping and tunneling operations like those shown in Figure 3. Tunneling requires encapsulating the legacy transport and payload packets within the InfiniBand architecture packet on the unified fabric, allowing the legacy protocol software at each end of the connection to remain unaltered. For example, a TCP/IP connection between a legacy network node and a unified fabric node can take place entirely over the TCP/IP stack. The unified fabric node needs the InfiniBand software only to remove the TCP/IP packet from the InfiniBand packet. Mapping, or *transport offloading,* actually removes the IP information from the InfiniBand packet and transmits it natively on the IP network. The unified fabric node can take full advantage of the InfiniBand transport features, but the TCA must translate between InfiniBand packets and IP packets—a potential bottleneck. Although InfiniBand was designed with transparency in mind, the performance and benefits of these methods will ultimately depend on the design and implementation of TCA solutions.

The high bandwidth of InfiniBand makes it possible for several legacy protocols to coexist on the same unified connection. InfiniBand currently supports multiple levels of bandwidth. Single-link widths, also called 1X widths, support a bidirectional transfer rate of up to 500 MB/sec. Two other link widths, 4X and 12X, support rates of up to 2 GB/sec and 6 GB/sec, respectively. Compared with today's speeds of up to 125 MB/sec (1 Gbps) for Ethernet and 250 MB/sec (2 Gbps) for Fibre Channel connections, InfiniBand bandwidth has enough capacity to support both types of traffic through a single connection. However, as legacy network bandwidth improves, unified fabric solutions must scale to meet the enhanced speeds of traditional I/O networks. InfiniBand must increase its bandwidth or unified I/O solutions must use multiple InfiniBand connections to garner a higher aggregate bandwidth.

As with other packet-based protocols, InfiniBand employs safety measures to help ensure proper disassembly, transmission, and reassembly of transported data. This can be considered fault tolerance in its most basic form. Furthermore, just as other protocols specify fault tolerance between connection ports, InfiniBand allows for two ports on the same channel adapter to be used in a fault-tolerant configuration. Therefore, InfiniBand can help enable a unified fabric solution to implement fault tolerance and failover features. Fault tolerance at the legacy technology layer, however, must still be provided by legacy software stacks. If a TCA is used to connect a legacy network to the InfiniBand fabric, the unified fabric solution must provide fault-tolerant links between the two. Again, the performance and benefits of internetwork fault tolerance will depend on the design and implementation of unified fabric solutions.

InfiniBand meets most of the suitability requirements for a unified I/O fabric primarily because it was conceived and designed to function as a single, ubiquitous interconnect fabric.

InfiniBand was designed with a goal of multiple physical computing devices existing on the same network. Different physical connections might be necessary for different elements of the unified network, and different I/O access models might be

[1] For more information, see www.infinibandta.org/ibta.

required for different legacy protocols. Fortunately, the InfiniBand Trade Association (IBTA) has set forth several options for connectivity. Physical connections can use copper or optical interconnects, and connection modules can use channel-based I/O protocols for legacy operations or zero-copy RDMA to reduce latency and CPU overhead. As a variety of unified fabric solutions becomes available, the IBTA plans to help ensure a standardized playing field that will allow interoperability with third-party management and software solutions.

Because the high-performance characteristics of InfiniBand allow for the transport of multiple legacy networks on a single connection, unified InfiniBand fabric solutions can help reduce the number of physical ports.

Because the high-performance characteristics of InfiniBand allow for the transport of multiple legacy networks on a single connection, unified InfiniBand fabric solutions can help reduce the number of physical ports. This reduction can simplify the physical view of the data center configuration, and result in a lower risk of physical port failure and a faster configuration time—which can translate to lower costs for network management. IBTA specifies standards for managing subnets within the InfiniBand network, providing network management solutions with a method of monitoring and controlling InfiniBand network configuration and performance. By utilizing these management standards, organizations can realize the value of unified I/O fabrics with InfiniBand.

Migrating high-speed networks to unified I/O architectures

As other I/O technologies evolve, unified I/O architectures will have more options for the underlying high-speed fabric. The advent of TCP/IP Offload Engines (TOEs) allows traditional Ethernet fabrics to utilize RDMA and hardware-based TCP/IP stacks, thereby reducing CPU overhead. Storage technologies such as Internet SCSI (iSCSI) can then efficiently utilize Ethernet as the unified fabric. InfiniBand, however, is a leading choice for immediate adoption of a unified fabric. Even though some InfiniBand components may not be widely available or competitively priced with legacy networks, the technology itself is viable and proven. Based on previous technology rollouts, mass adoption of InfiniBand over time can help increase its availability and lower its cost. The transparency benefits and legacy support built into this architecture can help drive rapid adoption within data centers, and InfiniBand's high-capacity switched architecture can provide reliable performance for multiple I/O models.

Solutions that utilize the high performance and simplified management of unified I/O are already entering the marketplace, and network architects are planning for data centers in which servers can be dynamically matched with I/O resources via a unified high-performance fabric. With the maturation of high-speed network architectures like InfiniBand, the benefits of unified I/O fabrics cannot be denied. It is only a matter of time before these fabrics become a requirement for the scalable enterprise.

J. Craig Lowery, Ph.D., is a senior engineering development manager on the Enterprise Solutions Engineering team within the Dell Product Group. His team is currently responsible for developing products that realize the Dell vision of the scalable enterprise. Craig has an M.S. and a Ph.D. in Computer Science from Vanderbilt University and a B.S. in Computing Science and Mathematics from Mississippi College.

David Schmidt is a systems engineer on the Enterprise Solutions Engineering team within the Dell Product Group, where he develops solutions for the scalable enterprise. Previously, David worked in the systems management group developing the Dell OpenManage™ Deployment Toolkit. David has a B.S. in Computer Engineering from Texas A&M University.

FOR MORE INFORMATION

InfiniBand Trade Association:
www.infinibandta.org

Progressive Degrees of Automation Toward the Virtual Data Center

As IT organizations make the transition toward fully automated virtual data center (VDC) operations, numerous paths can lead to environments that look good on paper but are premature, overly expensive, or proprietary. This article presents an overview of the VDC model and discusses degrees of maturity, readiness requirements, and best practices for each step in the progression toward full data center automation.

BY JIMMY PIKE AND TIM ABELS

The emerging virtual data center (VDC) has at its core an open standards architecture. At its highest stage, the VDC functions as a closed-loop management system that is designed to become increasingly self-managed over time. The key, high-level functions of the VDC and their interrelationships are as follows:

- **Enterprise management:** This framework typically includes higher-order management packages, including service-level expectations and operational management.
- **Monitoring:** Operational data includes various parameters that can be analyzed to determine whether operations are aligned with current business policies.
- **Orchestration:** Controlling software provides the capability to synthesize operational data (collected during monitoring) into knowledge about operational behavior, and to use that knowledge to understand how resources should be managed for the most desirable operation.
- **Resource managers:** The use of controllers helps administrators allocate and manage assets, and map their relationships to one another.
- **Support systems:** Self-contained items aid orchestration and resource managers in controlling system-wide behavior.
- **Mapping:** A system-wide directory tracks relationships between resources.
- **Element management:** The element management function consists of software that helps administrators manage specific resource types—such as servers, storage, and fabrics—and may include software that directly interfaces with physical resources as well as programs required to configure or operate the hardware.
- **Applications and operating systems:** Operating systems provide the capability to use the physical resources in the architecture, while applications and services offer the functionality required to complete the desired tasks in an enterprise environment.
- **Standard servers, storage, and fabrics:** Standard servers, storage, and fabrics comprise the physical resources in the architecture. The term *standard* indicates that these resource types conform to generally accepted industry standards.

Figure 1 shows various VDC implementation possibilities and opportunities. The remainder of this article focuses on the steps that lead toward a fully automated VDC.

Steps leading toward the virtual data center

The development of a VDC progresses in five distinct steps, or degrees. Not every organization will find it desirable to move beyond a specific degree of automation. Such decisions are based on individual organizational needs and capabilities, and any stage has the potential to be an end state. The five degrees of maturity progressing toward the VDC are as follows:

- **Degree 1—Solution-based automation:** The first step in creating a VDC is focused on the standardization of specific computing tasks to fulfill various organizational needs such as messaging or human resources tasks. It may include deploying major enterprise applications from vendors such as Oracle, SAP, Microsoft, and PeopleSoft using migration best practices. Discrete scalable solutions may also start small and follow a prescriptive method for scaling as needed.
- **Degree 2—Automated common infrastructure:** The second step in creating a VDC is achieved through the addition of infrastructure that supports cooperation between groups of discrete solutions. At this stage, server virtualization can help enhance equipment utilization, and a reduction in the number of physical servers can lower operational costs and staffing. A common infrastructure is achieved through the addition of infrastructure that automatically manages systems and operations common to the included groups.

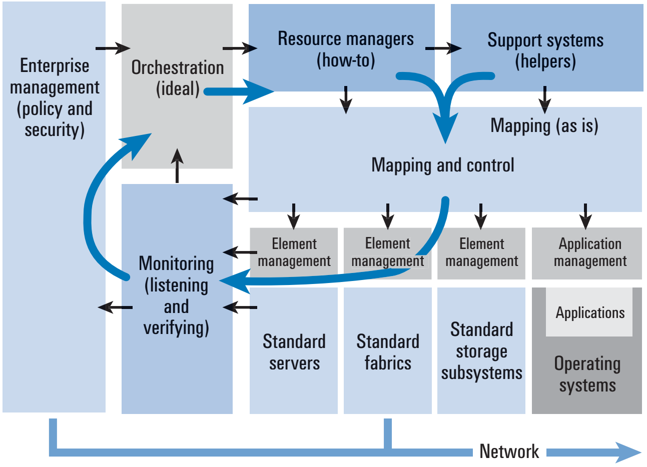


Figure 1. Virtual data center schematic

- **Degree 3—Dynamic automation:** This step is crucial in realizing the VDC because it begins to enable end-to-end, service-oriented automation by establishing enterprise management and orchestration among groups. Specifically, this step introduces tools that allow the data center to automatically allocate or reallocate resources based on a set of operational criteria defined by service needs. To achieve this degree of automation, organizations must first establish standards in both business policy and interoperability.
- **Degree 4—Business-based computing:** This step represents business-based computing. While aspects of policy based on business drivers are apparent in the previous stage, this degree of automation represents a model in which business goals and motivations become the key IT driver. This step represents a highly flexible, standardized infrastructure in which technical services become the commodity. As business policies increasingly drive every aspect of the way services are provided, organizations can expect significant changes in the interfaces between users, applications, and services—including a high degree of automation built around business needs.

The first step in creating a VDC is focused on the standardization of specific computing tasks to fulfill various organizational needs such as messaging or human resources tasks.

- **Degree 5—Virtual data center:** The VDC represents the ultimate in IT evolution. At this stage of automation, the connection between an application and the underlying physical assets on which it is executed will have little relevance. The goal of the VDC is to automate every aspect of an organization's business operations, and enable such operations to be configured or reconfigured as needed to achieve maximum business effectiveness. Much of the actual construction and operation of a fifth-stage VDC remains to be determined.

Before attempting to implement a given degree of automation, organizations should identify and understand what can be expected at that stage. Figure 2 indicates the readiness requirements and associated best practices for all five degrees of automation.

Both the infrastructure readiness requirements and associated best practices build a logical progression toward the VDC, with a distinct and realizable value proposition at each degree of maturity.

Degree	Readiness requirements	Associated best practices
1	Standardized servers, switches, storage, operations, systems, and applications	Use of best practices for application scaling, migration, and operations; conformance to IT infrastructure library (ITIL) and Microsoft® Operations Framework (MOF) operational standards
2	Boot from SAN; extensive monitoring; multifabric usage; deployment of high-bandwidth, low-latency interconnects	Server, storage, and network virtualization; common configurations in multiple groups; limited intergroup cooperation
3	Cooperative management tools, centralized relationship mapping, diskless operation	High degree of automation and consolidation, including provisioning and reprovisioning of the target environment
4	Authoritative management and orchestration, one authoritative agent per resource, centralized relationship mapping that includes business attributes	Business-based automation for all major solutions, full equipment interoperability, full reusability and provisioning based on changing business needs
5	Data center is fully abstracted as a "black box" service center	Automatic provisioning of application or service requests to best support business operations, fully automated and self-managed data center

Figure 2. Degree of automation, organizational readiness, and best practices for the VDC

This information enables choice, helping organizations maintain control over when and how to best evolve an IT environment so that it is consistent with business goals and practices.

Early preparation for achieving the virtual data center

While the evolution toward the VDC is progressing quickly, the target degree of automation for each data center must be based on each organization's individual needs. Enterprises should carefully assess the readiness of their infrastructure and adhere to the associated best practices for each stage. Although solutions for degrees 4 and 5 remain proprietary, degrees 1, 2, and 3 involve standardization. Thus, organizations can focus on laying the groundwork for the VDC today with degrees 1, 2, and to some extent, 3.

Dell's concept of the scalable enterprise is a staged, prescriptive, and pragmatic approach that can help provide organizations with an easy transition to emerging capabilities as they become standard—such as those envisioned in degrees 3, 4, and 5. Dell offers guidance to help smooth this journey and deliver maximum value for IT investments.

Jimmy Pike is a director and distinguished engineer in Dell's Server Architecture and Technology Group, where he serves as an enterprise architect. He is responsible for the strategic system architecture of Dell's enterprise product line. Jimmy is a seasoned veteran with more than 25 years of experience in the server industry. He has received numerous patents through extensive work in the area of symmetric multiprocessing in both large and small systems.

Tim Abels is a senior software architect currently developing scalable enterprise computing systems at Dell. Tim has an M.S. in Computer Science from Purdue University.

Deploying and Managing Oracle RAC with Oracle Enterprise Manager 10*g* Grid Control

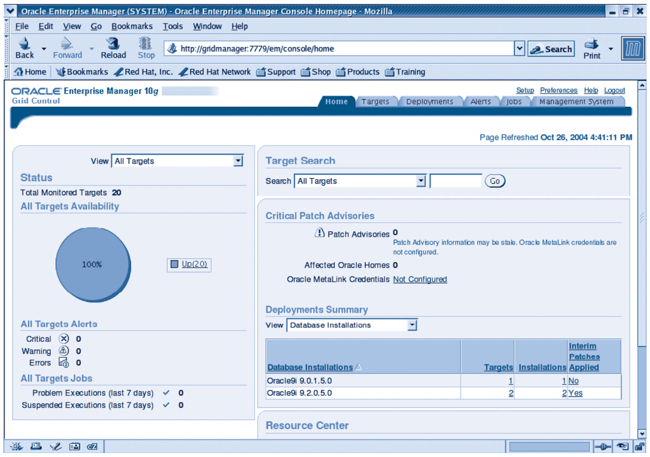
Oracle® Real Application Clusters 10*g* provides enterprises with a highly available, scalable, and cost-effective way to deploy their information systems: Oracle Enterprise Manager 10*g* Grid Control. This robust management framework can be used to provision, clone, and automate patching—and it is designed to scale well in heterogeneous environments. This article explains the setup and usage of this tool for administering and monitoring database cluster systems.

BY RUDRAMUNI B, C.S. PRASANNA NANDA, AND UDAY DATTA SHET

Oracle Enterprise Manager (EM) management software offers a management framework called Grid Control, which provides a centralized, integrated approach for managing different versions of Oracle products in the enterprise. Grid Control uses HTTP and HTTP over Secure Sockets Layer (HTTPS) protocols to provide administrators with a browser-based management interface. The EM suite is bundled with a Grid Control management console; a management service, which includes management repository components; and a management agent. The self-monitoring features of EM help ensure that critical components of Grid Control are always available and functional.

The Grid Control framework has a three-tiered architecture comprising the management console, management service, and management agent, as follows:

- **Management console:** This is the central console from which administrators can manage the Grid Control framework (see Figure 1). The console is Web-based and can be connected from any supported Web browser. The browser communicates with the management service over standard protocols such as HTTP and HTTPS to enable communication within EM.
- **Management service:** The management service is the middle tier in the EM suite. The management service and management repository generally reside on a system that is separate from the database server and application server. The management service communicates with the management agents, which are deployed on target nodes, and collects host and database-related statistics.
- **Management agent:** Managed servers are the final components in the EM suite. A management agent runs on each monitored target. The management agent is a lightweight process that is responsible for monitoring different services and host parameters on the host on which it is deployed.



Oracle Enterprise Manager (SYSTEM) - Oracle Enterprise Manager Console Homepage - Mozilla

http://gridmanager:7779/em/console/home

Page Refreshed **Oct 26, 2004 4:41:11 PM**

View: All Targets

Status

Total Monitored Targets **20**

All Targets Availability

Up(20)

All Targets Alerts

Critical	0
Warning	0
Errors	0

All Targets Jobs

Problem Executions (last 7 days)	✓ 0
Suspended Executions (last 7 days)	✓ 0

Target Search

Search: All Targets (Go)

Critical Patch Advisories

Patch Advisories **0**
Patch Advisory information may be stale. Oracle MetaLink credentials are not configured.
Affected Oracle Homes **0**
Oracle MetaLink Credentials Not Configured

Deployments Summary

View: Database Installations

Database Installations	Targets	Installations	Interim Patches Applied
Oracle9i 9.0.1.5.0	1	1	No
Oracle9i 9.2.0.5.0	2	2	Yes

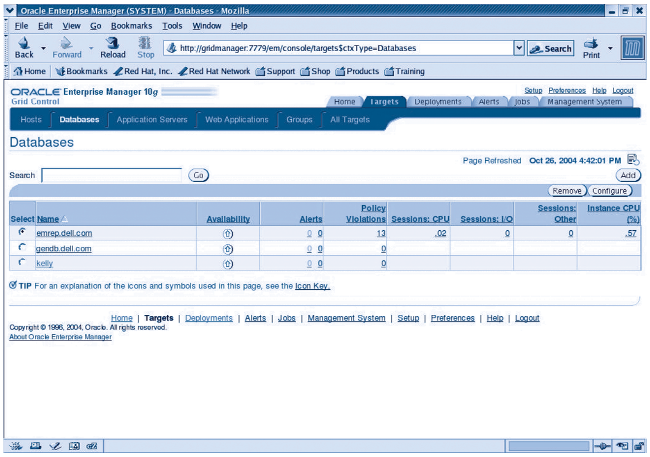
Resource Center

Figure 1. Grid Control home page

Features of the Grid Control framework

Grid Control provides a robust, reliable, and scalable management framework across the enterprise. The key Grid Control features are as follows:

- **Scalability:** Grid Control has been designed for scalability. In heterogeneous environments, which include databases and application servers of different versions on different platforms, the Grid Control framework scales seamlessly even when the environment has a large number of servers (see Figure 2). To add a new system to the management target list, the administrator simply installs the management agent on the system. As the IT infrastructure grows, administrators can add more management servers for server load balancing.
- **Consolidated management:** Every managed server appears as a target on the Grid Control management console. On the console, each server has a home page that provides a consolidated view of server parameters such as resource utilization (CPU, memory, and disk); performance characteristics; and configuration options. Each database (cluster and noncluster) also appears as a target on the Grid Control home page and can be monitored and administered through the console after the administrator logs in to the database.
- **Logical grouping of targets:** The targets on the console, including the database servers and application servers, can be grouped logically for ease of management. For example, administrators can create a group for all the hosts running Oracle Real Application Clusters (RAC) 10*g* in the enterprise. Similarly, another group can be defined for all the hosts running Oracle9*i*™ on Linux® platforms.



Oracle Enterprise Manager (SYSTEM) - Databases - Mozilla

http://gridmanager:7779/em/console/targets$ctxType=Databases

Databases

Page Refreshed **Oct 26, 2004 4:42:01 PM**

Search (Go) (Add)

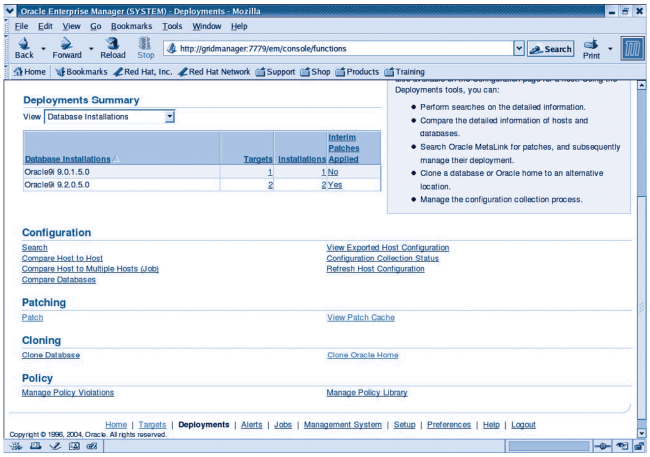
(Remove) (Configure)

Select	Name	Availability	Alerts	Policy Violations	Sessions: CPU	Sessions: I/O	Sessions: Other	Instance CPU (%)
●	emrep.dell.com	⇧	0 0	13	.02	0	0	.57
○	gendb.dell.com	⇧	0 0	0				
○	kelly	⇧	0 0	0				

TIP For an explanation of the icons and symbols used in this page, see the Icon Key.

Home | **Targets** | Deployments | Alerts | Jobs | Management System | Setup | Preferences | Help | Logout
Copyright © 1996, 2004, Oracle. All rights reserved.
About Oracle Enterprise Manager

Figure 2. Available databases



Oracle Enterprise Manager (SYSTEM) - Deployments - Mozilla

http://gridmanager:7779/em/console/functions

Deployments Summary

View: Database Installations

Database Installations	Targets	Installations	Interim Patches Applied
Oracle9i 9.0.1.5.0	1	1	No
Oracle9i 9.2.0.5.0	2	2	Yes

Deployments tools, you can:

- Perform searches on the detailed information.
- Compare the detailed information of hosts and databases.
- Search Oracle MetaLink for patches, and subsequently manage their deployment.
- Clone a database or Oracle home to an alternative location.
- Manage the configuration collection process.

Configuration

Search
Compare Host to Host
Compare Host to Multiple Hosts (Job)
Compare Databases
View Exported Host Configuration
Configuration Collection Status
Refresh Host Configuration

Patching

Patch
View Patch Cache

Cloning

Clone Database
Clone Oracle Home

Policy

Manage Policy Violations
Manage Policy Library

Home | Targets | **Deployments** | Alerts | Jobs | Management System | Setup | Preferences | Help | Logout
Copyright © 1996, 2004, Oracle. All rights reserved.

Figure 3. Deployment page

- **Automation of tasks:** Automation of various tasks is one of the most salient features of Grid Control. Each automated task is scheduled in the form of a job to EM. Automated tasks may include modeling applications, backing up databases, deploying custom SQL scripts, or deploying patch updates (see Figure 3). To enable the Grid Control management console to automatically download and apply patches, administrators must first configure Grid Control with a MetaLink[1] login, password, and other credentials. MetaLink is an Oracle Web site that allows registered users to download Oracle database patches and patch sets. EM can also be configured with Simple Mail Transport Protocol (SMTP) to send the logs and alerts to the system administrator via the management console.
- **System monitoring and diagnosis:** EM can be used for continuous monitoring of targets such as hosts, databases, and

[1] The MetaLink Web site is available at metalink.oracle.com.

application servers, and can be configured to send automated alerts and warnings when the target resource utilization reaches threshold values established by the system administrator.

Grid Control installation and configuration

This article describes the installation and configuration of Grid Control on the Red Hat® Enterprise Linux 3 operating system (OS). Best practices dictate installing Grid Control on a system that does not have an Oracle database installed and is not being used as an application server.[2]

Download the three Grid Control gzip files and unzip them to a directory to which the Oracle users have read and write privileges. Unzip the files and obtain the `cpio` archives:

```
# gzip -d linux_grid_control_10_1_0_2_Disk1.cpio.gz
# gzip -d linux_grid_control_10_1_0_2_Disk2.cpio.gz
# gzip -d linux_grid_control_10_1_0_2_Disk3.cpio.gz
```

After unzipping the files and obtaining the `cpio` archives, copy the files from the archive using the `cpio` command:

```
# cpio -idmv < linux_grid_control_10_1_0_2_Disk1.cpio
# cpio -idmv < linux_grid_control_10_1_0_2_Disk2.cpio
# cpio -idmv < linux_grid_control_10_1_0_2_Disk3.cpio
```

The preceding steps create three directories called Disk1, Disk2, and Disk3. The administrator should open an xterm session and execute the command `runInstaller` from within the Disk1 directory to start the Grid Control installer.

Installation of management service. The management service installation includes the management service and management repository. For a new installation, administrators should select the option "Enterprise Manager 10*g* Grid Control using a new database." If the environment already has an Oracle database to house the management repository, administrators should select the option "Enterprise Manager 10*g* Grid Control using an existing database." Administrators can then complete either installation by following the on-screen instructions. After installation of the management service, the Grid Control management console is accessible from the browser at *HostName*:7777/em.

Installation of management agent. The management agent must be installed on all the nodes that require monitoring. Administrators can select the option "Additional management agent" from the Product Selection menu during installation. The management agent requires credentials such as host name and port number of the server that is running the management service. After specifying these details, administrators should follow the on-screen instructions to complete the installation. After installation, the management agent should automatically communicate with the server running the management service and appear as a target in the Grid Control management console.

Management of database cluster nodes using Grid Control

Grid Control has features to help database administrators manage multiple database instances. Using Grid Control, an administrator can dynamically modify parameters of the database instances, such as buffer size and sort area, depending on the workload on the systems. Grid Control provides a console that displays real-time data for server activities. If an excessive load develops on a particular group of instances, the administrator can manually add an extra system and create another database instance without affecting the other systems.[3]

Cloning. Grid Control simplifies database cloning. Cloning creates a copy of the existing database on a new set of target servers and helps create test instances or backup instances of a production database. A Grid Control feature known as multicast allows the administrator to select an `$ORACLE_HOME` on a system and to clone the existing `$ORACLE_HOME` to multiple destination servers. Environment-specific parameters such as host name and IP address are automatically updated as part of the cloning process. The only requirement is that target servers must have the Grid Control management agent up and running. The entire cloning operation can be scheduled using the built-in EM job scheduler. The sidebar "Cloning a database with Grid Control" in this article provides an example scenario of Grid Control cloning capabilities.

Automated upgrades and patching. Grid Control can help automate the upgrade process for database instances. The critical patch advisory feature proactively notifies the administrator of important patches that are available on metalink.oracle.com, which is the patch repository for Oracle products. Grid Control also maintains an inventory of the different hosts as well as the patch sets and interim patches that administrators have applied to them. When configured properly, the critical patch advisory feature of EM can download the relevant patches from MetaLink and maintain them in a patch cache. The patch cache is a staging area for patches, which can be used to apply patches to other systems at a later time. Patch installation for different systems can be scheduled for a particular date and time using EM job scheduling mechanisms.

