DELL POWER SOLUTIONS · JUNE 2004