The critical patch advisory feature also checks for sufficient disk space on systems before applying the patches, and issues an alert when failures occur. Pre- and post-patch processing steps can also

[2] For more detailed installation information, review the section "Preinstallation Requirements for Enterprise Manager" in the *Oracle Enterprise Manager Grid Control Installation and Basic Configuration Guide,* which is available at download-east.oracle.com/docs/html/B12012_03/toc.htm.

[3] For more information about real-time diagnostics and tuning of Oracle database instances using EM Grid Control, see "Exploiting Automated Database Diagnostics in Oracle Database 10*g*" by Ramesh Rajagopalan, Uday Datta Shet, C.S. Prasanna Nanda, and Bharat Sajnani in *Dell Power Solutions,* February 2005.

be automated by writing custom scripts that are invoked by the critical patch advisory feature during patch deployment.

Monitoring of applications and service-level management. In addition to the database, Grid Control monitors mid-tier software and applications. The tool provides a holistic view of the systems on which the management agents are running, which helps administrators proactively monitor systems and anticipate any major issues. Grid Control also provides features that allow administrators to monitor and record critical application-specific business transactions that may be relevant for system tuning.

The Policy Manager feature in Grid Control enables administrators to set an enterprise-wide set of policies to be enforced on all database instances. Some examples of policies include:

- Disabling initialization parameters on undocumented or deprecated systems
- Enabling the use of an OS password file for authentication

Grid Control also enables administrators to create a system baseline. After setting up the system and ensuring that the behavior of the system is optimal, administrators can record a set of baseline metrics for the system. These baseline metrics can be used to determine the optimal values for various system thresholds (such as CPU and memory utilization, available file system space, and response time). An alert displays on the Grid Control management console if the system reaches any of these thresholds. Administrators can alert appropriate IT personnel by configuring relevant notification mechanisms using SMTP traps, e-mail, or user-defined scripts.

Automation for effective, scalable database management

Database administrators generally manage multiple databases, and database deployment consumes much of their available time. Oracle Enterprise Manager Grid Control is advanced software that provides features to help manage and deploy multiple database instances and Oracle RAC clusters. Enterprise Manager Grid Control can help automate and schedule many of the routine activities involved in database management, enabling database administrators to manage and scale out to a large number of database instances efficiently.

Rudramuni B is an engineering manager for the Enterprise Solutions and OS Engineering teams in the Dell Product Group in Bangalore, India. He has an M.Tech. in Systems Analysis and Computer Applications from Karnataka Regional Engineering College in Surathkal, India.

C.S. Prasanna Nanda is an engineering analyst on the Dell Database and Applications team in the Dell Product Group. He has a B.E. in Computer Science from the Birla Institute of Technology and Science in Pilani, India.

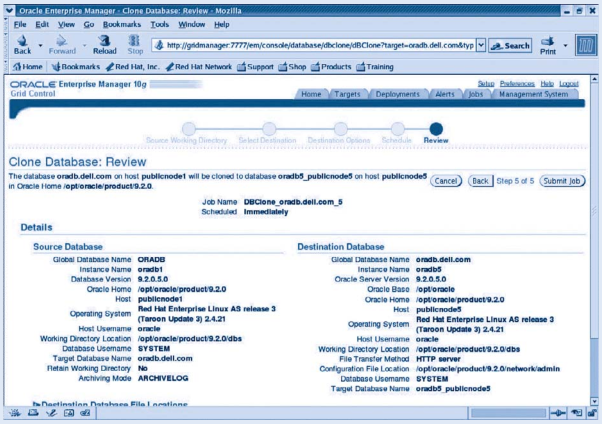
Uday Datta Shet is a senior engineering analyst on the Dell Database and Applications team in the Dell Product Group. He has a B.E. in Computer Science and is also an Oracle Certified Professional (OCP).

CLONING A DATABASE WITH GRID CONTROL

In this example scenario, an administrator has configured four Dell™ PowerEdge™ servers running an Oracle9*i* RAC database. The administrator has purchased two new PowerEdge servers and plans to scale out the existing infrastructure in response to an increase in workload.

Using Oracle Enterprise Manager Grid Control, the administrator can migrate the existing Oracle RAC database to the new nodes with just a few clicks. After installing the operating system and Oracle9*i* RAC database on the two additional server nodes, the administrator installs the Grid Control agent on both nodes. Once the two nodes start communicating with the server running Grid Control, the administrator performs the following six simple steps:

1. **Select the source database.** On the Grid Control home page Deployments tab, click "Clone Database." Then select the running instance of the database to clone.
2. **Supply the credentials.** Enter the username, password, role, and Oracle user credentials to log in to the target database.
3. **Select the destination.** Supply the global database name of the destination host and the destination host name.
4. **Create the schedule.** Determine a schedule to run the database clone job. This job can be scheduled immediately or can be scheduled to run later.
5. **Review.** After all the necessary parameters have been supplied, Grid Control displays a summary of the job (see Figure A).
6. **Submit.** After reviewing the job settings in the preceding step, submit the job for execution. The job will start executing at the start time set in step 4. Administrators can monitor a running job by viewing the Status column on the Jobs tab.



http://gridmanager:7777/em/console/database/dbclone/dBClone?target=oradb.dell.com&typ

Clone Database: Review

The database **oradb.dell.com** on host **publicnode1** will be cloned to database **oradb5_publicnode5** on host **publicnode5** in Oracle Home **/opt/oracle/product/9.2.0**.

Job Name **DBClone_oradb.dell.com_5**
Scheduled **Immediately**

Details

Source Database

Global Database Name	ORADB
Instance Name	oradb1
Database Version	9.2.0.5.0
Oracle Home	/opt/oracle/product/9.2.0
Host	publicnode1
Operating System	Red Hat Enterprise Linux AS release 3 (Taroon Update 3) 2.4.21
Host Username	oracle
Working Directory Location	/opt/oracle/product/9.2.0/dbs
Database Username	SYSTEM
Target Database Name	oradb.dell.com
Retain Working Directory	No
Archiving Mode	ARCHIVELOG

Destination Database

Global Database Name	oradb.dell.com
Instance Name	oradb5
Oracle Server Version	9.2.0.5.0
Oracle Base	/opt/oracle
Oracle Home	/opt/oracle/product/9.2.0
Host	publicnode5
Operating System	Red Hat Enterprise Linux AS release 3 (Taroon Update 3) 2.4.21
Host Username	oracle
Working Directory Location	/opt/oracle/product/9.2.0/dbs
File Transfer Method	HTTP server
Configuration File Location	/opt/oracle/product/9.2.0/network/admin
Database Username	SYSTEM
Target Database Name	oradb5_publicnode5

Destination Database File Locations

Figure A. Database cloning

Exploiting Automated Database Diagnostics in Oracle Database 10*g*

The task of monitoring and managing database system performance has become increasingly important as enterprise applications grow in complexity. This article presents a Dell study that demonstrates how administrators can use Automatic Database Diagnostic Monitor, a feature of Oracle® Database 10g, to monitor and manage the performance of Oracle Real Application Clusters 10g.

BY RAMESH RAJAGOPALAN, UDAY DATTA SHET, C.S. PRASANNA NANDA, AND BHARAT SAJNANI

A key management feature in Oracle Database 10*g* is Automatic Database Diagnostic Monitor (ADDM). ADDM proactively monitors the state of the production database environment, reporting on performance issues and recommending potential solutions to those issues. ADDM is designed to perform a top-down analysis of the discovered problem, resulting in a set of findings that includes the root cause of the problem and suggestions for resolution.

This article provides an overview of a common scenario for automating management processes using ADDM. In August and September 2004, through a controlled laboratory simulation of a midsized workload on Oracle Database 10*g* that used out-of-the-box settings, Dell engineers demonstrated how the self-diagnostic features of the database engine could help identify performance issues. This article analyzes some of the recommendations that ADDM provided in response to the performance issues discovered during the test.

Configuration of the test environment

The test environment comprised the Oracle Database 10*g* Enterprise Edition 10.0.2.0 database server running on two Dell™ PowerEdge™ 6600 servers, each with dual Intel® Xeon™ processors at 1.9 GHz, 1 MB level 2 (L2) cache, and 4 GB physical memory. The servers ran the Red Hat® Enterprise Linux® 3 Update 2 operating system (OS). For storage, the test team configured a Dell PowerVault™ 22xS SCSI enclosure with ten 36 GB 15,000 rpm disks.

Order-entry processing workload simulation

The study used an order-entry processing workload. The database was about 5 GB, and the largest table in the database had 15 million rows. The second largest table had 1.5 million rows. The test team used a driver program to simulate 25 concurrent users who executed five different transactions against the database such as select, insert, delete, and update operations. The workload exhibited online transaction processing (OLTP) characteristics—that is, short-duration transactions involving random and small-block reads and writes. Dell engineers executed the workload for one hour, which allowed two to three snapshots to be gathered for analysis during a steady state.

Automatic Workload Repository functionality

The Oracle database engine is designed to accumulate statistics such as Wait Events and Time Model Statistics in internal database tables—a capability known as dynamic performance views. However, the values in dynamic performance views are reset on instance startup and, therefore, performance data that could be used for comparisons is lost on startup. Automatic Workload Repository (AWR) is a feature of Oracle Database 10*g* that automatically stores the cumulative and delta values for a majority of the statistics in persistent storage to aid in proactive analysis of performance issues. The collection of a single data set is called an AWR snapshot. A pair or a sequence of snapshots serves as a baseline, which can be compared with snapshots captured during the occurrences of a performance issue to help isolate the issue.

> ADDM proactively monitors the state of the production database environment, reporting on performance issues and recommending potential solutions to those issues.

In studying the overall test environment, time was chosen as the common unit of measurement. The parameter `db_time` was the most important statistic. The `db_time` parameter represented the total time spent by an application in the database as it executed transactions. This parameter was the sum of CPU and wait times of all sessions, excluding idle user sessions. A major goal of tuning an Oracle system is to minimize bottlenecks such as inefficient SQL statements, which can potentially decrease performance and throughput.

AWR captures database statistics as well as system statistics. By default, it automatically generates snapshots of the database performance statistics once every hour (this is the snapshot interval) and retains the data for seven days (this is the snapshot retention). The degree of statistics collection is controlled by the initialization parameter `statistics_level`. By default, it is set to `typical`, which is sufficient for most cases.

To access AWR from Oracle Enterprise Manager 10*g* management software, administrators should perform the following steps:

1. From the Enterprise Manager startup screen, select the Targets tab.
2. Choose the Databases tab, and select the desired database.
3. Click "Administration."
4. Click "Automatic Workload Repository" to open the Automatic Workload Repository page.

From the Automatic Workload Repository page, administrators can manage snapshots or modify AWR settings. Configurable parameters for AWR are snapshot retention, snapshot interval, and collection level.

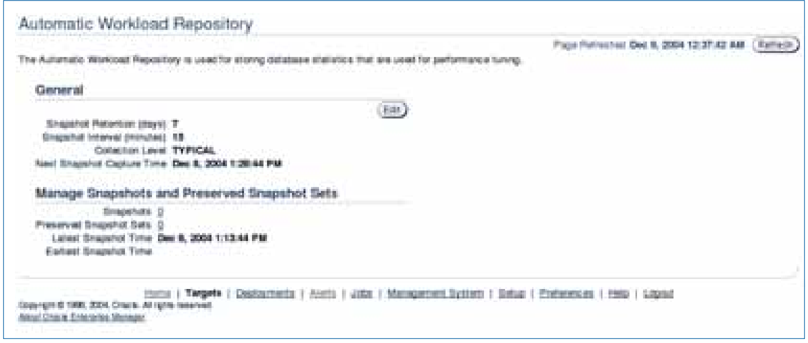
Figure 1 displays the current settings for these three configurable parameters. Administrators can change the settings by pressing the Edit button. The Automatic Workload Repository page also displays information about the number of snapshots available and the time of the earliest and most recent snapshots since the database instance was started. For the study discussed in this article, Dell engineers set the snapshot interval to 15 minutes and used the default settings for the other two parameters.

Automatic Database Diagnostic Monitor functionality

ADDM periodically analyzes the AWR data, locates the root causes of performance problems, and provides recommendations for resolving the problems. It also identifies non-problem areas of the systems running Oracle Database 10*g* to help administrators avoid misdiagnoses. Oracle Database 10*g* performs an ADDM analysis every time AWR takes a snapshot.

Even with ADDM, tuning is an iterative process; fixing one problem can cause another to arise. Administrators may find they require multiple cycles to achieve acceptable performance levels. ADDM enables proactive system monitoring, however, which enables administrators to take corrective action on problems efficiently and accurately. The types of issues that ADDM helps diagnose include the following:

- CPU bottlenecks
- Size of Oracle Database 10*g* memory structures
- Lack of I/O capacity
- Oracle Real Application Clusters (RAC)–related considerations such as hot global cache blocks and interconnect latency
- Lack of concurrency
- Application performance
- Lack of optimal configuration
- High PL/SQL execution time



Automatic Workload Repository

Page Refreshed Dec 8, 2004 12:37:42 AM

The Automatic Workload Repository is used for storing database statistics that are used for performance tuning.

General

Snapshot Retention (days): 7
Snapshot Interval (minutes): 15
Collection Level: TYPICAL
Next Snapshot Capture Time: Dec 8, 2004 1:28:44 PM

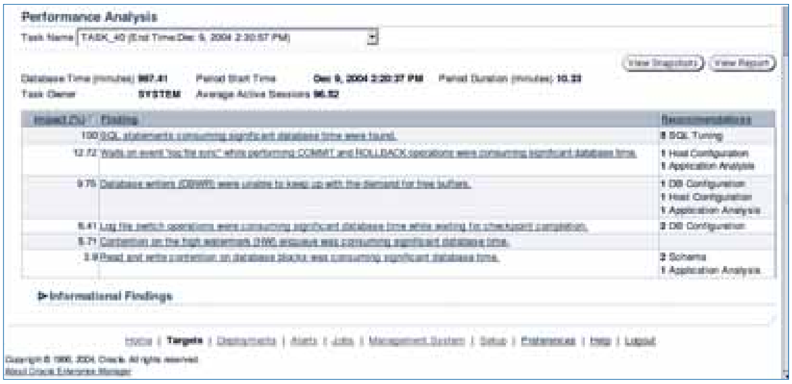
Manage Snapshots and Preserved Snapshot Sets

Snapshots: 0
Preserved Snapshot Sets: 0
Latest Snapshot Time: Dec 8, 2004 1:13:44 PM
Earliest Snapshot Time

Copyright © 1996, 2004, Oracle. All rights reserved.
About Oracle Enterprise Manager

Refresh
Edit
Home | Targets | Deployments | Alerts | Jobs | Management System | Setup | Preferences | Help | Logout

Figure 1. Automatic Workload Repository page



Performance Analysis

Task Name TASK_40 (End Time:Dec 9, 2004 2:30:57 PM)

Database Time (minutes) 987.41 Period Start Time Dec 9, 2004 2:20:37 PM Period Duration (minutes) 10.33
Task Owner SYSTEM Average Active Sessions 95.52

Impact (%)	Finding	Recommendations
100	SQL statements consuming significant database time were found.	8 SQL Tuning
12.72	Waits on event "log file sync" while performing COMMIT and ROLLBACK operations were consuming significant database time.	1 Host Configuration 1 Application Analysis
9.75	Database writers (DBWR) were unable to keep up with the demand for free buffers.	1 DB Configuration 1 Host Configuration 1 Application Analysis
6.41	Log file switch operations were consuming significant database time while waiting for checkpoint completion.	2 DB Configuration
5.71	Contention on the high watermark (HW) enqueue was consuming significant database time.	
3.9	Read and write contention on database blocks was consuming significant database time.	2 Schema 1 Application Analysis

▶ Informational Findings

Home | Targets | Deployments | Alerts | Jobs | Management System | Setup | Preferences | Help | Logout

Copyright © 1996, 2004, Oracle. All rights reserved.
About Oracle Enterprise Manager

Figure 2. Performance Analysis page

ADDM analysis reports include the following components:

- Problem (quantified by db_time)
- Symptom
- Recommendation (action and rationale)
- Informational findings (symptoms that do not particularly affect performance)

ADDM is enabled by default, and depends on the performance data gathered by AWR. ADDM I/O performance analysis is driven by the initialization parameter dbio_expected, which specifies the average time in microseconds required to perform a physical database read.

To access ADDM from Oracle Enterprise Manager 10*g*, administrators should perform the following steps:

1. From the Enterprise Manager startup screen, select the Targets tab.
2. Choose the Databases tab, and select the desired database.
3. Click "Advisor Central."
4. Click "ADDM" to open the Create ADDM Task page.

The Create ADDM Task page lets administrators create an ADDM task by specifying an appropriate start time and end time and clicking the OK button. Doing so displays a Performance Analysis page, which shows the performance analysis and informational findings.

Figure 2 shows the findings detected by ADDM when Dell engineers executed the workload. Several of these findings addressed performance. The informational findings, also shown in Figure 2, provided data regarding non-problems.

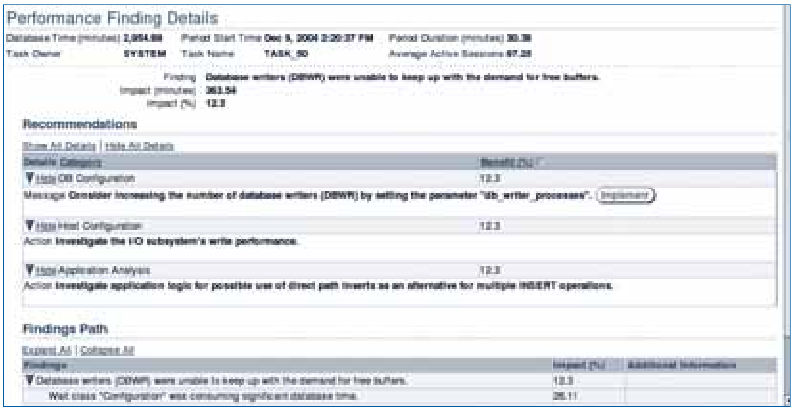
For performance issues, ADDM provides an estimated impact and summary of recommendations, which are classified into categories such as SQL Tuning, DB Configuration, Host Configuration, and so on.

Figure 3 shows details and recommendations for one of the performance findings for the multiuser workload: "Database writers (DBWR) were unable to keep up with the demand for free buffers."

Because the OLTP workload incurred a mix of read and write transactions, the test team expected the database buffer cache to become full rapidly. The database_cache parameter was set to 1.2 GB and the size of the database was about 5 GB. The finding "Database writers (DBWR) were unable to keep up with the demand for free buffers" addressed the insufficient buffer cache issue. Because the test workload was write intensive, the number of dirty buffers in the cache increased rapidly as the test team executed the workload.

The first recommendation provided by ADDM suggested increasing the number of db_writer_processes so that the dirty buffers could be flushed and freed up for reuse. The second recommendation pointed out potential limitations of the I/O subsystem. Following the second recommendation, administrators could check the Memory Adviser for buffer cache usage or review the Automatic Storage Management (ASM) performance monitor to quickly determine which component was the bottleneck, and then take corrective action. Both the Memory Adviser and ASM are features of Oracle Enterprise Manager 10*g*. The third recommendation indicated that administrators should examine the application logic to optimize the way the records are inserted.

For another test finding, "Log file switch operations were consuming significant database time while waiting for checkpoint completion," ADDM provided a specific recommendation: increase the size of the log files to 1221 MB to hold at least 20 minutes of redo information. Using this recommendation, administrators could tune the log files accordingly without having to go through



Performance Finding Details

Database Time (minutes) 2,954.88 Period Start Time Dec 9, 2004 2:20:37 PM Period Duration (minutes) 30.38
Task Owner SYSTEM Task Name TASK_50 Average Active Sessions 97.28

Finding Database writers (DBWR) were unable to keep up with the demand for free buffers.
Impact (minutes) 363.54
Impact (%) 12.3

Recommendations

Details Category	Benefit (%)
▼Hide DB Configuration	12.3
Message Consider increasing the number of database writers (DBWR) by setting the parameter "db_writer_processes".	
▼Hide Host Configuration	12.3
Action Investigate the I/O subsystem's write performance.	
▼Hide Application Analysis	12.3
Action Investigate application logic for possible use of direct path inserts as an alternative for multiple INSERT operations.	

Findings Path

Findings	Impact (%)	Additional Information
▼Database writers (DBWR) were unable to keep up with the demand for free buffers.	12.3	
Wait class "Configuration" was consuming significant database time.	26.11	

Figure 3. Performance Finding Details page

an iterative process to determine the optimal value for the size of the log files.

For the finding, “Waits on event ‘log file sync’ while performing COMMIT and ROLLBACK operations were consuming significant database time,” ADDM recommended that administrators investigate application logic for possible reduction in the number of COMMIT operations by increasing the size of transactions, and also investigate the possibility of improving the I/O performance to the online redo log files.

ADDM also provides information to point administrators to the application-level SQL statement corresponding to a given finding. For example, clicking on the findings for which the recommendation category is “SQL Tuning” provides the details of the statements that need to be tuned. The projected impact and benefit of tuning the statement are provided as percentage values. ADDM also displays the Explain Plan for the SQL statement. The Explain Plan is the data access path chosen by the database server for executing a SQL query. These details can help database administrators and application developers to optimize the SQL code. Administrators can obtain similar details using SQL (see the “Using SQL to create an AWR or ADDM report” sidebar in this article).

ADDM also points administrators to the application-level SQL statement corresponding to a given finding.

USING SQL TO CREATE AN AWR OR ADDM REPORT

The AWR report provides a summary of database events and statistics for a given snapshot interval. AWR reports can be obtained by running `awrrpt.sql` as a user with database administrator privileges:

```
SQL> @$ORACLE_HOME/rdbms/admin/awrrpt
```

The preceding command prompts the administrator with options for generating the report in either HTML or plain text. The command also prompts the administrator with the list of snapshot IDs. The administrator should enter the beginning and ending snapshot IDs, and either accept the default file name for the report or specify an alternate name.

SQL can also generate an ADDM report through the following command:

```
SQL> @$ORACLE_HOME/rdbms/admin/addmrpt
```

Proactive monitoring of database servers

Using a midsized, multiuser workload on Dell PowerEdge servers running Oracle Database 10*g*, Dell engineers demonstrated how the AWR and ADDM features of Oracle Database 10*g* are designed to work together to proactively monitor the health of the database server and provide recommendations and information to help administrators address performance issues. The self-diagnostic features of Oracle Database 10*g* can help reduce total cost of ownership by proactively monitoring the performance of the database server and significantly lightening the burden on database administrators who manage increasingly complex and global database environments.

Acknowledgments

The authors would like to thank their colleagues Paul Rad and Anthony Fernandez for their valuable comments and help.

Ramesh Rajagopalan is a lead software engineer on the Database and Applications Engineering team of the Dell Product Group. His current areas of focus include Oracle RAC solutions and performance analysis of Dell cluster solutions. Ramesh has a B.E. in Computer Science from the Indian Institute of Science in Bangalore, India.

Uday Datta Shet is a senior engineering analyst on the Database and Applications Engineering team of the Dell Product Group. His current interests include databases, operating systems, and clustering. Uday has a B.E. in Computer Science and is also an Oracle Certified Professional (OCP).

C.S. Prasanna Nanda is an engineering analyst on the Database and Applications team of the Dell Product Group. His primary areas of interest include databases and the Linux operating system. Nanda has a B.E. in Computer Science from the Birla Institute of Technology and Science in Pilani, India.

Bharat Sajnani is a systems engineer on the Database and Applications team of the Dell Product Group. His primary areas of interest include enterprise software and applications as well as Linux-based distributed systems. Bharat has a B.S. in Computer Engineering and a master’s degree in Computer Engineering from The University of Texas at Austin.

FOR MORE INFORMATION

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

Dell and Oracle supported configurations:
www.dell.com/oracle

Migrating Oracle Database 10*g*

from Sun Servers to Dell Servers

Oracle® Database 10*g*, the latest version of Oracle Database, is designed to make database migration from one platform to another easy. To better understand this process, a team of Dell engineers migrated a 100 GB Oracle database from a Sun server running the Sun Solaris operating system to a Dell™ PowerEdge™ server running either the Microsoft® Windows Server™ 2003 operating system (OS) or the Red Hat® Enterprise Linux® OS. This article explains the details and best practices involved in successful Oracle database migrations.

BY TODD MUIRHEAD; DAVE JAFFE, PH.D.; AND PAUL RAD

To improve the process of migrating data from one database to another or from one platform to another, two key Oracle tools for data migration—Data Pump and Transportable Tablespaces—were enhanced in Oracle Database 10*g*. Data Pump is an improved version of the Oracle Import and Oracle Export utilities from previous versions of Oracle Database. The improvements are designed to add capabilities and decrease the time required for exporting and importing Oracle Database 10*g* data.

While Data Pump exports the Oracle data into another form for migration, the Transportable Tablespaces feature allows for the actual data files to be copied or transported to another system and then simply attaches them to the target database. In previous releases of Oracle Database, the Transportable Tablespaces feature allowed a tablespace to be moved across Oracle databases running on the same processor architecture. Transportable Tablespaces in Oracle Database 10*g* enables data to be moved across different processor architectures. This feature can facilitate data migration from one hardware platform to another.

This article examines testing performed in June 2004 by a team of Dell engineers, in which a 100 GB Oracle database was migrated from a Sun Fire V440 server running Sun Solaris to a Dell PowerEdge 6650 server running either Microsoft Windows Server 2003 or Red Hat Enterprise Linux. Data Pump was used to easily export the database from one Oracle Database 10*g* instance running on the Sun server and import it into another Oracle Database 10*g* instance running on the Dell PowerEdge server. In this study, Transportable Tablespaces was used to migrate data from the Sun Fire V440 server—which used UltraSPARC IIIi processors—to the Dell PowerEdge 6650 server, which used Intel® Xeon™ processors MP. With previous versions of Oracle, it would not have been possible to use Transportable Tablespaces to accomplish this migration.

	Sun Fire V440	Dell PowerEdge 6650
Operating system	Solaris 9 12/03	Microsoft Windows Server 2003, Enterprise Edition, or Red Hat Enterprise Linux AS 3
CPU	Four UltraSPARC IIIi processors at 1.28 GHz and with 1 MB of L2 cache	Four Intel Xeon processors MP at 3.0 GHz and with 4 MB of L3 cache
Memory	16 GB	16 GB (16 × 1 GB dual in-line memory modules, or DIMMs)
Internal disks**	Four 73 GB 10,000 rpm Ultra320 SCSI	Four 73 GB 10,000 rpm Ultra320 SCSI
NICs	Two 10/100/1000 Mbps (internal)	Two 10/100/1000* Mbps (internal)
Disk controller	On-board SCSI	PowerEdge Expandable RAID Controller 3, Dual Channel (PERC 3/DC)
Fibre Channel host bus adapter	Two QLogic QLA2340	Two QLogic QLA2340
Fibre Channel storage	34 GB to 73 GB 10,000 rpm disks on Dell/EMC CX600	34 GB to 73 GB 10,000 rpm disks on Dell/EMC CX600

*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.

**For hard drives, GB means 1 billion bytes; actual capacity varies with preloaded material and operating environment, and will be less.

Figure 1. Sun Fire V440 and Dell PowerEdge 6650 server configurations

Configuring the hardware and software

To test migration from Sun to Dell, the Dell team installed and configured a Sun Fire V440 server and a Dell PowerEdge 6650 server. As shown in Figure 1, both servers were configured with four processors and 16 GB of RAM as well as with several internal drives and on-board network interface cards (NICs).

A Dell/EMC CX600 storage array provided the storage for the database and was connected to both Sun and Dell servers via a storage area network (SAN). Each server was assigned three 10-disk RAID-10 logical storage units (LUNs) to be used for data and two 2-disk RAID-1 LUNs to be used for logs. An additional 5-disk RAID-5 LUN was also assigned to each server for data staging during migration and initial database creation.

Both servers were configured with current operating systems and Oracle Database 10*g*. Oracle's Universal Installer tool can be used on both Sun and Dell servers to install Oracle Database 10*g*, enabling administrators to select the same options during installation on different platforms.

Solaris 9 12/03 was installed on the Sun Fire V440 server using Sun Factory JumpStart. Two installations were performed on the same PowerEdge 6650 server to test two operating systems: Red Hat Enterprise Linux AS 3 and Microsoft Windows Server 2003, Enterprise Edition. Both were installed using Dell OpenManage™ Server Assistant software. Oracle Database 10*g* installation guides and release notes for the respective platforms were then used to verify that the correct software levels were used and that the necessary settings were changed.

The new Automatic Storage Manager (ASM) feature of Oracle Database 10*g* was used during the creation of the initial database on both Sun and Dell servers to improve the performance and management of the data files. The three large RAID-10 LUNs were assigned in ASM to be part of a single file group for the data files, while each of the RAID-1 LUNs was assigned to a separate ASM file group for logs. An initial database instance was created on each of the servers using the Oracle Database Creation Assistant. ASM was used for the storage on both instances.

Transportable Tablespaces in Oracle Database 10g enables data to be moved across different processor architectures.

The 100 GB DVD store (DS) database schema was loaded onto the Sun system. It comprised four main tables and one other table (see Figure 2).

The Customers table was prepopulated with 200 million customers. The Orders table was prepopulated with 10 million orders per month for nine months, the Orderlines table was prepopulated with an average of five items per order, and the Products table contained 1 million DVD titles. In addition, the Categories table listed 16 DVD categories.

Using Data Pump to prepare the migration

Data Pump export and Data Pump import are additional utilities in Oracle Database 10*g*, but have a similar look-and-feel to the original export and import utilities. Data Pump export is a utility for unloading data and metadata from the database into a set of operating system files called dump file sets. A dump file set can be moved to another system and loaded by the Data Pump import utility. Data Pump import is used to load the metadata and data stored in an export dump file set into the database on the target system. Three methods can be used to interact with Data Pump

Table	Columns	Number of rows
Customers	CUSTOMERID, FIRSTNAME, LASTNAME, ADDRESS1, ADDRESS2, CITY, STATE, ZIP, COUNTRY, REGION, EMAIL, PHONE, CREDITCARD, CREDITCARDEXPIRATION, USERNAME, PASSWORD, AGE, INCOME, GENDER	200 million
Orders	ORDERID, ORDERDATE, CUSTOMERID, NETAMOUNT, TAX, TOTALAMOUNT	90 million
Orderlines	ORDERLINEID, ORDERID, PROD_ID, QUANTITY, ORDERDATE	450 million
Products	PROD_ID, CATEGORY, TITLE, ACTOR, PRICE, QUAN_IN_STOCK, SPECIAL	1 million
Categories	CATEGORY, CATEGORYNAME	16

Figure 2. DVD store database schema

export and import utilities: command line, parameter file, and interactive command mode. Figure 3 shows the process of moving data between servers using Data Pump.

Migration steps for Data Pump tool

The migration of the database using the Data Pump export and Data Pump import tools comprises three steps: export the data into a dump file on the source server with the `expdp` command; copy or move the dump file to the target server; and import the dump file into Oracle on the target server by using the `impdp` command.

The Data Pump export command (`expdp`) is used to export the data. Data Pump import and export require that the location of the dump file be specified via a database directory object. The directory object data_pump_dir must be created on both the source and target systems before running the export or import command from a SQL*Plus® session:

```
SQL> create directory data_pump_dir as '/data/
    expdata';
```

The data that the team needed to move resided entirely in the DS schema, which dictated that the export command would be the following:

```
expdp system/oracle DIRECTORY=data_pump_dir
    DUMPFILE=sundvd.dmp SCHEMAS=DS
```

The export of approximately 100 GB of data completed in 19 minutes and resulted in a 45 GB dump file with the name sundvd.dmp. FTP was used to transfer the exported 45 GB dump file to the target Windows Server 2003 system in 55 minutes and 46 seconds.

To complete the migration, a Data Pump import was performed with the following command from the directory where the dump file was located on the target Windows Server 2003 system and on the target Red Hat Enterprise Linux system:

```
impdp system/oracle SCHEMAS=DS DIRECTORY=data_pump_
    dir DUMPFILE=sundvd.dmp LOGFILE=sunimport.log
```

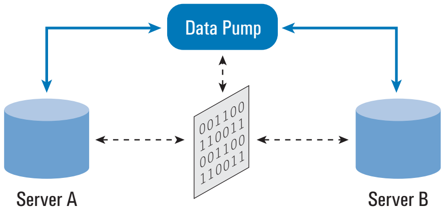


Figure 3. Using Data Pump to export and import data and metadata between databases

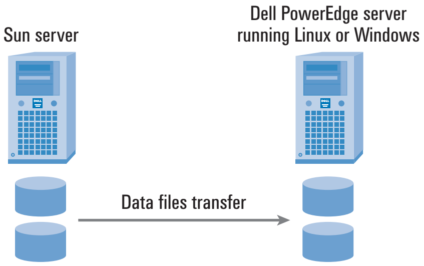


Figure 4. Transporting a tablespace from a Sun server to a Dell server

The command for the Data Pump import on both Windows Server 2003 and Red Hat Enterprise Linux is the same. The creation of the directory object was only slightly different because of the use of drive letters on Windows Server 2003 and the absence of drive letters on Linux when specifying the directory path.

The import was completed in 1 hour and 34 minutes, resulting in a total migration time of 2 hours and 49 minutes. For comparison, an export was performed with the Oracle Export tool on the Sun Fire server and took 4.5 hours.

Using Transportable Tablespaces to complete the migration

With previous releases of Oracle Database, the Transportable Tablespaces feature allowed a tablespace to be moved across Oracle databases running on the same architecture. Transportable Tablespaces in Oracle Database 10*g* allows actual data files to be copied from one database server to another across different operating system platforms (see Figure 4), and then simply plugged into the new database.

The simplicity of Data Pump allows for quick migrations without significant work to prepare for the actual movement of the data.

To transport a tablespace from one platform to another, data files belonging to the tablespace set must be converted to a format that can be understood by the destination database. Although disk structures conform to a common format with Oracle Database 10*g*, the source and target platforms can use different endian formats (byte orders). If the source and target platforms use different endian formats, the RMAN utility's `CONVERT` command can be used to convert byte ordering when transporting to a different endian platform. For platforms that use the same endian format, no conversion is necessary.

The V$TRANSPORTABLE_PLATFORM view can be used to determine whether endian ordering is the same on both platforms.

The following SELECT statement can be used to determine the endian format on both platforms:

```
SELECT endian_format
FROM v$transportable_platform tp, v$database d
WHERE tp.platform_name = d.platform_name;
```

On a Sun Solaris system, the SELECT statement produces the following output:

```
ENDIAN_FORMAT
--------------
Big
```

On a Linux or Windows system, the SELECT statement produces the following output:

```
ENDIAN_FORMAT
--------------
Little
```

The different endian format indicates that a conversion must be performed as part of a migration using Transportable Tablespaces.

Both databases also must use exactly the same character set. Solaris and Linux versions of Oracle Database 10*g* default to the same database character set of WE8ISO8859P1, but Windows must be explicitly set to use this character set during Oracle Database 10*g* installation.

Migration steps for Transportable Tablespaces tool

A migration using Transportable Tablespaces involves three major steps: exporting the metadata and creating a copy of the tablespaces; copying the tablespaces to the target server; and plugging in the copied data files as tablespaces on the target server. To prepare for the export and copy of the tablespace, administrators must check the referential integrity of the tablespaces to be moved by using the following command:

```
SQL> execute DBMS_TTS.TRANSPORT_SET_
   CHECK('CUSTTBS,DS_MISC,INDXTBS,ORDERTBS', TRUE);
SQL> select * FROM transport_set_violations;
```

All of the tablespaces to be moved must be set to read-only. For example:

```
SQL> alter tablespace custtbs read only;
```

A directory object must be created for Data Pump to use during export of the metadata:

```
SQL> create directory data_pump_dir as ‘/data/
   expdata’;
```

The Data Pump command is now ready to be used to generate the metadata export file:

```
expdp system/oracle DUMPFILE=expdat.dmp
   TRANSPORT_TABLESPACES=custtbs,ds_misc,
indxtbs,ordertbs TRANSPORT_FULL_CHECK=Y
```

Because the migration from Solaris to Linux or Solaris to Windows mandates a switch from big endian format to little endian format, Oracle Database 10*g*’s RMAN utility must be used to accomplish the endian change. The other factor to be considered is that ASM does not allow for operating system commands such as a file copy to be used directly against the data files, so an RMAN convert command is also required to get a copy of the data files out of ASM. On the source Sun Fire server, a single RMAN convert command can accomplish both of these requirements:

```
RMAN> convert tablespace
   custtbs,ds_misc,indxtbs,ordertbs
2> TO PLATFORM 'Microsoft Windows IA (32-bit)'
3> FORMAT '/data/expdata/%U';
```

In the Dell test, the endian conversion and copy with ASM, which occurred using the preceding command, completed in 37 minutes and 5 seconds.

At this point, the tablespaces on the source server should be changed back to read/write with the following command:

```
SQL> alter tablespace ordertbs read write;
```

The tablespaces and exported metadata file should now be copied to the target server. Any method of transferring files across the network can be used, but in the Dell test, FTP was used and the transfer of 100 GB of data files was completed in 1 hour and 52 minutes.

The third major phase of the migration with Transportable Tablespaces should begin by using an RMAN convert command to copy the data files into ASM on the target server:

```
RMAN> convert datafile
2> 'm:\expdata\cust_1.dbf',
3> 'm:\expdata\cust_2.dbf',
4> 'm:\expdata\dsmisc.dbf',
5> 'm:\expdata\indx_1.dbf',
6> 'm:\expdata\indx_2.dbf',
7> 'm:\expdata\order_1.dbf',
8> 'm:\expdata\order_2.dbf'
9> db_file_name_convert="m:\expdata\","+DATA1/";
```

In the Dell test, the RMAN `convert` command completed in 31 minutes and 2 seconds. The same command was used on the target Red Hat Enterprise Linux system with the exception of the paths to the data files. The path on the Linux server was `/data/expdata/*.dbf`.

The next steps prepare the database for the insertion of the tablespaces. Administrators should create the user that owns the tablespaces as well as the directory object to be used by the Data Pump import command for importing the metadata and plugging in the tablespaces.

In the Dell test, the following Data Pump import command was used to plug in the tablespaces:

```
impdp system/oracle dumpfile=expdat.dmp
   directory=data_pump_dir
transport_datafiles=+Data1/cust_1.dbf,+Data1/cust_
   2.dbf,
+Data1/dsmisc.dbf,+Data1/indx_1.dbf,+Data1/indx_
   2,+Data1/order_1.dbf, +Data1/order_2.dbf
```

Following the completion of the command, which usually takes less than a minute, a few more steps are required to finish the migration. The default tablespace for the DS user should be set to be the newly transported ds_misc tablespace; the sequences and data types must be created; and then the stored procedures should be created to complete the migration. The preceding three steps are completed by editing the database creation scripts and including only the needed sections. Finally, the tablespaces on the target system should be set to read/write with an `alter tablespace` command for each tablespace. The following is an example of this command:

Using either Data Pump or Transportable Tablespaces can be an effective way to migrate from one platform to another.

```
SQL> alter tablespace ordertbs read write;
```

Moving databases from Sun servers to Dell servers

Using either Data Pump or Transportable Tablespaces can be an effective way to migrate from one platform to another. The simplicity of Data Pump allows for quick migrations without significant work to prepare for the actual movement of the data. Using Transportable Tablespaces potentially requires more work to start and complete the migration but also offers the possibility of a very fast move. Figure 5 summarizes how long it took the Dell test team to move the database from a Sun Fire V440 server to a Dell PowerEdge 6650 server.

	Export	File transfer	Import	Total time
Data Pump	19 minutes	55 minutes and 46 seconds	1 hour and 34 minutes	2 hours and 49 minutes
Transportable Tablespaces	37 minutes and 5 seconds	1 hour and 52 minutes	31 minutes and 2 seconds	3 hours

Figure 5. Time required for Oracle Database 10*g* migration from a Sun Fire V440 server to a Dell PowerEdge 6650 server

In the Dell study, if ASM was not used on the source or target host, then the time needed to use Transportable Tablespaces would basically equal the time needed to convert and copy the data files from one server to another. Even though ASM requires that the RMAN utility be used to move the data files in and out of ASM, the additional benefits that ASM provides in terms of ongoing data file management make it worth the extra step associated with using RMAN during the migration.

Todd Muirhead is an engineering consultant on the Dell Technology Showcase team. He specializes in SANs and database systems. Todd has a B.A. in Computer Science from the University of North Texas and is Microsoft Certified Systems Engineer + Internet (MCSE+I) certified.

Dave Jaffe, Ph.D., is a senior consultant on the Dell Technology Showcase team who specializes in cross-platform solutions. Previously, he worked in the Dell Server Performance Lab, where he led the team responsible for Transaction Processing Performance Council (TPC) benchmarks. Before Dell, Dave spent 14 years at IBM in semiconductor processing, modeling, and testing, and in server and workstation performance. He has a Ph.D. in Chemistry from the University of California, San Diego, and a B.S. in Chemistry from Yale University.

Paul Rad is a senior software engineer in the Dell Database and Application Engineering Department of the Dell Product Group. Paul's current interests include operating systems, database systems, clustering, storage technologies, and virtualization. Paul has master's degrees in both Computer Science and Computer Engineering from The University of Texas at San Antonio.

FOR MORE INFORMATION

Dell and Oracle supported configurations:
www.dell.com/oracle

Oracle database products:
www.oracle.com/database

Microsoft Exchange Server 2003 Scale-Out Performance

on a Dell PowerEdge High-Availability Cluster

Microsoft® Exchange Server 2003 clusters built with industry-standard Dell™ PowerEdge™ servers and Dell/EMC storage can provide high-availability messaging services by enabling applications to fail over from one cluster node to another with minimal downtime. This article provides an overview of Exchange Server 2003 scale-out performance on a Dell PowerEdge cluster in a 3+1 active/passive configuration running the Microsoft Windows Server™ 2003, Enterprise Edition, operating system.

BY ARRIAN MEHIS, ANANDA SANKARAN, AND SCOTT STANFORD

Organizations seeking to achieve stable service levels for their business-critical applications can turn to high-availability server clusters. The primary benefits of this clustering approach include high availability, scalability, and manageability. High availability can be provided by enabling applications to fail over from one cluster node to another. Scalability can be achieved by incrementally adding nodes to the cluster. Manageability can be enhanced through the ability to perform maintenance tasks—such as operating system (OS) upgrades, patches, and service pack installations—on the server nodes in the cluster without incurring downtime. These benefits can be important considerations for messaging installations deployed on Microsoft Exchange Server 2003.

Understanding clustering capabilities in Windows Server 2003

The Microsoft Windows Server 2003 cluster service supports up to eight server nodes. Also referred to as Microsoft Cluster Service (MSCS), this service follows a shared-nothing model wherein common cluster resources such as shared drives are selectively owned by only one cluster server (node) at a time. The resource can be moved to another node when the owning node fails or must undergo maintenance. The cluster nodes are interconnected via a private local area network (LAN) and send a periodic signal—referred to as a heartbeat—to determine whether each node is functional. The cluster requires a shared storage system to share disk resources and to host the quorum resource, which is the cluster's shared database. The shared storage system should be SCSI-based because the cluster uses the SCSI protocol to manage the shared disk drives. Figure 1 illustrates the physical architecture of a typical MSCS-based high-availability cluster.

The application services running on a server cluster are exposed to clients as virtual servers. The virtual server can be hosted on any node in the cluster and appears as a physical server to the client. If virtual servers are

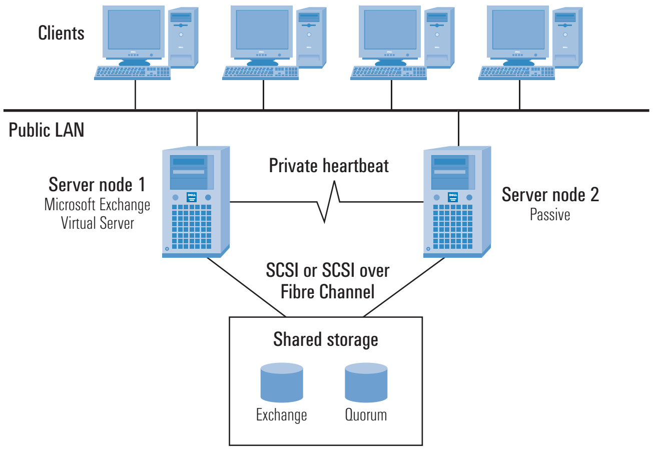


Figure 1. Architecture of an MSCS-based high-availability cluster

moved across cluster nodes because of failures or maintenance, the clients will not be aware of which physical node is hosting any given virtual server. All application resources, including the virtual server resources, are grouped into a resource group. The resource group serves as the unit of failover within a cluster. *Failover* is the process of moving an application (resource group) from one cluster node to another. Server clusters can exist in both active/active and active/passive configurations. In an active/active configuration, all the server nodes in the cluster are active and run virtual application servers. In an active/passive ($N + 1$) configuration, one or more nodes in the cluster are left idle to take over the applications of the active nodes if a failure occurs. Exchange Server 2003 supports up to seven Exchange Virtual Servers on a Windows Server 2003 cluster.

Examining the components of Dell PowerEdge Fibre Channel clusters

High-availability clustering helps ensure that the physical server hosting an application is not a single point of failure by enabling that application to be restarted on one of several other servers. Dell's high-availability cluster solutions are designed and tested to help ensure that no single point of failure exists. Each cluster component—server, storage subsystem, and OS—acts as a module that can be integrated seamlessly with other modules to form an end-to-end cluster solution (see Figure 2). Dell high-availability cluster solutions are built with industry-standard, off-the-shelf components, including the stand-alone server, storage system with fabric switches and host bus adapters (HBAs), OS, cluster service, and applications. Dell PowerEdge servers have been designed with a broad array of redundant features to maximize availability.

The shared storage required for Fibre Channel clustering can be configured as a storage area network (SAN) attached via Fibre Channel switches or it can be attached directly to the cluster servers. Dual redundant switches are required in SAN configurations to avoid a single point of failure. Each clustered server has dual HBAs to provide multiple redundant active paths to the storage for path failover and load balancing. For example, the Dell PowerEdge Cluster FE400 described in this article is a high-availability Fibre Channel cluster solution that encompasses a particular set of server, storage, switch, and HBA cluster components (based on their generation and/or component revisions).

Scaling out is achieved by adding more server nodes seamlessly into the cluster's existing storage subsystem (the SAN). In addition, hardware components in the cluster can be scaled up to meet increasing demands. For example, the server can be scaled up to add more processing power (CPUs) or memory, and the storage subsystem can be scaled up to add more storage capacity.

Implementing Exchange Virtual Servers: Scalability and performance tests

One benefit of Exchange Server 2003 clusters is the ability to easily add compute nodes to meet business needs. However, incrementally increasing the workload while simultaneously maintaining equivalent system latencies, or *linear scaling,* is difficult to achieve. Linear scaling cannot be readily achieved in a clustered environment for two key reasons: a shared storage system and intra-site messaging latencies.

In clustered messaging environments typically dominated by highly randomized I/O patterns, the shared storage system using SAN technologies and building blocks—including storage processors, storage processor cache memory, and disk resources—is subject to some latency inherent to random read and write I/O activity.

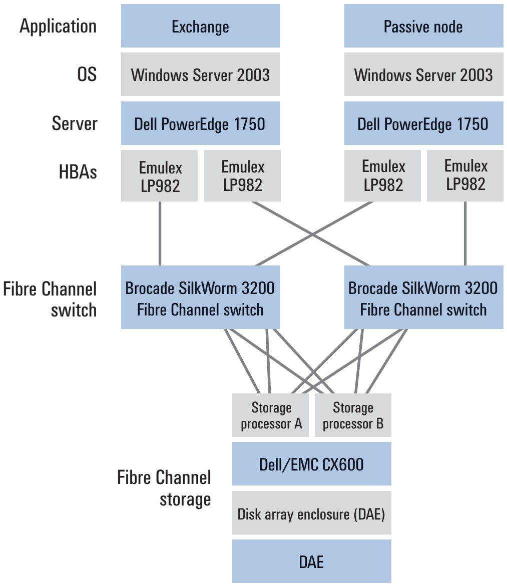


Figure 2. Dell PowerEdge Fibre Channel cluster

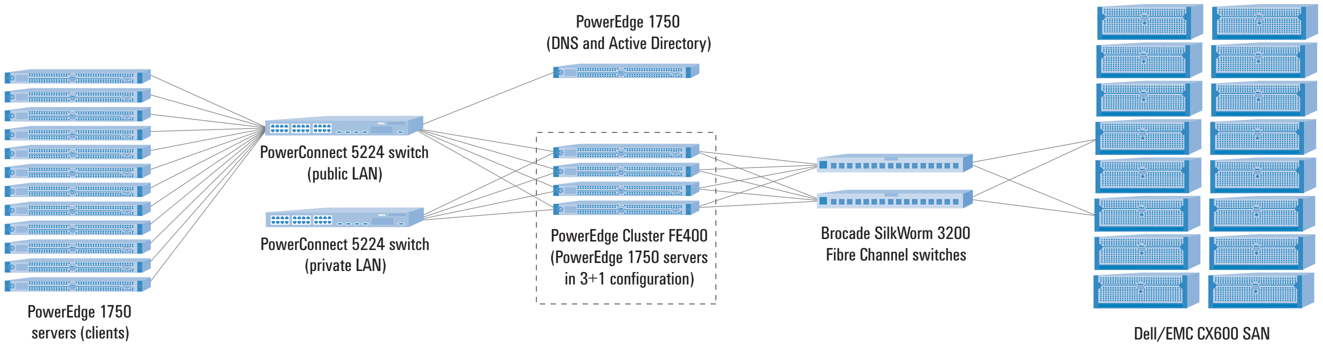


Figure 3. Microsoft Exchange Server 2003 test configuration

As additional cluster server nodes connect to the shared storage system and as those server nodes in turn introduce more work, the time required to seek out a specific data block increases.

Microsoft Exchange administrators familiar with the concept of single instance ratio recognize the latency implications that intra-site messaging has on message delivery times and message queues. Single instance ratio, often touted as one way to reduce the storage requirements for a message destined for more than one user on the same server, also reduces the amount of network traffic beyond a single Exchange Server 2003 Virtual Server when large distribution lists are present. However, intra-site messaging latencies arise when messages travel between users whose mailboxes reside on different servers in the Exchange organization. For example, target user names and properties often require resolution from a Microsoft Active Directory® domain controller before messages can be delivered. By design, the resolution mechanism introduces delays, so some latency increases can be expected in an Exchange organization that hosts thousands of mailboxes spread across numerous physical servers that may not be in the same Active Directory site.

Establishing a baseline for Exchange Server 2003 performance

Before examining overall Exchange Server 2003 cluster scaling and performance data, administrators should establish and explore a baseline system and understand how it is affected by a random messaging workload. The Microsoft LoadSim 2003 test configuration can be used for conducting MAPI (Messaging Application Programming Interface) Messaging Benchmark 3 (MMB3) tests. In June and July of 2004, the Dell Scalable Enterprise Computing team tested the scalability of Microsoft Exchange Server 2003 using LoadSim 2003 and MMB3 on a PowerEdge Cluster FE400 consisting of four PowerEdge 1750 servers in a 3+1 configuration and a Dell/EMC CX600 SAN. As shown in Figure 3, this test configuration also consisted of an additional 12 PowerEdge 1750 servers, which acted as clients; two Dell PowerConnect™ 5224 switches; another PowerEdge 1750 that served as the Domain Name Server (DNS) and Active Directory server; and two Brocade SilkWorm 3200 switches.

Although RAID-0 is not recommended for mission-critical systems in a production environment, its use in a lab testing environment can help minimize I/O bottlenecks and allow other subsystems such as memory, the network, and processors to achieve higher utilization levels. The highest Exchange Server 2003 MMB3 results are typically achieved on systems running RAID-0.

Pushing server and storage building blocks to high utilization levels with tools like LoadSim 2003 can help performance engineers evaluate how individual building blocks affect and contribute to overall system and application performance and scaling. Engineers and system administrators can then leverage this insight and apply those principles to the design of scalable, reliable, and high-performance production-level messaging systems.

Scaling out within the cluster

To measure scale-out performance, three active Exchange Virtual Servers were activated. Figures 4, 5, 6, and 7 show performance data for the three test configurations: a single active Exchange Server 2003 cluster node (5,000 MMB3 users); two active cluster nodes (10,000 MMB3 users); and three active cluster nodes (15,000 MMB3 users).

In Figure 4, average processor utilization is compared across the three configurations. As the results show, the single active node ran at just over 50 percent processor utilization, but with three active

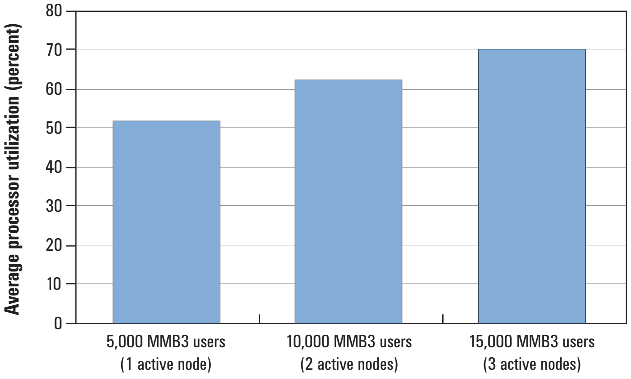


Figure 4. Average processor utilization as workload increases (cluster aggregate)

nodes, the cluster reached 70 percent processor utilization. Thus, the average cluster host CPU utilization increased by less than 20 percent as additional Exchange Virtual Servers were brought online and the workload tripled.

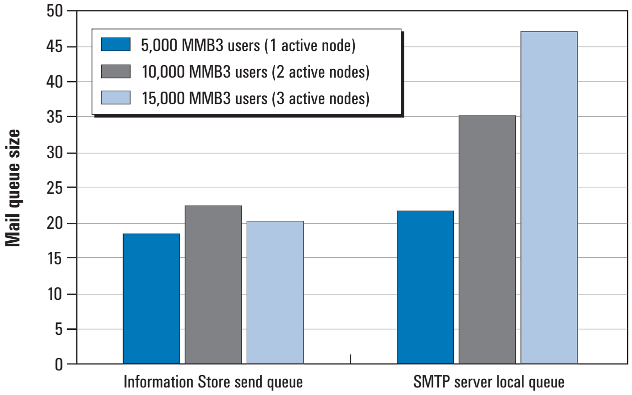


Figure 5. Mail queue sizes for Information Store send queue and SMTP server local queue

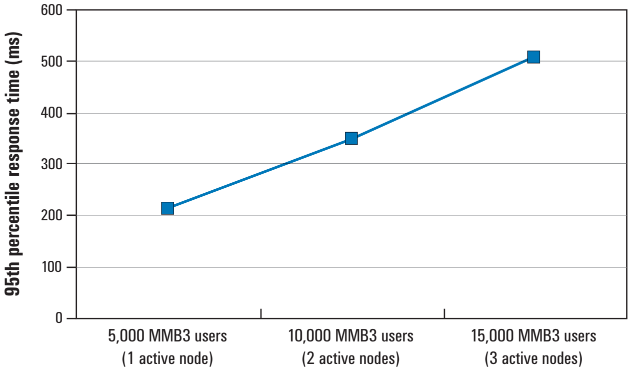


Figure 6. 95th percentile overall response times

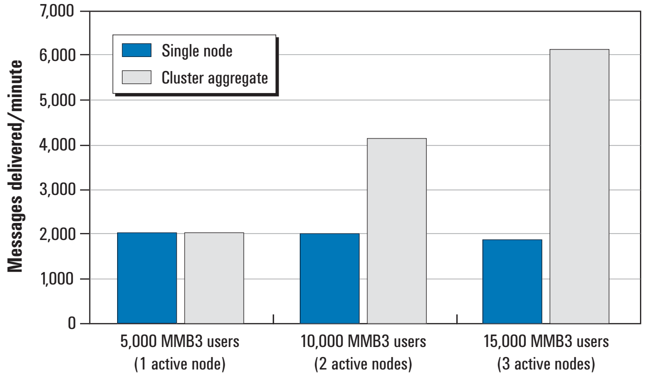


Figure 7. Messaging throughput: single node versus cluster aggregate

Figure 5 compares mail queue sizes for the Information Store send queue and the Simple Mail Transport Protocol (SMTP) local queue. While queue sizes remained relatively stable across the configurations for the Information Store send queue, the cluster was able to handle larger SMTP local queue sizes as the workload increased.

Figure 6 shows the aggregate response time for the cluster. Response time increased as more MMB3 users were added with their respective nodes. This is expected as a function of shared storage.

Figure 7 shows the messaging throughput—the rate at which messages were delivered to all recipients—for all three workloads, both as single nodes and as a cluster. The "Messages delivered/minute" results indicate that, while throughput for the single-node workloads remained relatively static, the cluster sustained a high level of aggregate messaging throughput even as the messaging workload and overall server response time (shown in Figure 6) increased.

Several factors such as I/O subsystem latency and intra-server communications affected linear scaling in the clustered Exchange environment. As shown by Figures 4, 5, 6, and 7, the PowerEdge Cluster FE400 was able to sustain consistent levels of messaging throughput, even as overall server and storage subsystem utilization levels increased.

Scaling out with high-availability Dell clusters

Dell PowerEdge clusters are designed to offer organizations several options for high availability and high performance. The test results discussed in this article show how the Dell PowerEdge Cluster FE400 can scale well and provide a solid, high-availability platform for active/passive Exchange Server 2003 configurations.

Arrian Mehis is a systems engineer on the Server and Storage Performance Analysis team in the Dell Enterprise Product Group. His current focus includes Microsoft Exchange Server single-node and high-availability cluster performance analysis on Dell server and SAN solutions. Arrian has a B.S. in Computer Engineering with a minor in Information Systems from the Georgia Institute of Technology.

Ananda Sankaran is a systems engineer in the High-Availability Cluster Development Group at Dell. His current interests related to high-availability clustering include storage systems, application performance, business continuity, and cluster management. Ananda has a master's degree in Computer Science from Texas A&M University.

Scott Stanford is a systems engineer on the Scalable Enterprise Computing team in the Dell Enterprise Product Group. His current focus is on performance analysis and characterization for Dell/VMware solutions. He has an M.S. in Community and Regional Planning from The University of Texas at Austin and a B.S. from Texas A&M University.

FOR MORE INFORMATION

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

Streamlining Beowulf Cluster Deployment with NPACI Rocks

For high-performance parallel-processing applications, Beowulf clusters that comprise industry-standard, two-processor and four-processor servers can be a cost-effective alternative to symmetric multiprocessing computer systems and supercomputers. This article discusses NPACI Rocks, an open source cluster computing software stack that can be used to improve the deployment, management, and maintenance of Beowulf clusters.

BY RINKU GUPTA, YUNG-CHIN FANG, AND MUNIRA HUSSAIN

Because they are built from cost-effective, industry-standard components such as Dell™ PowerEdge™ servers, Beowulf clusters can provide a price/performance ratio that makes them a compelling alternative to more expensive symmetric multiprocessing (SMP) systems for many high-performance parallel computing systems. However, administration of clusters can become problematic for IT organizations as clusters grow.

Along with the popularity and customer acceptance of Beowulf clusters has come the need for a robust and comprehensive cluster computing software stack to simplify cluster deployment, maintenance, and management. This article introduces NPACI Rocks, an open source, Linux®-based software stack for building and maintaining Beowulf high-performance computing (HPC) clusters.

The NPACI Rocks toolkit was designed in November 2000 by the National Partnership for Advanced Computational Infrastructure (NPACI). The NPACI facilitates collaboration between universities and research institutions to build cutting-edge computational environments for future scientific research. The organization is led by the University of California, San Diego, and the San Diego Supercomputer Center.

NPACI Rocks is designed to make clusters easy to deploy, manage, maintain, and scale. The Rocks package is built on standard and mostly open source components and is available as a free download on the NPACI Rocks Web site.[1]

Rocks software components and capabilities

NPACI Rocks provides a collection of integrated software components that can be used to build, maintain, and operate a cluster. Its core functions include the following:

- Installing the Linux operating system (OS)
- Configuring the compute nodes for seamless integration with the cluster
- Constructing a database of cluster-wide information
- Delivering middleware libraries and tools that build programs to run on the cluster
- Monitoring the cluster
- Managing the cluster
- Providing for system integration, packaging, and documentation

Installing the Linux OS. Rocks is based on the Red Hat® Linux OS. The availability of kickstart tools and the RPM™

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

(Red Hat Package Manager) under Red Hat Linux served as a great impetus for the selection of Red Hat Linux as the base OS. The OS is installed on the compute nodes using the Preboot Execution Environment (PXE), in which the client nodes perform a network boot to obtain the OS. Rocks recompiles the available Red Hat RPMs and uses them in the Rocks package.

Configuring the compute nodes for seamless integration with the cluster. Like other open source and commercial cluster packages, Rocks uses a master node—called a front-end node in the Rocks terminology—for centralized deployment and management of a cluster. The Rocks front-end node helps administrators to specify the cluster-wide private cluster network configuration and networking parameters, including IP addresses assigned to the compute nodes. These parameters are specified during the front-end node installation process. The front-end node then provides the IP addresses to the compute nodes when the compute nodes contact the front-end nodes via network PXE boot.

Constructing a database of cluster-wide information. Many services at and above the OS level—for example, job schedulers and the Dynamic Host Configuration Protocol (DHCP) server—require a global knowledge of the cluster to generate service-specific configuration files. The front-end node of a Rocks cluster maintains a dynamic MySQL database back end, which stores the definitions of all the global cluster configurations. This database forms the core of the NPACI Rocks package, and can be used to generate database reports to create service-specific configuration files such as /etc/hosts and /etc/dhcpd.conf.

Delivering middleware libraries and tools that build programs to run on the cluster. Rocks delivers code using two types of CD. The Rocks Base CD contains the core Rocks package. Rocks Roll CDs are add-on software packages that augment the cluster with specific capabilities. The Rocks Roll software packages, designed to seamlessly integrate with the base Rocks installation, provide a mechanism to enable administrators and vendors to add software that provides extra functionality to Rocks clusters. Administrators can create their own Rolls, independent from the Rocks Base CD, to contain their domain-specific software and applications.

The following components come packaged in a separate HPC Roll CD: middleware (for example, MPICH,[2] a portable implementation of Message Passing Interface [MPI], and Parallel Virtual Machine [PVM][3]); cluster performance monitoring software such as Ganglia;[4] and common performance benchmark applications such as Linpack. Other Rolls provided by NPACI include Intel® compilers,[5] the Maui[6] scheduler, and Portable Batch System (PBS)[7] resource manager functionality.

Monitoring the cluster. Cluster monitoring is an important task for helping system administrators to proactively understand and troubleshoot issues occurring in a cluster. Rocks is packaged with open source software tools such as Ganglia, which helps provide in-band global monitoring of the entire cluster. Ganglia is an open source tool developed by the University of California at Berkeley and the San Diego Supercomputer Center under National Science Foundation (NSF) NPACI funding.

Managing the cluster. Efficient tools for cluster management are an important part of any cluster computing package. NPACI Rocks provides easy-to-use command-line interface (CLI) commands for adding, replacing, and deleting nodes from the cluster. Rocks treats compute nodes as soft-state machines—that is, machines that have no application information or data stored on their hard drives—and uses fresh OS reinstallation as a way to ensure uniformity across all the nodes in the cluster. This OS-reinstallation approach works well for large clusters because the OS is installed rapidly and automatically. As a result, no time is wasted in performing exhaustive checks to debug problems that might exist in a system on which the OS has already been installed. For cluster-wide operations, NPACI Rocks comes with a parallel command (`cluster-fork`), which is capable of executing query-based operations.

Rocks cluster installation

NPACI Rocks cluster installation consists of two parts. Figure 1 shows a typical Rocks cluster layout.

Front-end node installation

Front-end node installation requires the administrator to install the Rocks Base CD along with any of the optional Rocks Roll CDs. As of Rocks 3.3.0, the Rocks Base CD along with the HPC Roll CD and the Kernel Roll CD are required to build a functional Rocks-based HPC cluster.

A front-end node typically has two network interface cards (NICs). One of the NICs connects to the external, public network. The other NIC is used to connect to the private network of the cluster. Rocks interactive front-end node installation allows the administrator to configure the external-network NIC. It also allows the administrator to specify the network configuration details (such

[2] For more information about MPICH, visit www-unix.mcs.anl.gov/mpi/mpich.

[3] For more information about PVM, visit www.csm.ornl.gov/pvm.

[4] For more information about Ganglia, visit ganglia.sourceforge.net.

[5] For more information about Intel compilers, visit www.intel.com.

[6] For more information about Maui, visit www.clusterresources.com/products/maui.

[7] For more information about PBS, visit www.openpbs.org.

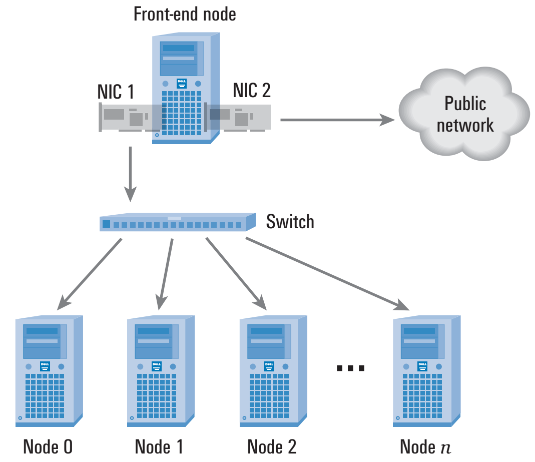


Figure 1. A typical Rocks cluster layout

as IP address and subnet mask) of the private network NIC. These details are used to assign IP addresses to the client nodes of the cluster in the latter part of the cluster installation process.

Rocks is self-contained, including the necessary tools, scripts, and services for subsequent installation phases. Rocks front-end node installation takes place on bare-metal hardware. The Rocks CDs install the OS on the front-end node during the installation procedure.

Compute node installation

The rest of the Rocks cluster is installed from the front-end node, using a test-based utility called insert-ethers (see Figure 2). To build the compute nodes, administrators can launch the insert-ethers utility from the front-end node. The insert-ethers utility is completely database driven. It allows the administrator to configure each node as a compute node, to be used for computational purposes; a Parallel Virtual File System (PVFS)[8] node, to be used only for I/O purposes; or a combination compute and PVFS node, to be used for both computational and I/O purposes.

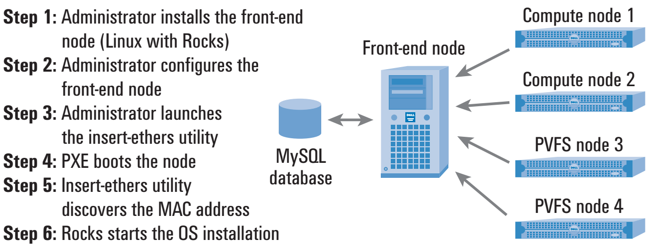
The insert-ethers utility is also used to populate the nodes table in the MySQL database. To this end, the utility continuously parses the Linux OS log files (such as the /var/log/messages file) to extract DHCPDISCOVER messages and Media Access Control (MAC) addresses. When the compute node is PXE-booted, the DHCP request is sent to the front-end node. When the front-end node finds the MAC address in the parsed files, and if the MAC address is not present in the database, then the front-end node accepts the new node and installs it using the Rocks kickstart mechanism. However if the MAC address is already present in the database, the front-end node reinstalls the compute node by assigning the same IP address that it previously had. Hence, the insert-ethers utility merely reimages the compute node by matching previous information present in the database. Front-end node scripts in Rocks generate the kickstart file on the fly depending on the information—such as architecture and CPU count—provided by the compute node through the modified Rocks-supplied Red Hat Anaconda installer.

Rocks is designed to make compute node installation easy. Rocks can also detect NICs (such as Myricom Myrinet[9] cards) on the compute nodes and compile the correct drivers for these NICs during installation. Rocks is based on XML and the Python[10] programming language. This design, combined with the MySQL database back end, helps provide great flexibility for administrators to modify compute node settings such as disk partitioning; enable selected services such as Remote Shell (RSH); and install third-party packages. An administrator can customize the target node environment by modifying select XML files in Rocks and simply re-creating a new Rocks distribution, which can then be used for installation.

Rocks cluster management and monitoring

Rocks uses the basic mechanism of OS reinstallation for updating the nodes in the cluster. If a hardware failure occurs, or if a hard drive or NIC must be replaced, the compute node should be reinstalled from the front-end node; the front-end node's MySQL database is automatically updated if required. If additional software or update files need to be installed on the compute nodes, then the administrator should add these to the front-end node and create an updated Rocks distribution. All nodes should then be reinstalled using the updated distribution to help maintain consistency across the cluster. Reinstallation is fast and helps reduce or eliminate administrator time spent troubleshooting and debugging failed nodes.

The Rocks CLI is also used to replace and remove nodes from a cluster. An administrator can create a naming scheme that helps identify nodes by their cabinet and rack locations.



Step 1: Administrator installs the front-end node (Linux with Rocks)
Step 2: Administrator configures the front-end node
Step 3: Administrator launches the insert-ethers utility
Step 4: PXE boots the node
Step 5: Insert-ethers utility discovers the MAC address
Step 6: Rocks starts the OS installation

Figure 2. Steps to install a Rocks cluster

[8] For more information about PVFS, visit www.parl.clemson.edu/pvfs.

[9] For more information about Myrinet, visit www.myri.com.

[10] For more information about Python, visit www.python.org.

For monitoring, Rocks provides a set of Web pages, which are accessible from the front-end node of the cluster. The administrator can view the MySQL database (using a MySQL Web-based administration tool called phpMyAdmin) that the front-end node of the cluster creates and uses. Ganglia is the default software package bundled with Rocks for monitoring purposes. The Web pages provide a graphical user interface (GUI) for use on the Ganglia monitors running on each cluster node. These Ganglia monitors provide a host of information including CPU usage and load, memory usage, and disk usage. Rocks also provides a utility called Cluster Top, which can be viewed from a browser on the front-end node. The Cluster Top utility is a version of the standard `top` command for the cluster. The Cluster Top Web page presents process information from each node in the cluster.

In addition, administrators can use the `cluster-fork` parallel command packaged in Rocks. This CLI-based parallel command, when used in conjunction with queries against the MySQL database, is designed to provide comprehensive management functionality.

HPC packages and middleware in Rocks

Rocks, when installed with the HPC Roll and the Kernel Roll, installs many HPC cluster packages and automatically compiles typical MPI libraries such as MPICH and MPICH-GM (MPICH for GM environments). Benchmark software such as Linpack and IOzone are also compiled and installed for ready use.

Rocks 3.3.0 provides several Roll CDs that can be used to add functionality to the cluster. The Intel Compiler Roll installs all relevant Intel compilers. The Condor Roll installs the Condor job scheduler and the relevant packages,[11] and the Sun Grid Engine[12] Roll installs the Sun-sponsored Sun Grid Engine scheduler packages. A list of current Rolls and their intended functionality can be found at the Rocks Web site. In addition, Rocks allows administrators to create custom Rolls that include their own specific packages, thereby enabling organizations to flexibly adapt Rocks for their individual IT environments.

Platform Computing Rocks

NPACI Rocks, an open source Linux-based alternative to proprietary cluster solution packages, can help significantly reduce the complexity of building Beowulf HPC clusters. Moreover, its integrated bundle of software helps ease cluster use and maintenance, and the NPACI Rocks Group is working to enhance and extend this software to address diverse applications in the HPC field.

With the acceptance of NPACI Rocks in the cluster community, vendors are in the process of providing commercial, supported versions of NPACI Rocks. Platform Rocks, which is being developed by Platform Computing, is based on NPACI Rocks. Platform Rocks is a hybrid software stack featuring a blend of market-leading open source software technologies and proprietary products.[13]

Platform Rocks can be obtained from Platform Computing with an annual support subscription, which provides support, maintenance, fixes, and other value-added services. Platform Rocks includes the supported Red Hat OS version, as compared to the recompiled Red Hat OS used by NPACI Rocks.

Platform Rocks has been verified and validated on 8- to 256-node clusters of the latest-generation Dell HPC servers[14]—including the Dell PowerEdge 1850 server, PowerEdge SC1425 server, and PowerEdge 1855 blade server—using both x86 technology and Intel Extended Memory 64 Technology (EM64T), Red Hat operating systems, and Myrinet and InfiniBand[15] interconnects.

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.

Yung-Chin Fang is a senior consultant in the Scalable Systems Group at Dell. He specializes in cyberinfrastructure resource management and high-performance computing. He also participates in open source groups and standards organizations as a Dell representative. Yung-Chin has a B.S. in Computer Science from Tamkang University in Taiwan and an M.S. in Computer Science from Utah State University.

Munira Hussain is a systems engineer in the Scalable Systems Group at Dell. Her current research interests are in the areas of interconnects, IA-64 and EM64T, life sciences applications, and Linux and Microsoft® Windows® software stacks for high-performance computing. She has a B.S in Electrical Engineering and a minor in Computer Science from the University of Illinois at Urbana-Champaign.

FOR MORE INFORMATION

NPACI Rocks:
www.rocksclusters.org/Rocks

Platform Rocks:
www.platform.com

[11] For more information about Condor, visit www.cs.wisc.edu/condor.

[12] For more information about the Sun Grid Engine, visit gridengine.sunsource.net.

[13] For more information about Platform Rocks, visit www.platform.com.

[14] For test results, visit www.rocksclusters.org/rocks-register.

[15] For more information about InfiniBand, visit www.topspin.com.

Designing High-Performance Clusters

with the Dell PowerEdge SC1425 Server

Hardware selection for high-performance computing clusters is often driven by the characteristics of the parallel applications that will be deployed. This article discusses different classes of parallel applications and presents the Dell™ PowerEdge™ SC1425 server as a viable, low-cost platform on which to build clusters for different classes of applications.

BY RON PEPPER AND RINKU GUPTA

An important consideration when designing a high-performance computing (HPC) cluster is the characteristics of the parallel application that will run on the cluster. Application characteristics go a long way in determining the components needed for the cluster. For example, a long-running parallel application that exchanges many small messages between nodes in the cluster may require a special network communications infrastructure. In this case, the cluster design should specify that the compute servers be connected to each other by a fast network interconnect that can send and receive many small messages very quickly.

Defining the types of HPC applications

A parallel application runs on a distributed collection of nodes. Hence, all applications consist of communication steps (to communicate between themselves) and computation steps (to perform independent computation). Based on the degree of communication and computation, parallel applications in the HPC field fall into three broad categories.

Coarse-grained parallel applications. Beowulf clusters—that is, parallel-processing HPC clusters comprising industry-standard components—have traditionally been built to exploit coarse-grained parallel applications. Parallel applications for which the overall time spent in computation is much higher than the time spent in communication fall into this category. A coarse-grained application is an ideal candidate for running Beowulf clusters, because there is a great probability of obtaining higher performance at a lower price when additional servers or CPUs are added to the cluster. Because of their highly parallel nature, these applications are also called embarrassingly parallel applications. A classic example is the Monte Carlo simulation problem.

Medium-grained parallel applications. Applications for which the computation time is greater (but not much greater) than the communication time are considered medium-grained parallel applications. As compared to coarse-grained parallel applications, medium-grained applications have a much lower computation-to-communication ratio. The communication overhead is apparent for medium-grained parallel applications.

Fine-grained parallel applications. In a fine-grained application, the time intervals spent on communication tasks are greater than or equal to the time spent on computation tasks. Such an application is not considered the best candidate for Beowulf clusters. Even if the application is parallelized and run on a cluster of nodes, the scalability of the application may be limited.

Choosing the right cluster components

Components in a cluster are generally selected based on the generic applications that will be run on them. Every part of the cluster plays an important role in the cluster environment, depending upon the specific application's ability to exploit it. Two of the most important components in a cluster are server type and network interconnect. The servers in a cluster provide the computational power and are connected through the interconnect fabric.

For coarse-grained parallel applications, communication is generally less of a concern. Hence, organizations running a high-performing, coarse-grained parallel application can select high-performance servers, which offer good computation power. Gigabit networking can meet the interconnect requirements of coarse-grained parallel applications while maintaining a low price point. Using a traditional low-cost interconnect with a high-performance server can provide a low-cost HPC solution for these applications.

For medium-grained parallel applications, the communication overhead can be high. For such applications, organizations should consider a high-performance interconnect like Myricom Myrinet[1] or InfiniBand. Both interconnects offer low latency and high bandwidth. *Latency* is the amount of time taken to transmit a message from the source server to the destination server. *Bandwidth* is the rate at which information can be sent over the interconnect. A slow interconnect with a fast processing system will cause the interconnect to become a bottleneck for the application.

For fine-grained parallel applications running on Beowulf clusters, a fast interconnect like Myrinet or InfiniBand is recommended. Using a slower interconnect could cause the communication time to overshadow the computation time, rendering the parallelization of the application unsuccessful.

Server types and interconnects are two high-level components of a cluster. Choosing the most appropriate server is in itself a broad topic with many components such as memory subsystem, processor speed, and cache sizes to be considered. Dell supports a wide range of rack-mount servers—including the PowerEdge 1850, PowerEdge 1855, PowerEdge 2850, and PowerEdge 3250 servers—that are suitable for HPC clusters and offer varied architectures to satisfy computational needs.

For coarse-grained applications, choosing the appropriate industry-standard components can enable organizations to create a low-cost cluster that will still meet their application needs. A recently released server from Dell, the PowerEdge SC1425, is one such component that can help provide a viable low-cost alternative for HPC needs.

Testing the Dell PowerEdge SC1425 server as a cluster node

In October 2004, a team of Dell engineers tested cluster performance using a PowerEdge SC1425 server as the building block for a low-cost cluster. For this study, each PowerEdge SC1425 server was configured with two Intel® Xeon™ processors. However, the PowerEdge SC1425 can be configured with either single or dual CPUs and can run in either 32-bit mode or Intel Extended Memory 64 Technology (EM64T) mode for 64-bit applications. For memory subsystem needs, the PowerEdge SC1425 supports dual-banked double data rate 2 (DDR2) memory on an 800 MHz frontside bus (FSB). It can be equipped with either serial ATA (SATA) or SCSI hard drives. Because most of the data in this cluster performance test was shared from the master node, a local SCSI hard drive on each compute node was not necessary. Two embedded Gigabit[2] Ethernet controllers were included in the base system, which eliminated the need for any additional network hardware.

The following sections demonstrate the performance of the PowerEdge SC1425 cluster for different communication patterns. Using various well-known benchmarks—Pallas PingPong, NAS Parallel Benchmarks, IOzone, and STREAM—the team conducted tests on a single node, two nodes, and the entire cluster to profile the PowerEdge SC1425 for cluster performance. The lab setup consisted of four Dell PowerEdge SC1425 compute server nodes, each with 2 GB of main memory and dual symmetric multiprocessing (SMP) processors at 3.6 GHz. Two types of interconnect were examined: Gigabit Ethernet and Myrinet. The Dell PowerConnect™ 5324 switch was used to form the Gigabit Ethernet fabric, while Myricom Myrinet switches formed the Myrinet fabric.

The impact of interconnects on cluster performance

Using the Pallas[3] PingPong benchmark, the Dell test team examined latency at various message sizes for two PowerEdge SC1425 cluster nodes. The two nodes—one using Gigabit Ethernet and the other using Myrinet—sent data to each other via the Message Passing Interface (MPI). As shown in Figure 1, the latency for Myrinet with small messages of 4 bytes was about 6.5 microseconds (µs) compared to the latency of 35 µs for 4-byte messages using Gigabit Ethernet.

Figure 2 shows the MPI message throughput between these two cluster nodes at various message sizes. In this study, Gigabit

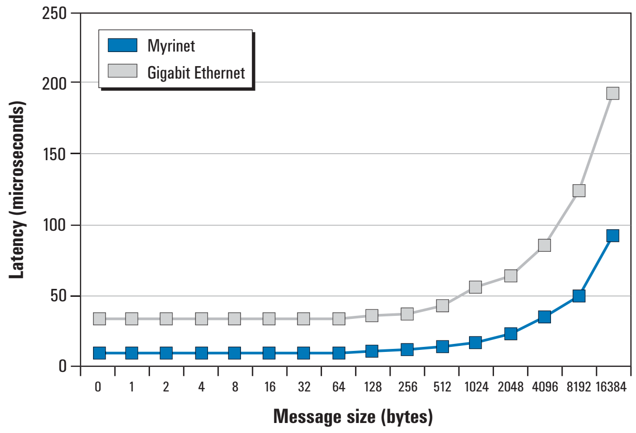


Figure 1. Latency between two PowerEdge SC1425 cluster nodes using different interconnects

[1] For more information about Myrinet, see www.myri.com.

[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.

[3] For more information about Pallas MPI benchmarks, see www.pallas.com/e/products/pmb/index.htm.

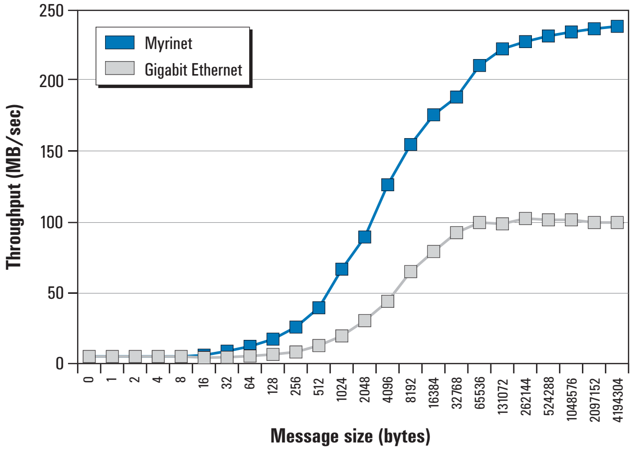


Figure 2. Throughput between two PowerEdge SC1425 cluster nodes using different interconnects

Ethernet scaled up to 95 percent of the theoretical bandwidth, offering close to 100 MB/sec. Myrinet, in contrast, demonstrated bandwidth of nearly 235 MB/sec (close to its theoretical peak of 250 MB/sec) for message sizes of up to 4 GB.

For coarse-grained applications, it is usually sufficient to have a Gigabit Ethernet communication fabric. The on-board Gigabit Ethernet network interface cards (NICs) built into each PowerEdge SC1425 server can be used for connecting the systems together. Gigabit Ethernet provides a good price/performance ratio for coarse-grained applications.

For medium- and fine-grained applications, high-speed, low-latency interconnects like Myrinet can be instrumental in improving the overall performance of an application. An example of this can be seen in Figure 3, which shows relative performance results from the NAS Parallel Benchmarks (NPB) suite.[4] The NPB suite is a set of eight programs derived from computational fluid dynamics code, and results are measured in millions of operations per second (MOPS). Each of the eight programs represents a particular function of highly parallel computation for aerophysics applications. NPB measures overall cluster performance, so the Dell team conducted the NPB tests on the entire four-node PowerEdge SC1425 cluster. Figure 3 shows the results of three NPB benchmarks on the four-node cluster with one instance on each node. The Embarrassingly Parallel (EP) benchmark from the NPB suite falls into the category of extreme coarse-grained application. The EP test generates pairs of Gaussian random deviates according to a specific scheme. Because EP does not perform any interprocessor communication, the results obtained in this study using different interconnects show the same performance characteristics—thereby supporting the assertion that clusters running applications similar to the EP benchmark suffice with a Gigabit Ethernet interconnect.

The Integer Sort (IS) benchmark from the NPB suite tests both integer computation speed and communication performance. It is a parallel integer sort program that is used in particle method codes. The IS benchmark involves no floating-point arithmetic but does have intense data communication. As shown in Figure 3, the type of interconnect used can significantly affect the performance of Integer Sort. In this study, the IS benchmark performed 1.75 times better on the Myrinet interconnect when compared to Gigabit Ethernet for a four-node cluster (with one instance). Hence, for such medium- and coarse-grained applications, low-cost clusters can benefit from Myrinet or InfiniBand interconnects. The same trend can be seen in the results of other NPB benchmarks like the Conjugate Gradient (CG) benchmark used in this study, whose performance increased by almost 50 percent (on a four-node cluster) when a Myrinet interconnect was used.

Disk I/O performance

Figures 4 and 5 show the relative performance of SCSI and SATA drives supported by the PowerEdge SC1425 server. The Dell test team generated this data using the IOzone benchmark[5] with a file size of 6 GB. IOzone measures the I/O performance of a single server, so the test was conducted on only one PowerEdge SC1425 cluster node.

Figure 4 shows the read performance with varying record sizes for a 6 GB file. In this study, the SCSI reads showed about 12 to 15 percent better performance as compared to the SATA reads. Figure 5 shows the write performance for varying record sizes. The SCSI writes showed more than 40 percent improvement as compared to the SATA writes.

The compute nodes in a cluster typically mount and use shared storage or storage located on the master node. When compute nodes do not use the local hard drives, it may be sufficient to use SATA drives, depending on specific application needs, and thus achieve a good price/performance ratio. However, if the PowerEdge SC1425

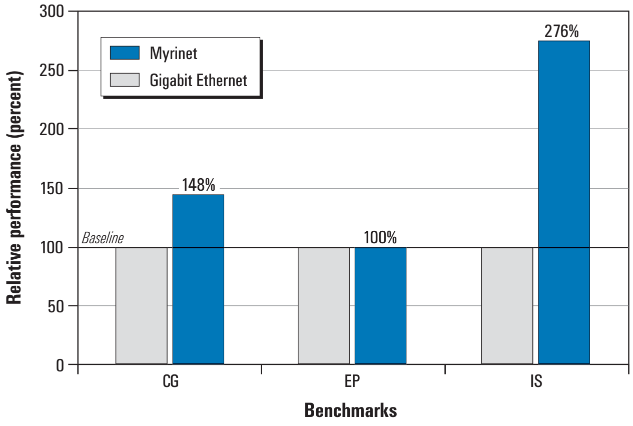


Figure 3. Results of the CG, EP, and IS benchmarks from the NPB suite for four nodes with one instance each

[4] For more information about the NAS Parallel Benchmark suite, see www.nas.nasa.gov/Software/NPB.

[5] For more information about IOzone, see www.iozone.org.

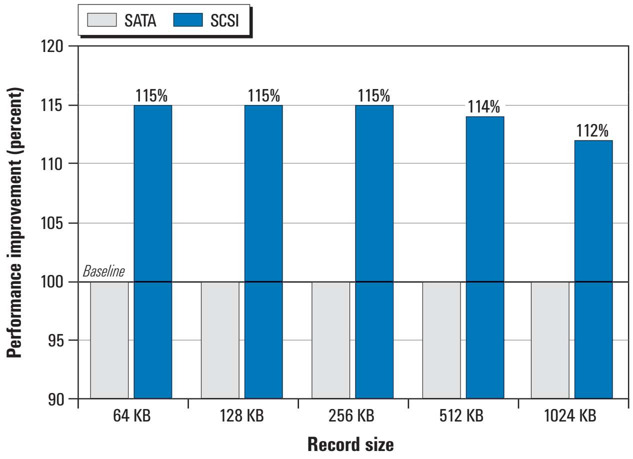


Figure 4. Comparison between SCSI and SATA drives: reads with IOzone benchmark

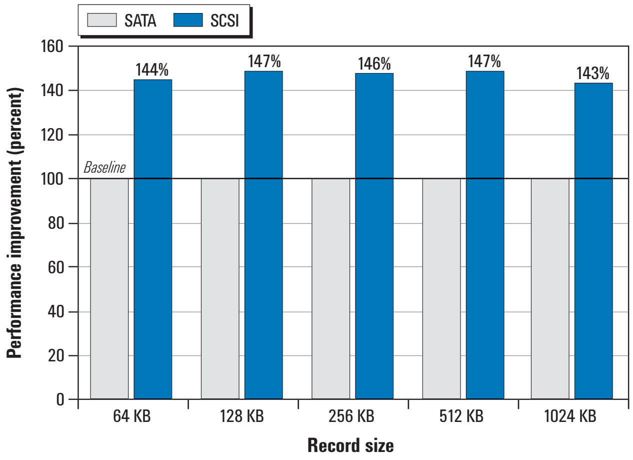


Figure 5. Comparison between SCSI and SATA drives: writes with IOzone benchmark

server is part of an I/O cluster in which each node uses its hard drives for local storage, then using SCSI drives can provide better performance than SATA drives.

Memory subsystem performance

Using the STREAM benchmark,[6] the Dell test team compared the memory subsystem performance of the PowerEdge SC1425 server with that of a PowerEdge 1750 server. STREAM is a synthetic benchmark that measures memory bandwidth in megabytes per second (MB/sec) and can be useful in understanding the speed at which the compute nodes can process large data sets that are too big to remain in the CPU cache. Memory bandwidth is defined as the amount of memory traffic that a computer system can move from memory to CPU.

For this memory subsystem performance study, the PowerEdge SC1425 was configured with an 800 MHz FSB and dual Intel Xeon processors at 3.6 GHz with 1 MB of level 2 (L2) cache; the PowerEdge 1750 was configured with a 533 MHz FSB and dual Intel Xeon processors at 3.2 GHz with 512 KB of L2 cache and 2 MB of level 3 (L3) cache. The PowerEdge SC1425 used 2 GB of 400 MHz error-correcting code (ECC) DDR2 memory; the PowerEdge 1750 used 4 GB of 266 MHz ECC DDR memory.

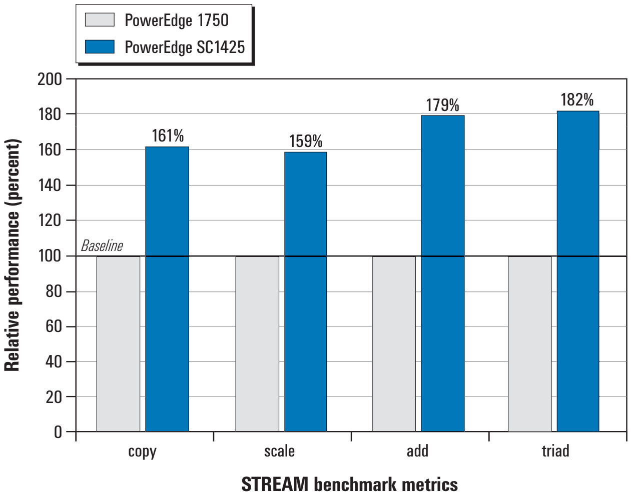


Figure 6. Comparison of memory subsystem performance in a PowerEdge 1750 server and a PowerEdge SC1425 server

Figure 6 shows that the PowerEdge SC1425 server demonstrated high memory bandwidth—achieving up to 82 percent more bandwidth than the PowerEdge 1750 server in this study. The 82 percent increase can be attributed mainly to the high-speed 800 MHz FSB and the 400 MHz DDR2 memory of the PowerEdge SC1425 as compared to the 533 MHz FSB and 266 MHz DDR memory of the PowerEdge 1750.

Deploying HPC clusters based on the PowerEdge SC1425 server

The PowerEdge SC1425 server can be an excellent choice for organizations that plan to deploy coarse-grained parallel applications in high-performance cluster environments. Dell supports standard configuration bundles with the PowerEdge SC1425 server, with node counts ranging from 8 to 256 nodes—and offers custom solution consulting services for larger node counts. The standard bundles support Gigabit Ethernet, Myrinet, and InfiniBand interconnects. PowerEdge SC1425 Gigabit Ethernet clusters can be optimal for coarse-grained applications that do not require a high-speed interconnect. In addition, the PowerEdge SC1425 server can be combined with high-speed interconnects to create a cluster suitable for a range of HPC application categories.

Ron Pepper is a systems engineer and advisor in the Scalable Systems Group at Dell. He works on the Dell HPC Cluster team developing grid environments. Ron attended the University of Madison at Wisconsin, where he worked on a degree in computer science; he is continuing his degree at Saint Edwards University.

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.

[6] For more information about STREAM, see www.cs.virginia.edu/stream.

Performance Characterization of BLAST on 32-bit and 64-bit Dell PowerEdge Servers

The Dell high-performance computing team recently explored the effect of Intel® Extended Memory 64 Technology (EM64T) on the performance of the basic local alignment search tool (BLAST), comparing BLAST performance on 32-bit and 64-bit architectures. This article discusses the performance results obtained when the team tested Dell™ PowerEdge™ servers using three Intel processor technologies: a PowerEdge 1750 server with 32-bit Intel Xeon™ processors; a PowerEdge 1850 server with Intel EM64T architecture; and a PowerEdge 3250 server with 64-bit Intel Itanium® 2 processors.

BY RAMESH RADHAKRISHNAN, PH.D.; RIZWAN ALI; GARIMA KOCHHAR; KALYANA CHADALAVADA; AND RAMESH RAJAGOPALAN

High-performance computing (HPC) applications have increasingly adopted clustered Intel architecture–based servers. This development has been largely due to several technological enhancements in Intel architecture–based servers, primarily substantial improvements in Intel processor and memory technology over the past few years.

Computing clusters built from industry-standard components such as Intel processors are becoming the fastest-growing choice for HPC systems. Twice yearly, the 500 most powerful computing systems in the world are ranked on the TOP500 Supercomputer Sites Web page at www.top500.org. In June 2002, 44 of the top 500 sites were using Intel processor–based systems; two years later, in June 2004, that number reached 287. And by November 2004, 318 of the top 500 sites were running on Intel processor–based systems.

Industry-standard Intel processor–based servers are changing the landscape of the enterprise server market and, in particular, the HPC market. Intel has continuously introduced processors at higher frequencies, with larger caches and faster frontside buses (FSBs). Technological innovations such as Intel Hyper-Threading Technology, Streaming SIMD (single instruction, multiple data) Extensions instructions, and Intel NetBurst® microarchitecture have significantly increased the performance of 32-bit Intel processors. The 64-bit Intel Itanium 2 processors are based on the Explicitly Parallel Instruction Computing (EPIC) architecture and have claimed the top spot in several industry-standard benchmarks.[1] Intel recently introduced a 90 nanometer (nm) version of the 32-bit Intel Xeon processor. The key architectural differences between the older 130 nm Intel Xeon processors and the 90 nm Intel Xeon processors include faster processor and FSB frequencies, larger caches, and support for Intel Extended Memory 64 Technology (EM64T). In addition, several bandwidth-enabling

[1] Intel Itanium 2 processors have achieved higher performance than RISC/x86 processors in the SPEC benchmarks; visit www.spec.org for archives of these benchmark results. For the latest Itanium benchmark performance information, visit www.intel.com/go/itanium2.

	PowerEdge 1750 (IA-32)	PowerEdge 1850 (IA-32)	PowerEdge 3250 (IA-64)
CPU	Dual Intel Xeon processors at 3.2 GHz (130 nm)	Dual Intel Xeon processors at 3.2 GHz and at 3.6 GHz (90 nm)	Dual Intel Itanium 2 processors at 1.5 GHz
FSB	64 bits wide; 533 MHz	64 bits wide; 800 MHz	128 bits wide; 400 MHz
Cache size	L2: 512 KB L3: 1 MB	L2: 1 MB	L2: 256 KB L3: 6 MB
Memory	Four 1 GB DDR at 266 MHz	Two 2 GB DDR2 at 400 MHz	Four 1GB DDR at 266 MHz (operating at 200 MHz)
Memory bandwidth	4.8 GB/sec	6.4 GB/sec	6.4 GB/sec
Operating system	32-bit Red Hat® Enterprise Linux® 3, Update 2	32-bit Red Hat Enterprise Linux 3, Update 2	Red Hat Enterprise Linux 3 for IA-64

Figure 1. Servers tested in BLAST performance study

technologies like double data rate 2 (DDR2) memory and Peripheral Component Interconnect (PCI) Express are available in Dell servers supporting the 90 nm Intel Xeon processors.

A continuation of the study reported in the October 2004 issue of *Dell Power Solutions,*[2] the study discussed in this article compared the performance characteristics of Dell PowerEdge server models equipped with 32-bit Intel Xeon processors, the new 90 nm Intel Xeon processors, and 64-bit Itanium 2 processors by using a scientific application—basic local alignment search tool (BLAST)—that is widely used in the field of bioinformatics. The aim of the study was to understand the impact of the different features and technologies that are available in Intel processors as well as the impact of memory technology on the performance of BLAST. To achieve this end, the Dell engineers used the STREAM benchmark to measure the memory system performance of the Dell servers. Furthermore, the test team studied the effect of the processor clock frequency on the performance of BLAST.

Figure 1 lists the servers tested in the Dell study and the processor technology used by each server. The Dell model names are used to avoid ambiguity between the current-generation 130 nm Intel Xeon and the more recent 90 nm Intel Xeon processor. Empirical studies have shown that small-scale symmetric multiprocessing (SMP) systems make excellent platforms for building HPC clusters. Thus, all the servers used in this study were two-processor systems.

Comparison of Intel processor–based systems

The processor and memory subsystems used in the test servers had the biggest impact on the performance of BLAST. The following sections discuss the architecture of these two subsystems.

Overview of processor architectures

The Intel NetBurst microarchitecture is the core of Intel's 130 nm technology, on which the 32-bit Intel Xeon processor—used in the Dell PowerEdge 1750 server—is based. The Intel NetBurst architecture uses a 20-stage pipeline that allows higher core frequencies than possible in previous-generation processors. The 130 nm Intel Xeon processors were introduced at speeds of 1.8 GHz and are currently available at speeds of up to 3.2 GHz. The FSB scales from 400 MHz in the initial 180 nm Intel Xeon processors to 533 MHz on the 3.2 GHz 130 nm processors.

The Intel Xeon processor is a superscalar processor that combines out-of-order speculative execution with register renaming and branch prediction to enhance performance. The processor uses an Execution Trace Cache that stores pre-decoded micro-operations. Streaming SIMD Extensions 2 (SSE2) instructions are used to speed up media types of workload.

The Dell PowerEdge 1850 server is the follow-on to the PowerEdge 1750 server, and uses the more recent Intel Xeon processor based on 90 nm technology. The 90 nm Intel Xeon processor is an extension of the 130 nm Intel Xeon processor. However, some architectural differences between these two Intel Xeon processors can have an impact on application performance. The 90 nm Intel Xeon processor is being introduced at a frequency of 2.8 GHz, coupled with a faster 800 MHz FSB. It uses a longer 31-stage processor pipeline that will facilitate higher frequencies in future versions.

The Dell PowerEdge 3250 system is based on the Itanium processor family, which uses a 64-bit architecture and implements the EPIC architecture. Instructions in groups, or *bundles,* are issued in parallel, depending on the available resources. The Itanium 2 processors differ from the Intel Xeon processors in the fact that they use software—a compiler—to exploit parallelism, as opposed to complex hardware to detect and exploit instruction parallelism. Software compilers provide the information needed to efficiently execute instructions in parallel.

Computing clusters built from industry-standard components such as Intel processors are becoming the fastest-growing choice for HPC systems.

Itanium 2 processors are available in frequencies ranging from 1.0 GHz to 1.9 GHz and use varying sizes of level 3 (L3) caches ranging from 1.5 MB to 9 MB. A 128-bit 400 MHz FSB is used to connect the processors. The Itanium 2 processor has a large number of registers compared to the Intel Xeon processors: 128 64-bit general-purpose registers (GPRs), which are used by the compiler to keep the six Integer Functional Units busy.

[2] See "Performance Characterization of BLAST on Intel EM64T Architecture" by Garima Kochhar, Kalyana Chadalavada, Amina Saify, and Rizwan Ali in *Dell Power Solutions,* October 2004.

Differences in memory subsystems

The Dell PowerEdge 1750 and Dell PowerEdge 3250 servers use DDR at 266 MHz (PC2100) memory. The PowerEdge 3250 server, however, operates at the speed of 200 MHz. The Dell PowerEdge 1850 server uses the new DDR2 memory running at 400 MHz (PC3200), which has a theoretical bandwidth of 3.2 GB/sec. DDR2 architecture is also based on the industry-standard dynamic RAM (DRAM) technology. The DDR2 standard contains several major internal changes that allow improvements in areas such as reliability and power consumption. One of the most important DDR2 features is the ability to prefetch 4 bits of memory at a time compared to 2 bits in DDR.

BLAST scaled well with increasing processor frequency on the Intel Xeon processors, especially on larger query sizes.

DDR2 transfer speed starts where the current DDR technology ends at 400 MHz. In the future, DDR2 is expected to support 533 and 667 mega-transfers/sec (MT/sec) to enable memory bandwidths of 4.3 GB/sec and 5.3 GB/sec. Currently, only DDR2 at 400 MHz is available in Dell PowerEdge servers, which is the memory technology used in the PowerEdge 1850 system.

Components of the test environment

The goal of the Dell study, which was conducted in July 2004, was to evaluate the impact of processor and memory architecture on the performance of BLAST. Three Dell PowerEdge servers configured similarly in terms of software and compilers were used. The main difference between the servers was in the processor architecture and, to some extent, in the memory technology. The use of three servers allowed the test team to compare three processor architectures, including the impact of processor and FSB (system bus) speeds and the influence of memory technology on BLAST performance.

The PowerEdge 1750 and PowerEdge 1850 servers, which use Intel Xeon processors, had Intel Hyper-Threading Technology turned off. The 90 nm Intel Xeon processors used in the PowerEdge 1850 support 64-bit extensions (EM64T). The 64-bit mode of the EM64T-capable Intel Xeon processor was not used because that was not the focus of this study.

Application background and characteristics

The BLAST family of sequence database-search algorithms serves as the foundation for much biological research. The BLAST algorithms search for similarities between a short query sequence and a large, infrequently changing database of DNA or amino acid sequences.

Version 8.0 Intel compilers for 32-bit applications were used to compile BLAST on the PowerEdge 1750 and PowerEdge 1850 servers. For the Itanium-based PowerEdge 3250 server, version 8.0 Intel compilers for 64-bit applications were used.

BLAST was executed on each of the four configurations—the PowerEdge 1850 was configured with two different processor frequencies, one at 3.2 GHz and one at 3.6 GHz. The test used a database of about 2 million sequences, with about 10 billion total letters. For this study, BLAST was executed against single queries of three lengths: 94,000 words; 206,000 words; and 510,000 words. Runs were conducted using both single and dual threads.

Performance evaluation and analysis

Before testing BLAST performance, the test team used the STREAM benchmark to measure memory bandwidth. STREAM measures real-world bandwidth sustainable from ordinary user programs as opposed to the theoretical peak bandwidth provided by vendors. By running four simple kernels, the benchmark measures traffic all the way from registers to main memory (and vice versa). Because the arrays are much too large to fit in caches, the benchmark measures a mixture of both read and write traffic. STREAM measures *programmer-perceived* bandwidth—that is, sustained bandwidth rather than raw or peak bandwidth.

Figure 2 shows the measured memory bandwidth using the STREAM benchmark. The PowerEdge 3250 server showed significant improvements over the PowerEdge 1750 server because of its wider system bus (128 bits). Similarly, the PowerEdge 1850 server showed improvements over the PowerEdge 3250 thanks to its faster memory clock speed (400 MHz) as well as a faster FSB (800 MHz).

Next, the performance of BLAST was evaluated using different query sizes and running single and dual threads. The importance of processor frequency, architecture, and memory subsystem design can be determined from the results obtained on the four tested system configurations.

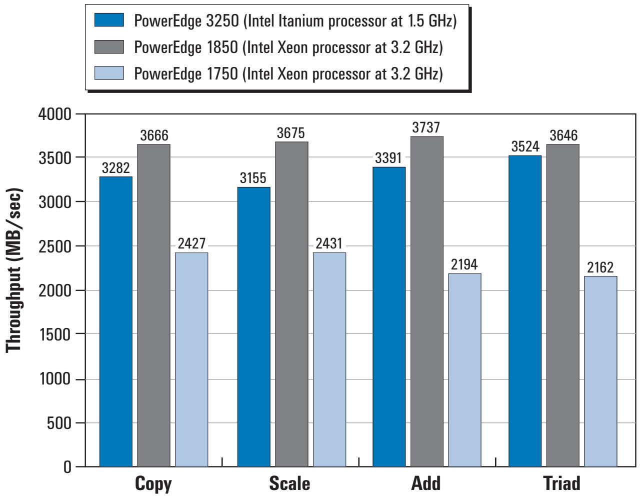


Figure 2. Sustainable memory bandwidth measured using STREAM benchmark

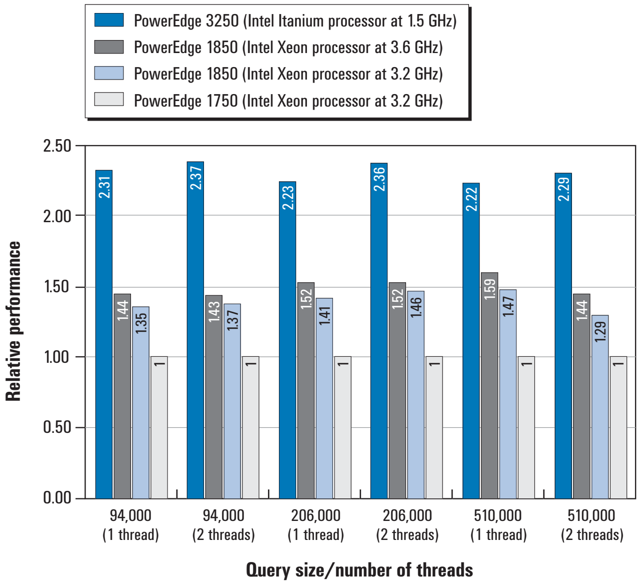


Figure 3. Performance speedup for the PowerEdge 3250 and PowerEdge 1850 servers over the PowerEdge 1750 server

Three query sizes—94,000 words; 206,000 words; and 510,000 words—were chosen to represent small, medium, and large data sets, respectively. The database against which these queries were matched remained constant. The medium and large query sizes were 2.2 and 5.4 times larger than the small query size.

Figure 3 illustrates the relative performance of the PowerEdge 3250 and PowerEdge 1850 servers compared to the PowerEdge 1750 system using dual 130 nm Intel Xeon processors at 3.2 GHz.

The PowerEdge 3250 server with the Itanium 2 processor had the best performance for all query sizes and exhibited speedup ranging from 122 percent to 137 percent over the PowerEdge 1750 server. BLAST scaled more efficiently on the PowerEdge 3250 server with the Itanium 2 processor. Therefore, in Figure 3 the speedup for Itanium 2 is slightly higher for dual-threaded runs. Figure 3 also shows that the frequency increase from 3.2 GHz to 3.6 GHz (13 percent frequency scaling) for the 90 nm Intel Xeon processor resulted only in a 6 to 11 percent improvement for small and medium query sizes. For the larger query size, improvement from frequency scaling was slightly higher at 12 to 15 percent. Therefore, BLAST scaled well with increasing processor frequency on the Intel Xeon processors, especially on larger query sizes.

The impact of memory and processors on BLAST performance

Memory performance is important for BLAST. On the PowerEdge 1750 and PowerEdge 1850 servers, the difference in performance occurred mainly because of the improved DDR2 technology and the faster FSB in the PowerEdge 1850 server. Performance was better on the PowerEdge 1850 server by 29 percent for the 3.2 GHz processor and 44 percent for the 3.6 GHz processor. The 29 percent speedup was primarily caused by the faster memory subsystem, and the 44 percent speedup can be attributed to the faster memory and frequency scaling.

Cache effectiveness is important for BLAST. Larger cache size on the Itanium 2 processor, along with the ability to execute a higher number of integer instructions per cycle, helped achieve speedups of 130 percent on the larger query size.

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

Rizwan Ali is a systems engineer working in the Scalable Systems Group at Dell. His current research interests are performance benchmarking and high-speed interconnects. Rizwan has a B.S. in Electrical Engineering from the University of Minnesota.

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

Kalyana Chadalavada is a senior engineer with the Dell Enterprise Solutions Engineering Group at the Bangalore Development Center. Kalyana has a B.S. in Computer Science and Engineering from Nagarjuna University in India. His current interests include performance characterizations on HPC clusters and processor architectures.

Ramesh Rajagopalan is a lead software engineer in the Database and Applications Engineering team of the Dell Product Group. His current areas of focus include Oracle® Real Application Clusters and performance analysis of Dell clusters. Ramesh has a Bachelor of Engineering in Computer Science from the Indian Institute of Science, Bangalore.

FOR MORE INFORMATION

Top 500 Supercomputer Sites:
www.top500.org

BLAST benchmark:
www.ncbi.nih.gov/BLAST

STREAM benchmark:
www.streambench.org

Eunice, Jonathan. "Itanium 2 Performance: Wow!" Illuminata Research Note. August 27, 2002.

Hsieh, Jenwei, Tau Leng, Victor Mashayekhi, and Reza Rooholamini. "Impact of Level 2 Cache and Memory Subsystem on the Scalability of Clusters of Small-Scale SMP Servers." *IEEE International Conference on Cluster Computing*. Chemnitz, Germany. November 2000.

Intel Corporation. "DDR2 Advantages for Dual Processor Servers." August 2004. www.memforum.org/memorybasics/ddr2/DDR2_Whitepaper.pdf.

Koufaty, David and Deborah T. Marr. "Hyper-Threading Technology in the NetBurst Microarchitecture." *IEEE Micro*. March/April 2003.

Sharangpani, Harash and Ken Arora. "Itanium Processor Microarchitecture." *IEEE Micro*. September/October 2000.

Getting the Best Performance from an HPC Cluster: A STAR-CD Case Study

High-performance computing (HPC) clusters represent a new era in supercomputing. Because HPC clusters usually comprise standards-based, commodity components, they differ primarily in node properties, interconnect configuration, file system type, and clustering middleware. This article explains how a multipurpose and multidiscipline computational fluid dynamics application can be used to understand how the performance of an HPC application may be affected by system components such as interconnects (bandwidth and latency), file I/O, and symmetric multiprocessing.

BY BARIS GULER; JENWEI HSIEH, PH.D.; RAJIV KAPOOR; LANCE SHULER; AND JOHN BENNINGHOFF

High-performance computing (HPC) clusters can deliver supercomputing-class performance using off-the-shelf, industry-standard components. As a result, HPC clusters differ primarily in their node properties, interconnect configuration, file system type, and clustering middleware. Application performance is highly sensitive to many characteristics of an HPC cluster, such as cache size, processor speed, memory latency and bandwidth, interconnect latency and bandwidth, file I/O, and so forth.[1] For example, administrators can configure a cluster with a local disk, Network File System (NFS), or a parallel file system and achieve very different performance and cost values for different applications.

Determining HPC cluster application performance through Dell and Intel collaboration

To determine the effect of various HPC cluster properties on application performance, from January to May 2004 Dell engineers tested the CD adapco Group STAR-CD, a popular industrial computational fluid dynamics (CFD) application, on an HPC cluster comprising Dell™ PowerEdge™ 3250 servers. The STAR-CD solver uses state-of-the-art numerical methodologies to achieve a high level of accuracy for complex unstructured meshes, in both steady and transient simulations.

Dell engineers cross-checked their results from two STAR-CD workloads with tests by the Intel HPC for Independent Software Vendor (HPC ISV) Enabling Team to verify performance and adjust configurations for optimal performance and scaling of STAR-CD. This cross-checking ranged from a comparison of single-node performance to scaling across a large cluster. Differences in benchmark results between Dell and Intel tests exposed performance issues, such as those arising from using a local file system instead of NFS.

[1] For more information about the effects of processor speed and cache size on application performance, see “Understanding the Behavior of Computational Fluid Dynamics Applications on Dell PowerEdge 3250 Servers” by Baris Guler and Rizwan Ali in *Dell Power Solutions*, October 2004.

Application	STAR-CD V3.150A.012
Compiler	Intel® Fortran and C++ Compilers 7.1
Middleware	MPICH 1.2.4 and MPICH-GM 1.2.5..10
Operating system	Red Hat Enterprise Linux AS 2.1, kernel 2.4.18-e.37smp
Protocol	TCP/IP, GM-2
Interconnect	Gigabit Ethernet, Myrinet
Platform	Dell PowerEdge 3250 servers in a 32-node cluster

Figure 1. Architectural stack of the computational test environment

Configuring the test environment

The test environment was based on a cluster comprising 32 Dell PowerEdge 3250 servers interconnected with nonblocking Gigabit Ethernet and Myricom Myrinet. The Gigabit Ethernet interconnect comprised one Myricom M3-E128 switch enclosure populated with 16 Myrinet M3-SW16-8E switch line cards, each having 8 Gigabit Ethernet ports (at 1 Gbps link speed) for 128 ports total. The Myrinet interconnect comprised one Myricom M3-E128 switch enclosure populated with 16 Myrinet M3-SW16-8F switch line cards, each having 8 fiber ports (at 2 Gbps link speed) for 128 ports total.

Each PowerEdge 3250 was configured with two Intel® Itanium® 2 processors running at 1.3 GHz with 3 MB of level 3 (L3) cache and 4 GB of double data rate (DDR) RAM operating on a 400 MHz front-side bus. Each PowerEdge 3250 had an Intel® E8870 chip set Scalable Node Controller (SNC), which accommodates up to eight registered DDR dual in-line memory modules (DIMMs). The operating system was Red Hat® Enterprise Linux® AS 2.1 with kernel version 2.4.18-e.37smp. Intel® Fortran and C++ Compilers 7.1 (Build 20030909) were used to link STAR-CD V3.150A.012 with MPICH 1.2.4 and MPICH-GM 1.2.5..10. Figure 1 shows the architectural stack of the computational test environment.

Measuring application efficiency

The efficiency in parallel applications is usually measured by speedup. The speedup of an application on a cluster with *N* processors is defined as t_1 / t_N, where t_1 is the execution time of the application running on one processor and t_N is the execution time of the application running across *N* processors. In theory, the maximum speedup of a parallel simulation with *N* processors is *N*—that is, the program runs *N* times faster than it would on a single processor. However, in reality, as the number of processors increases, application speedup is usually less than *N*. The disparity between theoretical and actual performance can be attributed to factors that include increasing parallel job initiation, interprocessor communication, file I/O, and network contention.

Running test cases

CD adapco Group provides a suite of six test cases selected to demonstrate the versatility and robustness of STAR-CD in CFD solutions and the relative performance of the application on different hardware platforms and different processor configurations.[2] CD adapco Group selected these test cases to be industry representative, categorizing them as small, medium, and large. Figure 2 lists benchmark names and brief descriptions of the data sets used in this study.

In this study, the small (Engine block) and the large (A-class) test cases were used to perform the analysis and draw the conclusions regarding the application performance and sensitivity to different hardware configurations. Each test case was performed on 1, 2, 4, 8, 16, and 32 processors to help assess the scalability of the application in different hardware configurations, such as using two processors versus one processor on each node, using the local file system versus NFS, and using Gigabit Ethernet versus Myrinet interconnects. All the benchmark runs were conducted using the double-precision version of the code and the conjugate gradient solver.

When running the application over the Myrinet interconnect using MPICH-GM 1.2.5..10, testers had to make sure that the RAMFILES and TURBO environment variables were properly passed to each node because the mpirun shell script does not handle this task the same way that MPICH 1.2.4 does. Initially, benchmark runs using Gigabit Ethernet were much faster than the runs using Myrinet. By setting the RAMFILES and TURBO environment variables, Dell engineers enabled STAR-CD to use a solver code optimized for Itanium 2 architecture and RAMFILES—resulting in improved application performance with the Myrinet interconnect.

Comparing the performance of the local file system versus NFS

Some applications perform several file I/O operations during execution, while other applications do most of the I/O at the end of the simulation and still other applications do very little I/O at all. Direct (local) or remote (NFS) access to the file system can affect the application's performance because file I/O is usually slower than other operations such as computation and communication. To observe the performance impact of NFS compared to the local file system, Dell engineers ran the Engine block and A-class test cases in

Class	Benchmark	Cells	Mesh	Description
Small	Engine block	156,739	Hexahedral	Engine cooling in automobile engine block
Large	A-class	5,914,426	Hybrid	Turbulent flow around A-class car

Figure 2. STAR-CD benchmarks used in the Dell test

[2] For more information about STAR-CD benchmarks and data sets, visit www.cd-adapco.com/support/bench/315/index.htm.

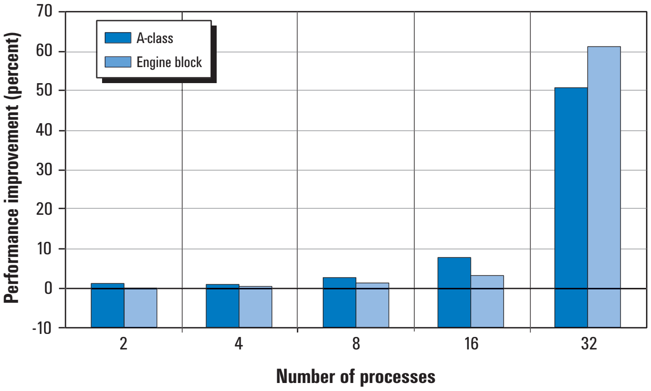


Figure 3. Performance improvement of the local file system compared to NFS

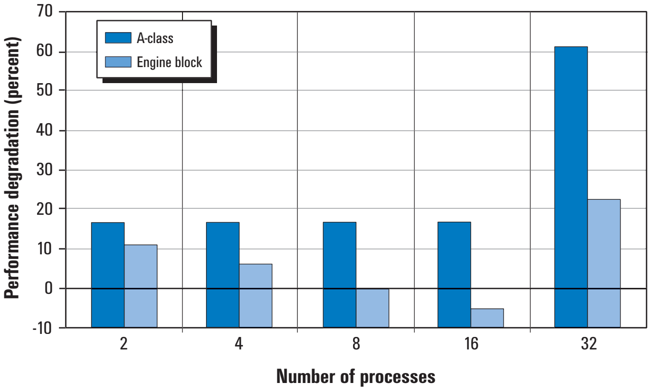


Figure 4. Performance degradation of two processors per node compared to one processor per node

the same cluster test environment with a single processor per node, first using NFS and later using the local file system for the STAR-CD input and output files.

Figure 3 shows the performance gain from using the local file system over NFS for different numbers of processes running on a single-processor node. Clearly, the file system's effect on the application's performance for both test cases was apparent at 16 processes (or processors), and a greater than 50 percent performance improvement was achieved using the local file system at 32 processes (or processors). In other words, using NFS instead of the local file system can drastically decrease the overall application performance at large processor counts. This behavior should be taken into account when designing an HPC cluster for a specific application.

Comparing single-processor-per-node versus dual-processor-per-node performance

A parallel application such as STAR-CD running on two processors (one from each of two nodes) usually delivers better performance than a similar parallel application running on two processors that reside on the same node. The better performance can be attributed primarily to the fact that the application does not need to share memory and I/O devices when it is running on multiple nodes. However, a symmetric multiprocessing (SMP) system in which the processors reside on the same node is often a less-expensive solution because of the many resources that can be shared between two or more processors. Besides sharing internal resources, SMP systems require fewer ports on the network switch because one network interface card (NIC) can be used for each SMP system. The need for fewer ports also helps to decrease the overall cost of an SMP-based cluster.

The A-class and Engine block test cases were executed in the same hardware and software environment. First, only one processor per node was used to perform the simulations. Then, both processors from each node were used to perform the same simulations. Results of this test are shown in Figure 4. The *x*-axis is the number of processes (equivalent to the number of processors) used, and the *y*-axis shows the percent degradation in overall parallel application performance for both test cases when using two processors per node, with respect to using one processor per node. Gigabit Ethernet was used as the cluster interconnect, and the local file system was used to perform file I/O.

Using up to 16 processors, the A-class test case exhibited a performance drop of approximately 17 percent. When 32 processors were used, the drop increased to over 60 percent because interprocessor communication took up a larger percentage of total time than in the 16-processor case. With two communicating processes using a shared resource, namely the network interface, there is an increased potential for contention. Network interfaces with lower latency and higher bandwidth can mitigate some of this contention. Demonstrating this mitigation is the comparison of Myrinet and Gigabit Ethernet performance in the two-processor-per-node case discussed in the section, "Comparing the performance of Gigabit Ethernet versus Myrinet interconnects."

On the other hand, the Engine block test case showed roughly 10 percent degradation when using two processors (one SMP node); the degradation diminished as more processors—up to 16 processors (eight SMP nodes)—were used. Because the same workload decomposed into smaller sub-workloads when more processors were used, each sub-workload started to benefit more from the cache. In addition, the memory contention problem inherent to SMP systems became less significant when each processor worked on a smaller data set. These factors contributed to less performance degradation (from a roughly 10 percent degradation to a 5 percent performance improvement) as indicated in Figure 4. However, when 32 processors were used, more than 20 percent performance degradation was observed because of the negative affect of high-latency interconnects (as discussed in the "Comparing the performance of Gigabit Ethernet versus Myrinet interconnects" section) and increased interprocessor communication.

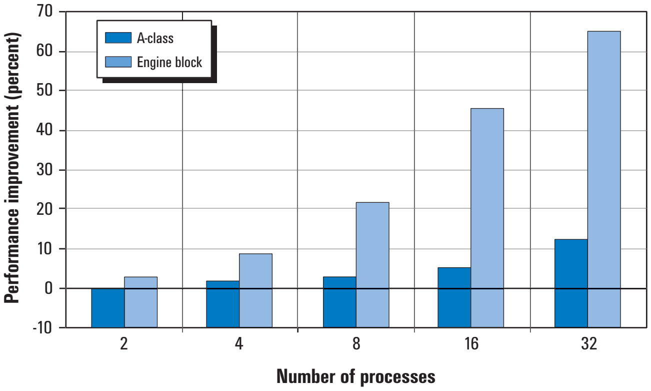


Figure 5. Performance improvement of Myrinet compared to Gigabit Ethernet

Comparing the performance of Gigabit Ethernet versus Myrinet interconnects

The interconnect is a crucial backbone of all clusters, and many interconnect technologies are available today with varying latency, bandwidth, and scalability characteristics. Each application has different communication characteristics and may be more sensitive to latency, bandwidth, both, or none. Dell conducted tests to find out whether a low-latency, high-bandwidth interconnect such as Myrinet is required for STAR-CD or whether Gigabit Ethernet is sufficient. Both test cases were conducted on the cluster using Gigabit Ethernet first and Myrinet second. To isolate the effect of the interconnect from other factors in the cluster, one processor per node was used with the local file system.

Figure 5 shows the percentage improvement in execution times of the test cases using Myrinet compared to the runtimes using Gigabit Ethernet. Obviously, the smaller test case, Engine block, benefited from a lower-latency interconnect—performance improved by more than 20 percent when more than 8 processors were used and up to 65 percent when 32 processors were used. The larger data set, A-class, did not show much improvement over Gigabit Ethernet when Myrinet was used. Even at 32 processors, the application performance improvement from Myrinet was little more than 10 percent. However, these runs were conducted using one processor per node, and results for both interconnects are expected to be much different when two processors per node are used.

Considering clusters of dual-processor or single-processor nodes

The main benefit of using a lower-latency, higher-bandwidth interconnect occurs when both processors from each node are used during simulation. In this situation, both processors on the same node share all resources in the system, including memory, frontside bus, NIC, and hard drive.

Figures 6 and 7 show the speedups for various cluster configurations when executing A-class and Engine block test cases, respectively, using the local file system. In the results for the two-processor-per-node simulations represented by the speedup lines labeled Myrinet (2PPN) and Gigabit Ethernet (2PPN), Myrinet showed much better scalability, especially when more than 16 processors were used with the large data set A-class test case (see Figure 6). When 32 processors were used, the performance improved more than 50 percent—16 times the speedup when Gigabit Ethernet was used and 25 times the speedup when Myrinet was used. However, if the problem size is smaller, as in the Engine block case (see Figure 7), the effect of using a low-latency interconnect appears earlier, beyond four processors. Figure 7 clearly indicates that Gigabit Ethernet does not enable the workload to scale properly beyond eight processors if the workload's size is small. In addition, when 32 processors were used, the speedup improved by more than 100 percent using Myrinet compared to using Gigabit Ethernet as the interconnect.

The results pose a potential question: Should enterprises invest in an HPC cluster of single-processor nodes interconnected with Gigabit Ethernet or an HPC cluster of dual-processor nodes interconnected with Myrinet? These two options may be similar in price because the first option requires twice as many ports in the network switch and additional costs per single-processor node (for example, the cost of buying a second single-processor node is typically more than

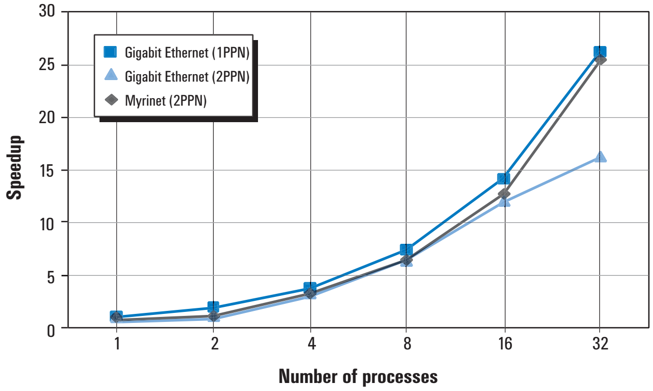


Figure 6. Speedup comparison for A-class test case

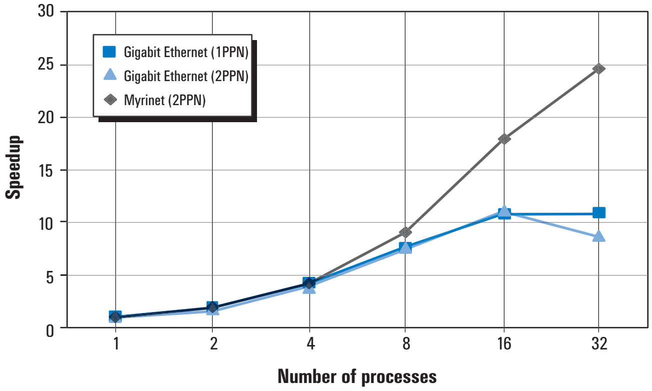


Figure 7. Speedup comparison for Engine block test case

the incremental cost of adding a second CPU), even though a special network such as Myrinet is used for the second option. When these two options are compared in Figures 6 and 7, where they are labeled Gigabit Ethernet (1PPN) and Myrinet (2PPN), the results indicate that both configurations may deliver very similar performance for the larger data set. However, speedup results for the smaller data set indicate that the lower-latency interconnect may be more important than putting two processes on one SMP node or using two separate single-processor nodes interconnected with Gigabit Ethernet. These results suggest that an HPC cluster based on a dual-processor node and a lower-latency, higher-bandwidth interconnect such as Myrinet, InfiniBand, or Quadrics QsNet may be a better choice for the STAR-CD application from a price/performance point of view.

Clusters can be fully optimized for a specific application only if the application's communication, computation, and I/O characteristics are properly addressed.

Optimizing the design of an HPC cluster

Based on the findings in this study, the performance of CFD applications such as STAR-CD may depend on the following factors:

- The larger the problem being solved, the less the need is for a low-latency, high-bandwidth interconnect as long as each processor spends more time computing than communicating.
- A low-latency interconnect becomes very important if a small data set must be solved in parallel using more than four processors, because each processor does very little computation.
- Performance characteristics of an application are highly dependent on the size and nature of the problem being solved.
- With up to 16 processors, SMP affects performance less than 20 percent if the data set is large and uses memory extensively. With more than 16 processors, when communication becomes a large contributor to overall processing time, the SMP use of a shared NIC can degrade performance. Use of a low-latency, high-bandwidth interconnect can help reduce the contention for the shared resource.
- Beyond 16 processors, the choice of file system for output files can make a significant difference in performance—up to 60 percent.

Designing a cluster is an art. Clusters can be fully optimized for a specific application only if the application's communication, computation, and I/O characteristics are properly addressed. Studies such as the one presented in this article can enable designers to plan their clusters for a special purpose or application by providing guidance that helps them decide whether to use single- or multiple-processor nodes; a low-latency, fast interconnect; and additional local storage in each node. In addition, designers can consider the findings in this study and the current prices for HPC cluster components to calculate the price/performance ratio for each configuration before investing in the hardware.

Acknowledgments

The authors would like to acknowledge the CD adapco Group for providing the STAR-CD application, benchmarks (test cases), and license used for this study. Many thanks go to the STAR-CD technical staff—Jean-Yves Tillier, Lucia Game, Dave Robinson, Ron Gray, and Michael Dickens—for their help and contribution to this study. The authors would like to extend their thanks to several members from the Intel HPC ISV Enabling Team; the Intel enterprise marketing team; and the Intel Parallel and Distributed Solutions Division, particularly Michael D'Mello, for their contribution during the course of the study. In addition, the authors appreciate the collaboration from the Intel software teams, particularly Robert A. Chesebrough, in providing the necessary compilers and tools to complete this study.

Baris Guler is a systems engineer and advisor in the Scalable Systems Group at Dell. His current research interests are parallel processing, diskless HPC clusters, performance benchmarking, reservoir engineering and simulation, and numerical methods. Baris has a B.S. in Petroleum and Natural Gas Engineering (PNGE) from the Middle East Technical University in Turkey, and an M.S. in PNGE from Pennsylvania State University. He is currently a Ph.D. candidate in Petroleum and Geosystems Engineering at The University of Texas at Austin.

Jenwei Hsieh, Ph.D., is an engineering manager in the Scalable Systems Group at Dell, where he is responsible for developing HPC clusters. His work in the areas of multimedia computing and communications, high-speed networking, serial storage interfaces, and distributed network computing has been published extensively. Jenwei has a Ph.D. in Computer Science from the University of Minnesota and a B.E. from Tamkang University in Taiwan.

Rajiv Kapoor is a software engineer in the Software Solutions Group at Intel Corporation. His interests are application performance analysis, optimization, and development of related tools.

Lance Shuler manages an Applications Engineering team at Intel focused on technical computing across Intel® Xeon™ processor and Itanium processor platforms. He has worked at Intel for eight years and in the technical computing field for 10 years.

John Benninghoff is a software engineer working on HPC performance analysis in the Performance Marketing team of the Intel Enterprise Products Group. He has worked at Intel for five years and in the software industry for over 20 years. Current projects include running and analyzing synthetic and application-based benchmarks covering many HPC application domains such as CFD, molecular modeling, atmospheric modeling, and Linpack.

Achieving Scalable I/O Performance

in High-Performance Computing Environments

Dell™ PowerEdge™ servers, when combined with a Dell/EMC CX700 storage array and the IBRIX high-performance parallel file system, are designed to provide a scalable, economical cluster solution for high-performance computing applications. This article discusses the performance results obtained using an IBRIX parallel file system on Dell/EMC storage connected to Dell PowerEdge servers.

BY AMINA SAIFY; RAMESH RADHAKRISHNAN, PH.D.; SUDHIR SRINIVASAN, PH.D.; AND ONUR CELEBIOGLU

For organizations implementing coarse-grained parallel solutions in scientific applications, high-performance computing (HPC) clusters made up of standards-based computers can offer a cost-effective alternative to expensive supercomputers. Figure 1 shows a generic HPC cluster that consists of several compute nodes and I/O nodes. The compute nodes perform CPU-intensive tasks while the I/O nodes perform data storage functions. The I/O nodes provide the compute nodes with shared access to files, and also interface with disk arrays, typically over a storage area network (SAN).

Several factors should be considered when designing a high-performance I/O subsystem: servers used as I/O nodes, storage technology, file systems, interconnects, and application behavior. SCSI and Fibre Channel technologies are available for back-end storage, and InfiniBand, Myrinet, and Gigabit Ethernet technologies are available for interconnects. For file systems that connect the entire I/O subsystem, Network File System (NFS), storage area network (SAN) file systems, and parallel file systems are available.

This article discusses the performance results obtained using an IBRIX parallel file system on Dell/EMC CX series storage connected to Dell PowerEdge servers. The configuration used for this performance study—which was conducted in November 2004 by the Dell HPC Cluster (HPCC) team and IBRIX Inc.—included eight Dell PowerEdge 1750 servers, which served as I/O nodes connected to a back-end Fibre Channel Dell/EMC CX700 storage array. The I/O nodes served data from the array to 16 compute nodes connected through a Gigabit Ethernet switch. Each I/O node used a QLogic QLA2340 host bus adapter (HBA) for Fibre Channel connectivity. The results of this study show that the cluster achieved near-linear scalability in read bandwidth (to over 1.1 GB/sec) as the number of I/O nodes was increased from one to eight.

Setting up the cluster

Dell PowerEdge 1750 servers were used for the I/O nodes of the cluster. The PowerEdge 1750 server can have up to two Intel® Xeon™ processors, which feature Intel NetBurst® microarchitecture and Intel Hyper-Threading Technology. This server can support up to 8 GB of 266 MHz error-correcting code (ECC) double data rate (DDR) SDRAM memory for expandable performance. It also offers Chipkill support with 512 KB, 1 GB, and 2 GB dual in-line memory modules (DIMMs) and spare-bank configuration to help ensure memory availability. The PowerEdge 1750 features a ServerWorks GC-LE chip set that utilizes a 533 MHz frontside bus (FSB), 2:1 memory interleaving for fast memory access, and two Peripheral Component Interconnect (PCI) slots. In addition, this server

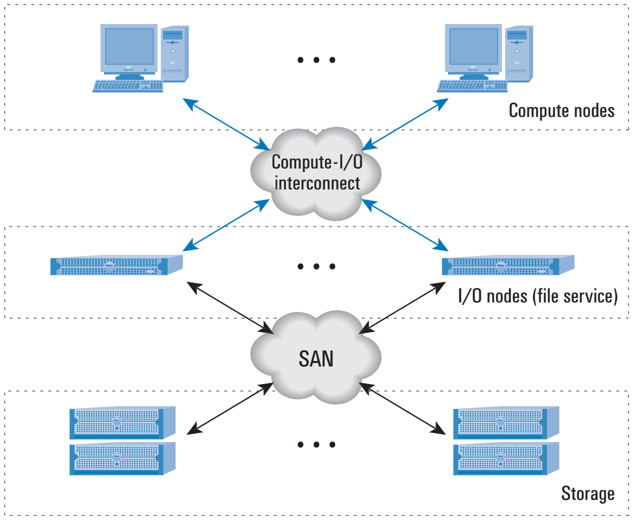


Figure 1. A generic HPC cluster architecture

includes two integrated Gigabit[1] Ethernet network interface cards (NICs) with load balancing, and offers failover support designed to provide fast Ethernet technology and to keep PCI slots open and the server available.

An overview of the Dell/EMC CX700 storage array

The Dell/EMC CX700 storage array features two dual-CPU storage processors, each with four 2 Gbps Fibre Channel loops in the back end to connect up to 240 disks (see Figure 2). The eight back-end loops per array provide a theoretical bandwidth of 1600 MB/sec, of which more than 1400 MB/sec can be achieved by hosts. Each storage processor has 4 GB of RAM, of which up to 3473 MB can be allocated for read or write caching, or both. The write cache is mirrored between the two storage processors to enable durability of writes if a storage processor fails. The array architecture favors storage processor cross-examination of write data over a simple mirrored-memory model to avoid the propagation of erroneous data. In a wide variety of applications, this advanced data protection scheme and the ability to withstand major component failures is generally well worth the trade-off with bandwidth. If a power failure occurs, a battery-backed power supply can flush the write cache to RAID-5 protected disks.

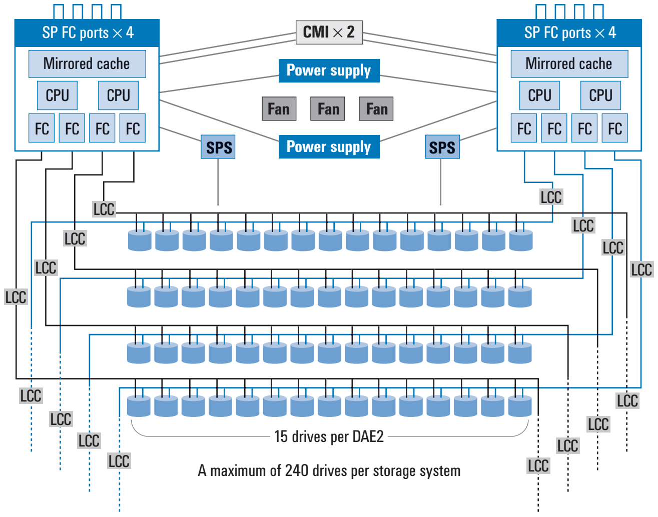
HPC clusters made up of standards-based computers can offer a cost-effective alternative to expensive supercomputers.

On the host side, each storage processor has four Fibre Channel ports, each of which can be connected to HBAs in hosts either directly or through a switch. Hosts communicate with the array using the SCSI over Fibre Channel protocol. Storage processors present logical storage units (LUNs) to I/O nodes. Physically, each LUN resides on a group of disks. Administrators can choose from RAID-1, RAID-3, RAID-5, or RAID-10 for each disk group. Resilience to one storage processor failure is provided by allowing LUN ownership to be transferred to the other storage processor. Host-based multipathing software can then route I/O requests to this storage processor.

Understanding the IBRIX file system architecture

IBRIX has developed a unique architecture called a segmented file system (U.S. Patent no. 6,782,389). In traditional parallel computing terms, this architecture may be described as a loosely-coupled approach to distributing metadata; essentially, the segmented architecture is based on the divide-and-conquer principle. Figure 3 highlights the operating principles of the segmented file system architecture. The following numbered descriptions correspond to the numbering in Figure 3:

1. The "file space" of the file system is a collection of segments. Each segment is simply a repository for files and directories with no implicit namespace relationships among them (specifically, a segment need not be a complete rooted directory tree). Segments can be of any size and different segments can be different sizes.



CMI: CLARiiON® Messaging Interface
DAE2: Dell/EMC 2 GB Disk Array Enclosure
FC: Fibre Channel
LCC: Link control card
SP: Storage processor
SPS: Standby power supply

Figure 2. Architecture of the Dell/EMC CX700 storage array

[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.

2. The location of files and directories within particular segments in the file space is independent of their respective and relative locations in the namespace. Thus, as shown in Figure 3, a directory can be located on one segment while the files contained in that directory are spread over other segments. The selection of segments for placement of files and directories occurs dynamically at the time of file/directory creation as determined by an allocation policy. The allocation policy is set by the system administrator in accordance with the anticipated access patterns and specific criteria relevant to the installation (performance, manageability, and so forth).
3. Individual files can also be distributed across multiple segments, in cases where very high throughput is desired from a single file and access patterns are regular and predictable.
4. Segment servers are assigned responsibility for management of individual segments of the file system. Each segment is assigned to a single segment server and each server can "own" multiple segments, as shown by the color coding in Figure 3. Segment ownership can be migrated from one server to another while the file system is actively in use. To meet growing performance needs, additional servers can be added to the system dynamically without adding more capacity by distributing the ownership of existing segments for proper load balancing and utilization of all servers. Conversely, additional capacity can be added to the file system while in active use without adding more servers—ownership of the additional segments is distributed among existing servers. Servers can be configured with failover protection, with other servers being designated as standby servers that automatically take control of a server's segments if a failure occurs.

Several factors should be considered when designing a high-performance I/O subsystem: servers used as I/O nodes, storage technology, file systems, interconnects, and application behavior.

5. Clients run the applications that use the file system (applications also can run on the segment servers). Clients can access the file system either as a locally mounted cluster file system using the IBRIX driver or using standard network attached storage (NAS) protocols such as NFS and Common Internet File System (CIFS).
6. Use of the IBRIX driver on the client has significant advantages over the NAS approach—specifically, the IBRIX driver is aware of the segmented architecture of the file system and, based on the file/directory being accessed, can route requests directly to the correct segment server, yielding balanced resource utilization and high performance.

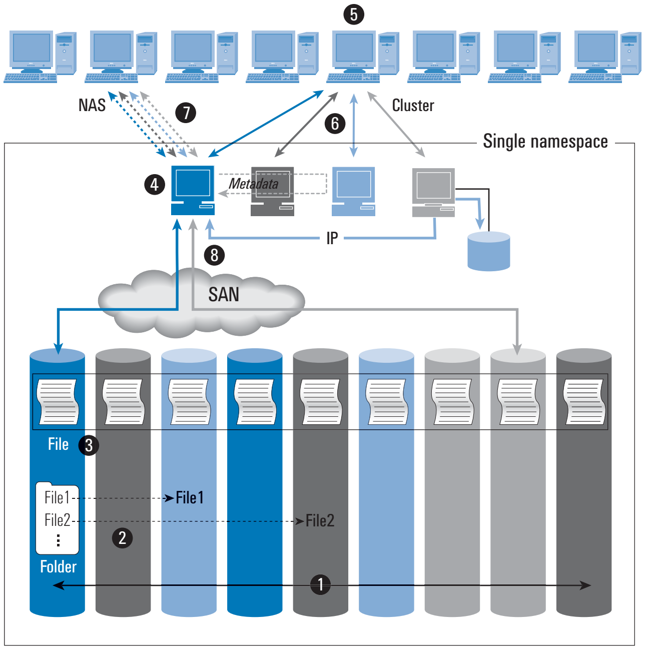


Figure 3. IBRIX segmented file system architecture

7. When using NAS protocols such as NFS and CIFS, a client must mount the file system from one (or more) of the segment servers. All requests are sent to the mounting server that performs the required routing. NAS protocols offer the benefits of multiplatform support and low administration cost for the client software, because the client drivers for these protocols are generally available with the base operating system.
8. Segment servers (and clients using the IBRIX driver) are SAN-friendly. A request at a segment server could be for a file on a segment that is either owned by the server, owned by another server but accessible by this server over the SAN, or owned by another server and not accessible by this server over the SAN. In the second scenario, the server obtains the relevant metadata from the owning server and performs the I/O directly over the SAN. In the third scenario, the I/O is performed through the owning server over the IP network.

The segmented architecture is also the basis for fault resilience—loss of access to one or more segments does not render the entire namespace inaccessible. Individual segments can thus be taken offline temporarily for maintenance operations and then returned to the file system.

IBRIX Fusion is a software-based, fully integrated solution comprising a scalable parallel file system for clusters, a volume manager, high-availability features such as multicomponent automatic failover, and a comprehensive management interface that includes a graphical user interface (GUI) and a command-line interface (CLI). IBRIX enables enterprises to pull together their I/O and storage systems into a single environment that is multipurpose and sharable across a multitude of applications and to manage that environment through a centralized interface. IBRIX Fusion is designed to be deployed in environments scaling to thousands of nodes while achieving linear scalability of aggregate throughput as I/O serving capacity is added.

Studying I/O performance in Dell HPC clusters

Figure 4 presents configuration details of the storage array used in the Dell performance study and the deviations from the default configuration settings. Figure 5 describes configuration of the client/server architecture used for this study.

Figure 6 shows how the clients, servers, and storage array were networked to form a typical HPC cluster configuration. The clients in Figure 6 represent the compute nodes shown in Figure 1. The Dell PowerEdge servers were the I/O nodes, and the CX700 storage array provided the storage. Clients used the IBRIX client software to access files through the IBRIX segment servers running on the I/O nodes. The disks used on the CX700 were evenly divided among the array's eight back-end loops and equally divided across its two storage processors.

Dell PowerEdge servers, combined with a Dell/EMC CX700 storage array and the optimized IBRIX parallel file system, can provide a high-performing, scalable, and economical cluster solution for HPC environments.

System performance was measured using the IOzone benchmarking tool. One IOzone process was run on each client, and the process accessed one file—all files were located in a single directory of the file system. Each process performed 64 KB sequential writes and reads. The bandwidth results reported by the IOzone processes were summed up to calculate the total system bandwidth shown in Figure 7. This method of calculating total bandwidth results in total bandwidth that is 1 to 2 percent higher than that reported by the storage system. This difference occurs because of the variation in bandwidth achieved by the different IOzone processes.

Figure 7 illustrates the total read and write bandwidth obtained when a varying number of clients accessed the storage

CX700 storage processor settings	Hardware configuration
Array-wide settings Memory: 4 GB per storage processor Read cache: 1168 MB for storage processor A; 1168 MB for storage processor B Write cache: 2048 MB Cache page size: 16 KB Low watermark: 40 High watermark: 60	**Disk drives** Eighty 10,000 rpm Fibre Channel drives
Prefetch settings for each LUN Prefetch multiplier: 4 Segment multiplier: 4 Maximum prefetch: 4,096 (in blocks)	**LUNs** 16 LUNs (each LUN comprises five drives configured using RAID-3) Element size: 128
	CX700 management software EMC® Navisphere® 6.6

Figure 4. Hardware and software configuration of the storage subsystem used in the Dell performance study

Segment servers (8)	Clients (16)
Each Dell PowerEdge 1750 One Intel Xeon processor at 3.06 GHz 533 MHz FSB 512 MB L2 cache 2 GB DDR at 266 MHz Operating system: Red Hat® Enterprise Linux® 3, Update 2 HBA: QLogic QLA2340 (driver 6.07.02-RH2 with SG_SEGMENTS set to 64) File system: IBRIX 1.3	**Each Dell PowerEdge 1750** Two Intel Xeon processors at 3.06 GHz 533 MHz FSB 512 MB L2 cache 2 GB DDR at 266 MHz Operating system: Red Hat Enterprise Linux 3, Update 2 File system: IBRIX 1.3

Figure 5. Hardware and software configuration of the clients and segment servers used in the Dell performance study

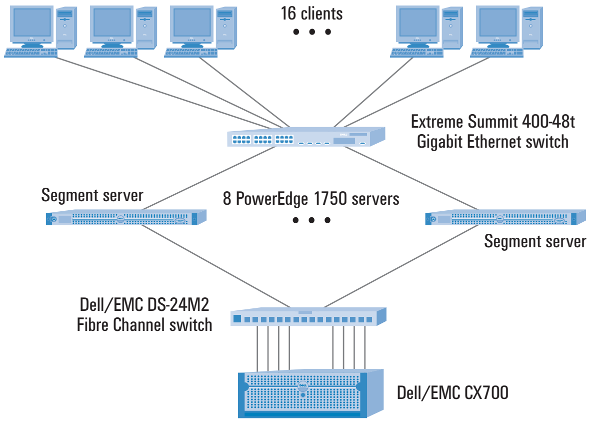


Figure 6. Cluster configuration used in the Dell performance study

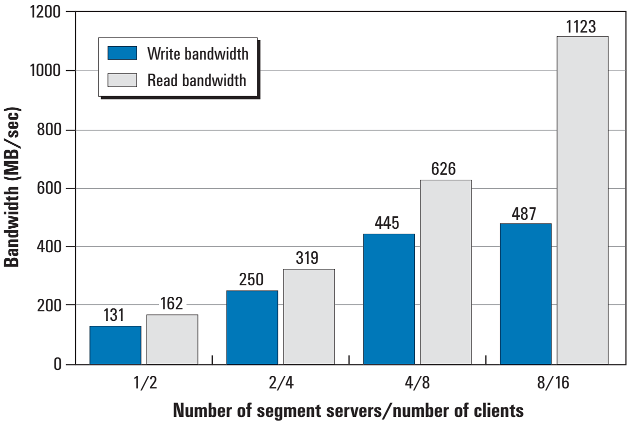


Figure 7. Measured bandwidth for the I/O subsystem using IOzone

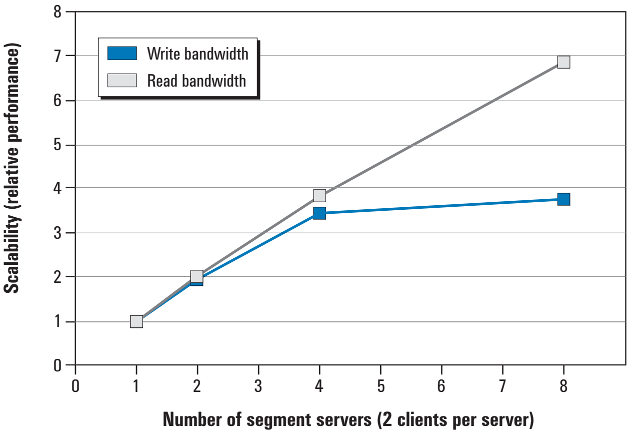


Figure 8. Performance scalability of the I/O subsystem

subsystem using one or more segment servers. Four configurations were used to test the scalability of the I/O subsystem—one segment server serving data to 2 clients, two segment servers for 4 clients, four segment servers for 8 clients, and eight segment servers for 16 clients. The I/O read bandwidth ranged from 162 MB/sec for the single segment server to 1123 MB/sec for the eight segment servers. Similarly, the write bandwidth ranged from 131 MB/sec to 487 MB/sec for the different configuration sizes of the cluster.

Figure 8 shows the scalability of the I/O subsystem as the size of the cluster increased from 2 clients to 16 clients and the number of segment servers increased from one to eight—maintaining a 2:1 client-to-segment-server ratio. These results show near-linear scalability of read bandwidth. Write bandwidth scales well for up to four segment servers, but beyond that, it is limited by overhead on the CX700 that results from the steps the array takes to provide data protection (as described in the "An overview of the Dell/EMC CX700 storage array" section in this article). The CX700 is designed to ensure that data written to the array (even data written only to the write cache) will survive any single failure. If I/Os to the array satisfy certain alignment and size conditions, then the write cache can be bypassed and systems can achieve higher write bandwidth than that obtained in this study.

Building a high-performance computing cluster with Dell systems

The Dell HPCC team used a Dell/EMC storage array and the IBRIX file system to help evaluate the performance scalability of an I/O subsystem for a commonly used HPC cluster scenario. The findings of this study indicate that Dell PowerEdge servers, combined with a Dell/EMC CX700 storage array and the optimized IBRIX parallel file system, can provide a high-performing, scalable, and economical cluster solution for HPC environments.

Acknowledgments

The authors would like to thank the Dell HPCC team and Thomas Eastham, Mike Wolak, and Wayne Paquette from IBRIX Inc. for their immense help in writing this article.

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.

Ramesh Radhakrishnan, Ph.D., is a systems engineer on the Dell HPCC team. His areas of interest are computer architecture and performance analysis. Ramesh has a Ph.D. in Computer Engineering from The University of Texas at Austin.

Sudhir Srinivasan, Ph.D., is the chief technology officer of IBRIX Inc. His interests include storage systems, operating systems, and distributed computing. He has a Ph.D. in Computer Science from the University of Virginia.

Onur Celebioglu is an engineering manager in the Scalable Systems Group at Dell and is responsible for developing HPC clustering products. His current areas of focus are networking and HPC interconnects. Onur has an M.S. in Electrical and Computer Engineering from Carnegie Mellon University.

FOR MORE INFORMATION

Dell HPC clusters:
www.dell.com/hpcc

Planning Considerations for Job Scheduling in HPC Clusters

As cluster installations continue growing to satisfy ever-increasing computing demands, advanced schedulers can help improve resource utilization and quality of service. This article discusses issues related to job scheduling on clusters and introduces scheduling algorithms to help administrators select a suitable job scheduler.

BY SAEED IQBAL, PH.D.; RINKU GUPTA; AND YUNG-CHIN FANG

Cluster installations primarily comprise two types of standards-based hardware components—servers and networking interconnects. Clusters are divided into two major classes: high-throughput computing clusters and high-performance computing clusters. High-throughput computing clusters usually connect a large number of nodes using low-end interconnects. In contrast, high-performance computing clusters connect more powerful compute nodes using faster interconnects than high-throughput computing clusters. Fast interconnects are designed to provide lower latency and higher bandwidth than low-end interconnects.

These two classes of clusters have different scheduling requirements. In high-throughput computing clusters, the main goal is to maximize throughput—that is, jobs completed per unit of time—by reducing load imbalance among compute nodes in the cluster. Load balancing is particularly important if the cluster has heterogeneous compute nodes. In high-performance computing clusters, an additional consideration arises: the need to minimize communication overhead by mapping applications appropriately to the available compute nodes. High-throughput computing clusters are suitable for executing loosely coupled parallel or distributed applications, because such applications do not have high communication requirements among compute nodes during execution time. High-performance computing clusters are more suitable for tightly coupled parallel applications, which have substantial communication and synchronization requirements.

A resource management system manages the processing load by preventing jobs from competing with each other for limited compute resources. Typically, a resource management system comprises a resource manager and a job scheduler (see Figure 1). Most resource managers have an internal, built-in job scheduler, but system administrators can usually substitute an external scheduler for the internal scheduler to enhance functionality. In either case, the scheduler communicates with the resource manager to obtain information about queues, loads on compute nodes, and resource availability to make scheduling decisions.

Usually, the resource manager runs several daemons on the master node and compute nodes including a scheduler daemon, which typically runs on the master node. The resource manager also sets up a queuing system for users to submit jobs—and users can query the resource manager to determine the status of their jobs. In addition, a resource manager maintains a list of available compute resources and reports the status of previously submitted jobs to the user. The resource manager helps organize submitted jobs based on priority, resources requested, and availability.

As shown in Figure 1, the scheduler receives periodic input from the resource manager regarding job queues and available resources, and makes a schedule that determines the order in which jobs will be executed. This is done while maintaining job priority in accordance with the site policy that the administrator has established for the amount and timing of resources used to execute jobs. Based on that information, the scheduler decides which job will execute on which compute node and when.

Understanding job scheduling in clusters

When a job is submitted to a resource manager, the job waits in a queue until it is scheduled and executed. The time spent in the queue, or *wait time,* depends on several factors including job priority, load on the system, and availability of requested resources. *Turnaround time* represents the elapsed time between when the job is submitted and when the job is completed; turnaround time includes the wait time as well as the job's actual execution time. *Response time* represents how fast a user receives a response from the system after the job is submitted.

Resource utilization during the lifetime of the job represents the actual useful work that has been performed. *System throughput* is defined as the number of jobs completed per unit of time. *Mean response time* is an important performance metric for users, who expect minimal response time. In contrast, system administrators are concerned with overall resource utilization because they want to maximize system throughput and return on investment (ROI), especially in high-throughput computing clusters.

In a typical production environment, many different jobs are submitted to clusters. These jobs can be characterized by factors such as the number of processors requested (also known as job size, or *job width*), estimated runtime, priority level, parallel or distributed execution, and specific I/O requirements. During execution, large jobs can occupy significant portions of a cluster's processing and memory resources.

System administrators can create several types of queues, each with a different priority level and quality of service (QoS). To make intelligent schedules, however, schedulers need information regarding job size, priority, expected execution time (indicated by the user), resource access permission (established by the administrator), and resource availability (automatically obtained by the scheduler).

In high-performance computing 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. The scheduler needs to ensure that nodes scheduled to execute parallel jobs are connected by fast interconnects to minimize the associated communication overhead. For parallel jobs, the job efficiency also affects resource utilization. To achieve high resource utilization for parallel jobs, both job efficiency and advanced scheduling are required. Efficient job processing depends on effective application design.

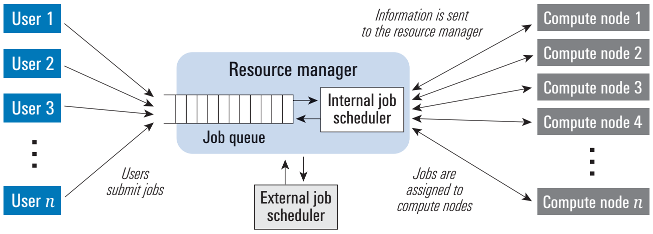


Figure 1. Typical resource management system

Under heavy load conditions, the capability to provide a fair portion of the cluster's resources to each user is important. This capability can be provided by using the *fair-share* strategy, in which the scheduler collects historical data from previously executed jobs and uses the historical data to dynamically adjust the priority of the jobs in the queue. The capability to dynamically make priority changes helps ensure that resources are fairly distributed among users.

Most job schedulers have several parameters that can be adjusted to control job queues and scheduling algorithms, thus providing different response times and utilization percentages. Usually, high system utilization also means high average response time for jobs—and as system utilization climbs, the average response time tends to increase sharply beyond a certain threshold. This threshold depends on the job-processing algorithms and job profiles. In most cases, improving resource utilization and decreasing job turnaround time are conflicting considerations. The challenge for IT organizations is to maximize resource utilization while maintaining acceptable average response times for users.

Figure 2 summarizes the desirable features of job schedulers. These features can serve as guidelines for system administrators as they select job schedulers.

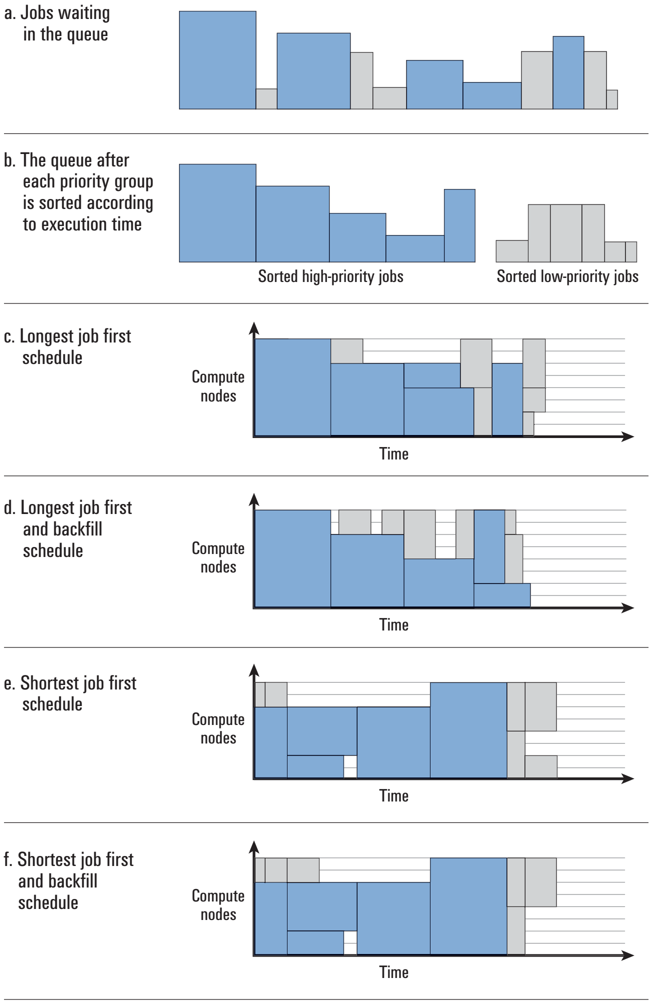
Using job scheduling algorithms

The parallel and distributed computing community has put substantial research effort into developing and understanding job scheduling algorithms. Today, several of these algorithms have been implemented in both commercial and open source job schedulers. Scheduling algorithms can be broadly divided into two classes: time-sharing and space-sharing. *Time-sharing* algorithms divide time on a processor into several discrete intervals, or *slots.* These slots are then assigned to unique jobs. Hence, several jobs at any given time can share the same compute resource. Conversely, *space-sharing* algorithms give the requested resources to a single job until the job completes execution. Most cluster schedulers operate in space-sharing mode.

Common, simple space-sharing algorithms are first come, first served (FCFS); first in, first out (FIFO); round robin (RR); shortest job first (SJF); and longest job first (LJF). As the names suggest, FCFS and FIFO execute jobs in the order in which they enter the queue. This is a very simple strategy to implement, and works acceptably well with a low job load. RR assigns jobs to nodes as they arrive in

Feature	Comments
Broad scope	The nature of jobs submitted to a cluster can vary, so the scheduler must support batch, parallel, sequential, distributed, interactive, and noninteractive jobs with similar efficiency.
Support for algorithms	The scheduler should support numerous job-processing algorithms—including FCFS, FIFO, SJF, LJF, advance reservation, and backfill. In addition, the scheduler should be able to switch between algorithms and apply different algorithms at different times—or apply different algorithms to different queues, or both.
Capability to integrate with standard resource managers	The scheduler should be able to interface with the resource manager in use, including common resource managers such as Platform LSF, Sun Grid Engine, and OpenPBS (the original, open source version of Portable Batch System).
Sensitivity to compute node and interconnect architecture	The scheduler should match the appropriate compute node architecture to the job profile—for example, by using compute nodes that have more than one processor to provide optimal performance for applications that can use the second processor effectively.
Scalability	The scheduler should be capable of scaling to thousands of nodes and processing thousands of jobs simultaneously.
Fair-share capability	The scheduler should distribute resources fairly under heavy conditions and at different times.
Efficiency	The overhead associated with scheduling should be minimal and within acceptable limits. Advanced scheduling algorithms can take time to run. To be efficient, the scheduling algorithm itself must spend less time running than the expected saving in application execution time from improved scheduling.
Dynamic capability	The scheduler should be able to add or remove compute resources to a job on the fly—assuming that the job can adjust and utilize the extra compute capacity.
Support for preemption	Preemption can occur at various levels; for example, jobs may be suspended while running. Checkpointing—that is, the capability to stop a running job, save the intermediate results, and restart the job later—can help ensure that results are not lost for very long jobs.

Figure 2. Features of job schedulers


a. Jobs waiting in the queue

b. The queue after each priority group is sorted according to execution time

c. Longest job first schedule

d. Longest job first and backfill schedule

e. Shortest job first schedule

f. Shortest job first and backfill schedule

Figure 3. Job scheduling algorithms

the queue in a cyclical, round-robin manner. SJF periodically sorts the incoming jobs and executes the shortest job first, allowing short jobs to get a good turnaround time. However, this strategy may cause delays for the execution of long (large) jobs. In contrast, LJF commits resources to longest jobs first. The LJF approach tends to maximize system utilization at the cost of turnaround time.

Basic scheduling algorithms such as these can be enhanced by combining them with the use of advance reservation and backfill techniques. *Advance reservation* uses execution time predictions provided by the users to reserve resources (such as CPUs and memory) and to generate a schedule. The *backfill* technique improves space-sharing scheduling. Given a schedule with advance-reserved, high-priority jobs and a list of 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 improves system utilization by running low-priority jobs in between high-priority jobs. To use backfill, the scheduler requires a runtime estimate of the small jobs, which is supplied by the user when jobs are submitted.

Figure 3 illustrates the use of the basic algorithms and the enhancements discussed in this article. Figure 3a shows a queue with 11 jobs waiting; the queue has both high-priority and low-priority jobs. Figure 3b shows these jobs sorted according to their estimated execution time.

The example in Figure 3 assumes an eight-processor cluster and considers only two parameters: the number of processors and the estimated execution time. This figure shows the effects of generating schedules using the LJF and SJF algorithms with and without backfill techniques. Sections c through f of Figure 3 indicate that backfill can improve schedules generated by LJF and SJF, either by increasing utilization, decreasing response time, or both. To generate the schedules shown, the low- and high-priority jobs are sorted separately.

Examining a commercial resource manager and an external job scheduler

This section introduces scheduling features of a commercial resource manager, Load Sharing Facility (LSF) from Platform Computing, and an open source job scheduler, Maui.

Platform Load Sharing Facility resource manager

Platform LSF is a popular resource manager for clusters. Its focus is to maximize resource utilization within the constraints of local administration policies. Platform Computing offers two products: Platform LSF and Platform LSF HPC. LSF is designed to handle a broad range of job types such as batch, parallel, distributed, and interactive. LSF HPC is optimized for HPC parallel applications by providing additional facilities for intelligent scheduling, which enables different QoS in different queues. LSF also implements a hierarchical fair-share scheduling algorithm to balance resources among users under all load conditions.

Platform LSF has built-in schedulers that implement advanced scheduling algorithms to provide easy configurability and high reliability for users. In addition to basic scheduling algorithms, Platform LSF uses advanced techniques like advance reservation and backfill.

Platform LSF and Platform LSF HPC both have a dynamic scheduling decision mechanism. The scheduling decisions under this mechanism are based on processing load. Based on these decisions, jobs can be migrated among compute nodes or rescheduled. Loads can also be balanced among compute nodes in heterogeneous environments. These features make Platform LSF suitable for a broad range of HPC applications. In addition, Platform LSF can dynamically migrate jobs among compute nodes. Platform LSF can also have multiple scheduling algorithms applied to different queues simultaneously. Platform LSF HPC can make intelligent scheduling decisions based on the features of advanced interconnect networks, thus enhancing process mapping for parallel applications.

The term *resource* has a broad definition in Platform LSF and Platform LSF HPC. Resources can be CPUs, memory, storage space, or software licenses. (In some sectors, software licenses are expensive and are considered a valuable resource.)

Platform LSF and Platform LSF HPC each have an extensive advance reservation system that can reserve different kinds of resources. In some distributed applications, many instances of the same application are required to perform parametric studies. Platform LSF and Platform LSF HPC allow users to submit a job group that can contain a large number of jobs, making parametric studies much easier to manage.

Platform LSF and Platform LSF HPC can interface with external schedulers such as Maui. External schedulers can complement features of the resource manager and enable sophisticated scheduling. For example, using Platform LSF HPC, the hierarchical fair-share algorithm can dynamically adjust priorities and feed these priorities to Maui for use in scheduling decisions.

Maui job scheduler

Maui is an advanced open source job scheduler that is specifically designed to optimize system utilization in policy-driven, heterogeneous HPC environments. Its focus is on fast turnaround of large parallel jobs, making the Maui scheduler highly suitable for HPC. Maui can work with several common resource managers including Platform LSF and Platform LSF HPC, and potentially improve scheduling performance compared to built-in schedulers.

Maui has a two-phase scheduling algorithm. During the first phase, the high-priority jobs are scheduled using advance reservation. In the second phase, a backfill algorithm is used to schedule low-priority jobs between previously scheduled jobs. Maui uses the fair-share technique when making scheduling decisions based on job history. *Note:* Maui's internal behavior is based on a single, unified queue. This maximizes the opportunity to utilize resources.

Typically, users are guaranteed certain QoS, but Maui gives a significant amount of control to administrators—allowing local policies to control access to resources, especially for scheduling. For example, administrators can enable different QoS and access levels to users and jobs, which can be preemptively identified. Maui uses a tool called QBank for allocation management. QBank allows multisite control over the use of resources. Another Maui feature allows charge rates (the amount users pay for compute resources) to be based on QoS, resources, and time of day. Maui is scalable to thousands of jobs, despite its nondistributed scheduler daemon, which is centralized and runs on a single node.

Maui supports job preemption, which can occur under several conditions. High-priority jobs can preempt lower-priority or backfill jobs if resources to run the high-priority jobs are not available. In some cases, resources reserved for high-priority jobs can be used to run low-priority jobs when no high-priority jobs are in the queue. However, when high-priority jobs are submitted, these low-priority jobs can be preempted to reclaim resources for high-priority jobs.

Maui has a simulation mode that can be used to evaluate the effect of queuing parameters on the scheduler performance. Because each HPC environment has a unique job profile, the parameters of the queues and scheduler can be tuned based on historical logs to maximize scheduler performance.

Satisfying ever-increasing computing demands

As cluster sizes scale to satisfy growing computing needs in various industries as well as in academia, advanced schedulers can help maximize resource utilization and QoS. The profile of jobs, the nature of computation performed by the jobs, and the number of jobs submitted can help determine the benefits of using advanced schedulers.

Saeed Iqbal, Ph.D., is a systems engineer and advisor in the Scalable Systems Group at Dell. He has a Ph.D. in Computer Engineering from The University of Texas at Austin.

Rinku Gupta is a systems engineer and advisor in the Scalable Systems Group at Dell. She 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.

Yung-Chin Fang is a senior consultant in the Scalable Systems Group at Dell. He specializes in cyberinfrastructure management and high-performance computing.

Understanding the Scalability of NWChem in HPC Environments

Dell™ PowerEdge™ servers can provide a suitable platform for deployment of applications that have been designed for efficient scaling on parallel systems. NWChem, a compute-intensive computational chemistry package, is one such application that can benefit from the performance and computational power provided by high-performance computing (HPC) clusters. This article introduces NWChem and explains the advantages of running NWChem on HPC clusters versus a single node.

BY MUNIRA HUSSAIN; RAMESH RADHAKRISHNAN, PH.D.; AND KALYANA CHADALAVADA

Clusters of standards-based computer systems have become a popular choice for building cost-effective, high-performance parallel computing platforms. High-performance computing (HPC) clusters typically consist of a set of symmetric multiprocessing (SMP) systems connected with a high-speed network interconnect into a single computational unit. The rapid advancement of microprocessor technologies and high-speed interconnects has facilitated many successful deployments of HPC clusters. HPC technology is now employed in many domains, including scientific computing applications such as weather modeling and fluid dynamics as well as commercial applications such as financial modeling and 3-D imaging.

HPC is traditionally associated with RISC-based systems. However, expensive, proprietary RISC-based systems can be difficult to afford for small-scale research establishments and academic institutions on tight budgets. Building an HPC cluster with standards-based Intel® processors such as 32-bit Intel Xeon™ processors or 64-bit Intel Itanium® processors and standards-based, off-the-shelf components can offer several advantages—including optimal performance, minimal costs, and freedom to mix and match technologies for an excellent price/performance ratio. In addition, HPC clusters built from industry-standard components are tolerant to component failures because they have no single point of failure, thus enhancing system availability and reliability.

NWChem,[1] an open source application developed and supported by the Pacific Northwest National Laboratory, is designed to use the resources of high-performance parallel supercomputers as well as HPC clusters built from standards-based systems. The application is built on the concepts of object-oriented programming and non-uniform memory access (NUMA). This architecture allows considerable flexibility in the manipulation and distribution of data on shared memory, distributed memory, and massively parallel hardware architectures—and hence, maps well to the architecture of HPC cluster systems.

[1] For more information about NWChem, visit www.emsl.pnl.gov/docs/nwchem/nwchem.html.

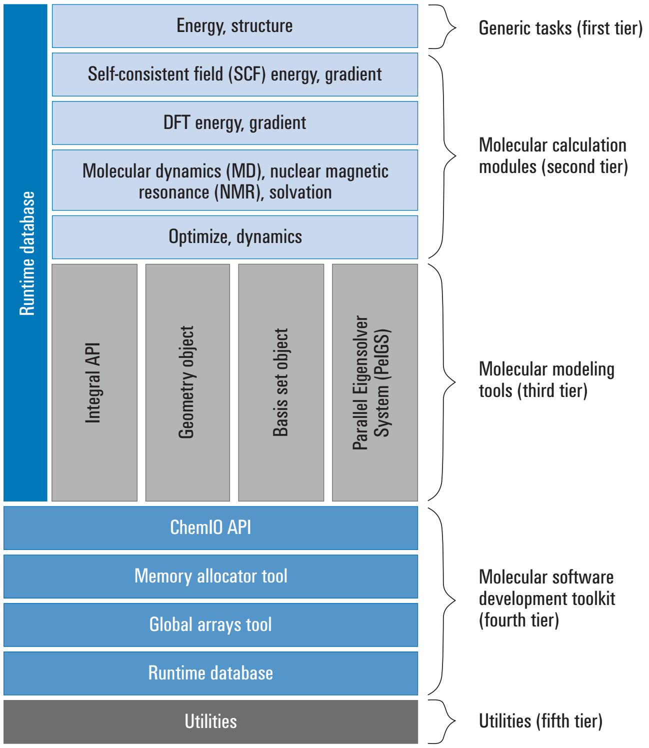

Runtime database
Energy, structure
Self-consistent field (SCF) energy, gradient
DFT energy, gradient
Molecular dynamics (MD), nuclear magnetic resonance (NMR), solvation
Optimize, dynamics
Integral API
Geometry object
Basis set object
Parallel Eigensolver System (PeIGS)
ChemIO API
Memory allocator tool
Global arrays tool
Runtime database
Utilities
Generic tasks (first tier)
Molecular calculation modules (second tier)
Molecular modeling tools (third tier)
Molecular software development toolkit (fourth tier)
Utilities (fifth tier)

Figure 1. Five-tier NWChem architecture

Understanding the architecture of NWChem

Molecular dynamics simulations such as NWChem are an important tool in the study and rational design of molecular systems and materials, providing information about the behavior of chemical systems that can be difficult to obtain by other means. Considerable computational resources are required even for small molecular systems, which have tens of thousands of atoms and short simulation periods in the range of nanoseconds.

NWChem has a five-tier, modular architecture (see Figure 1). The application incorporates several existing modules and toolkits at various levels in the architecture because of its adherence to object-oriented programming concepts. The five tiers are as follows:

- **Generic tasks:** At the first tier, the NWChem interface processes the input, sets up the parallel environment, and performs any initialization needed for the desired calculations. This tier basically serves as the mechanism that transfers control to the different modules in the second tier.
- **Molecular calculation modules:** At the second tier, these high-level programming modules accomplish computational tasks, performing particular operations using the specified theories defined by the input. Each module in this layer uses toolkits and routines that reside in the lower layers of the architecture to accomplish its tasks. These NWChem modules are independent and share data only through a disk-resident database, which allows modules to share data or share access to files containing data.
- **Molecular modeling tools:** The third tier contains tools that provide a basic functionality common to many of the algorithms used in the field of chemistry. These include symmetry, basis sets, grids, geometry, and integrals.
- **Molecular software development toolkit:** This fourth-tier toolkit makes up the foundation level of the five-tiered structure of the NWChem code, and enables the development of an object-oriented code that is constructed mainly in Fortran77.
- **Utilities:** At the lowest level of the NWChem architecture, the fifth tier contains several basic, odds-and-ends functions, including utility routines that most of the higher tiers require. Examples include the timing routines, the input parser, and the print routines.

To enable all aspects of the hardware—such as CPU, disk, and memory—to scale to a massively parallel computing architecture, NWChem uses NUMA to distribute the data across all nodes. Memory access is achieved through explicit message passing using the TCGMSG interface.[2] TCGMSG is a toolkit for writing portable parallel programs using the Message Passing Interface (MPI), a message-passing model. Designed for chemical applications, TCGMSG is relatively simple, having limited functionality that includes point-to-point communication, global operations, and a simple load-balancing facility. This simplicity contributes to the robustness of TCGMSG and its exemplary portability, as well as its high performance in a wide range of problem sizes.

HPC clusters built from industry-standard components are tolerant to component failures because they have no single point of failure, thus enhancing system availability and reliability.

Applications written using TCGMSG can be ported between environments without changes to the parallel code. The memory allocator (MA) tool is used to allocate memory that is local to the calling process. The global arrays (GA) tool is used to share arrays between processors as if the memory were physically shared. This

[2] For more information about the TCGMSG message-passing library, visit www.emsl.pnl.gov/docs/parsoft/tcgmsg/tcgmsg.html.

provides sufficient transparency to the programmer and is compatible with the message-passing model.

The complex I/O patterns required to accomplish efficient memory management are handled using the abstract programming interface ChemIO. ChemIO is a high-performance I/O application programming interface (API) designed to meet the requirements of large-scale computational chemistry problems. It allows the programmer to create I/O files that may be local or distributed.

Like TCGMSG, the runtime database is a component of the fourth tier. The runtime database is a persistent data storage mechanism designed to hold calculation-specific information for all the upper-level programming modules. Because it is not destroyed at the end of a calculation unless the user specifically requests its destruction, a given runtime database can be used in several independent calculations.

Examining the performance characteristics of NWChem

Every application behaves differently and has distinctive characteristics. For the study described in this article, which was conducted in November 2003, baseline performance was measured when running NWChem on a single Intel Itanium processor–based Dell PowerEdge 3250 server at various processor speeds and cache sizes; speedup was evaluated when running the application on multiple nodes in a Dell HPC cluster. The study also addressed the performance impact of running applications like NWChem on various configurations of Intel Itanium processor–based systems to obtain a better understanding of NWChem's dependencies on processor features such as cache size and clock frequency. Understanding the behavior and patterns of the NWChem application can help organizations identify how to design and allocate appropriate resources using standard data sets. This in turn helps to identify bottlenecks while running NWChem on a cluster.

All tests for this study were conducted using NWChem 4.5. The input file used for this study was siosi3.nw, a density functional theory (DFT) benchmark that calculates the DFT function on the siosi3 molecule and provides as output the various atomic energies related to the DFT function. This input file is publicly available and downloadable from the NWChem home page (www.emsl.pnl.gov/docs/nwchem/nwchem.html). For comparison purposes, the input file was kept constant throughout this study. Wall-clock time—that is, the real running time of the program from start to finish (in seconds)—was used to measure performance.

Test configuration	L3 cache	Processor clock frequency
1	4 MB	1.4 GHz
2	1.5 MB	1.4 GHz
3	1.5 MB	1.0 GHz

Figure 2. Test configurations comprising variable clock speeds and cache sizes

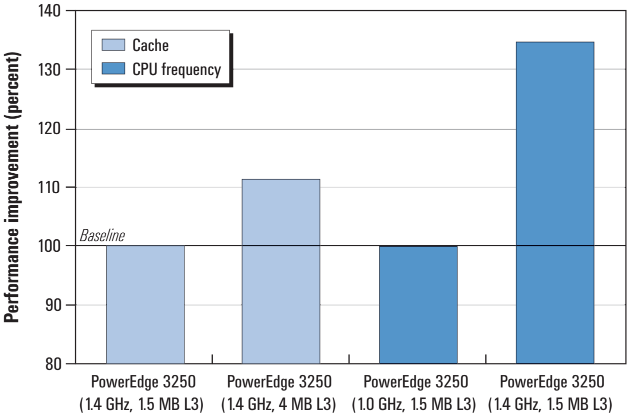


Figure 3. Impact of cache size and processor speed on NWChem performance

Measuring the effect of Itanium 2 processor features

To understand the sensitivity of the NWChem application to cache size and processor frequency, the Dell High-Performance Computing Cluster team conducted tests using the default data set (siosi3.nw). The team conducted multiple tests for various configurations such as constant processor clock frequency and different cache size versus constant cache size and different clock frequency.

Three configurations using 64-bit Intel Itanium processors were tested as follows: The first two test configurations represented the same clock speed but different level 3 (L3) cache sizes, while the second two test configurations represented the same cache size but different clock speeds (see Figure 2).

The results in Figure 3 were obtained on a single processor using the Red Hat® Enterprise Linux® AS 2.1 operating system with kernel 2.4.18-e31smp. For this particular test, Dell engineers used the precompiled binaries with default optimization that are available from the Red Hat Web site.

Figure 3 shows that when the clock speed was kept at 1.4 GHz and the L3 cache size increased from 1.5 MB to 4 MB, performance increased approximately 11 percent. Thus, a larger cache size helped achieve a performance improvement in this study when running NWChem's siosi3.nw benchmark. When running other data sets, the performance benefits from large caches will likely depend on the size of the data set. Larger data sets typically benefit more than smaller data sets from larger caches.

Figure 3 also shows performance benefits when the L3 cache size was kept constant at 1.5 MB and the CPU frequency was increased from 1.0 GHz to 1.4 GHz (a 40 percent increase in processor clock speed). The percentage performance gain for moving from the slower clock frequency to the higher clock frequency was approximately 35 percent. Thus, the results from this study signify that the NWChem application is highly compute intensive and can benefit from increasing processor clock speed because NWChem performance scaled well with CPU frequency.

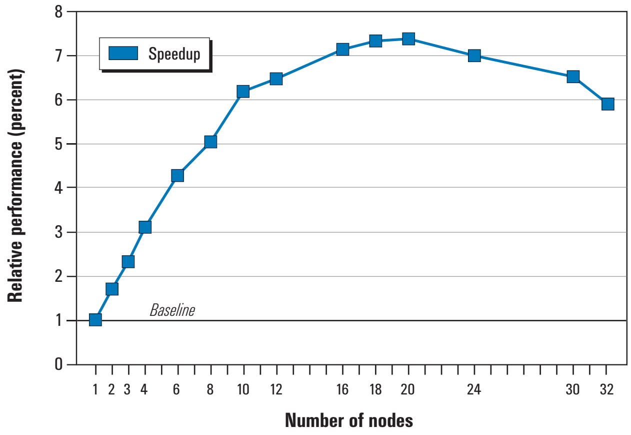


Figure 4. Performance improvement when running NWChem on multiple nodes in a cluster versus a single node (baseline)

In this study, performance gains were realized when L3 cache was increased and when CPU clock frequency was increased. Performance gains were more significant when clock speed was increased and cache size kept constant than when cache size was increased and clock speed kept constant. The benefits from cache size are likely to be more significant when using larger data sets for NWChem than the siosi3.nw input file used for this study.

The rapid advancement of microprocessor technologies and high-speed interconnects has facilitated many successful deployments of HPC clusters.

Measuring the scalability of NWChem on a Dell HPC cluster

The impact of running NWChem on a single node versus multiple nodes in a cluster is illustrated in Figure 4, which shows that in this study, NWChem scaled well on clusters as more nodes were used. Figure 4 shows that execution time was cut dramatically by running NWChem on multiple nodes versus a single node, which helps illustrate why NWChem is usually run on SMP systems and clusters as opposed to a single-processor system or platform. NWChem was developed to provide maximum efficiency and scalability on multiple nodes or processors.

Figure 4 also shows that the speedup from multiple nodes versus a single node was almost linear until 20 nodes were added. Thus, in this study, the NWChem application was not continuously scalable with respect to the amount of computation power available. Beyond 20 nodes, a performance bottleneck occurred, resulting in diminishing returns as the cluster size was increased beyond 20 nodes. This indicates that factors other than computational power play an important role in application performance. The bottleneck in this case could have been caused by I/O traffic, Network File System (NFS) performance, or the network interconnect—any of which can cause performance to diminish as nodes are added to the cluster. To improve performance and enable scaling to a larger number of nodes, administrators must detect and remove the bottleneck (for instance, by using faster interconnects or parallel file systems).

Enhancing NWChem performance using standards-based HPC clusters

Life sciences applications such as NWChem, which are based on quantum theories and electronic functions that calculate molecular and periodic energies and densities, are highly extensible and compute intensive. NWChem has been designed for efficient scaling on parallel systems, and as demonstrated by the study described in this article, can benefit from the performance and computational power provided by HPC clusters as compared to a single-node platform. Thus, an HPC cluster using standards-based Dell PowerEdge servers can provide a highly scalable and cost-effective platform on which to execute NWChem and similar applications.

Munira Hussain is a systems engineer in the Scalable Systems Group at Dell. Her current interests include interconnects, IA-64 and Intel Extended Memory 64 Technology (EM64T), life sciences applications, and Linux and Microsoft® Windows® software solution stacks for high-performance computing. She has a B.S. in Electrical Engineering and a minor in Computer Science from University of Illinois at Urbana-Champaign.

Ramesh Radhakrishnan, Ph.D., is a systems engineer on the Dell High-Performance Computing Cluster team. His areas of interest are computer architecture and performance analysis. Ramesh has a Ph.D. in Computer Engineering from The University of Texas at Austin.

Kalyana Chadalavada is a senior engineer with the Dell Enterprise Solutions Engineering Group at Bangalore Development Center (BDC). His current interests include performance characterization on HPC clusters and processor architectures. Kalyana has a bachelor of technology degree in Computer Science and Engineering from Nagarjuna University in India.

FOR MORE INFORMATION

NWChem:
www.emsl.pnl.gov/docs/nwchem/nwchem.html
www.llnl.gov/asci/platforms/bluegene/papers/28apra.pdf
viper.bii.a-star.edu.sg/nwchem-docs/user/node3.html

TCGMSG:
www.emsl.pnl.gov/docs/parsoft/tcgmsg/tcgmsg.html

MEGABYTE:

What not having a
Linux strategy can take out
of your bottom line.

If you're paying unreasonable licensing fees for software that constantly needs security patches, you're getting eaten alive. But there's a solution. With SUSE® LINUX, Novell® can help you unleash the cost-saving power of a flexible, end-to-end open source strategy. Only Novell supports Linux from desktop to server, across multiple platforms. We'll integrate our industry-leading security, management and collaboration tools seamlessly into your environment. We'll provide award-winning technical support 24/7/365, and train your IT staff to deploy Linux-based solutions. And we'll make sure your open source strategy actually meets your number-one business objective – making money. Call 1-800-215-2600 to put some teeth back into your tech strategy, or visit www.novell.com/linux ➔ **WE SPEAK YOUR LANGUAGE.**

suse
A NOVELL BUSINESS

Novell®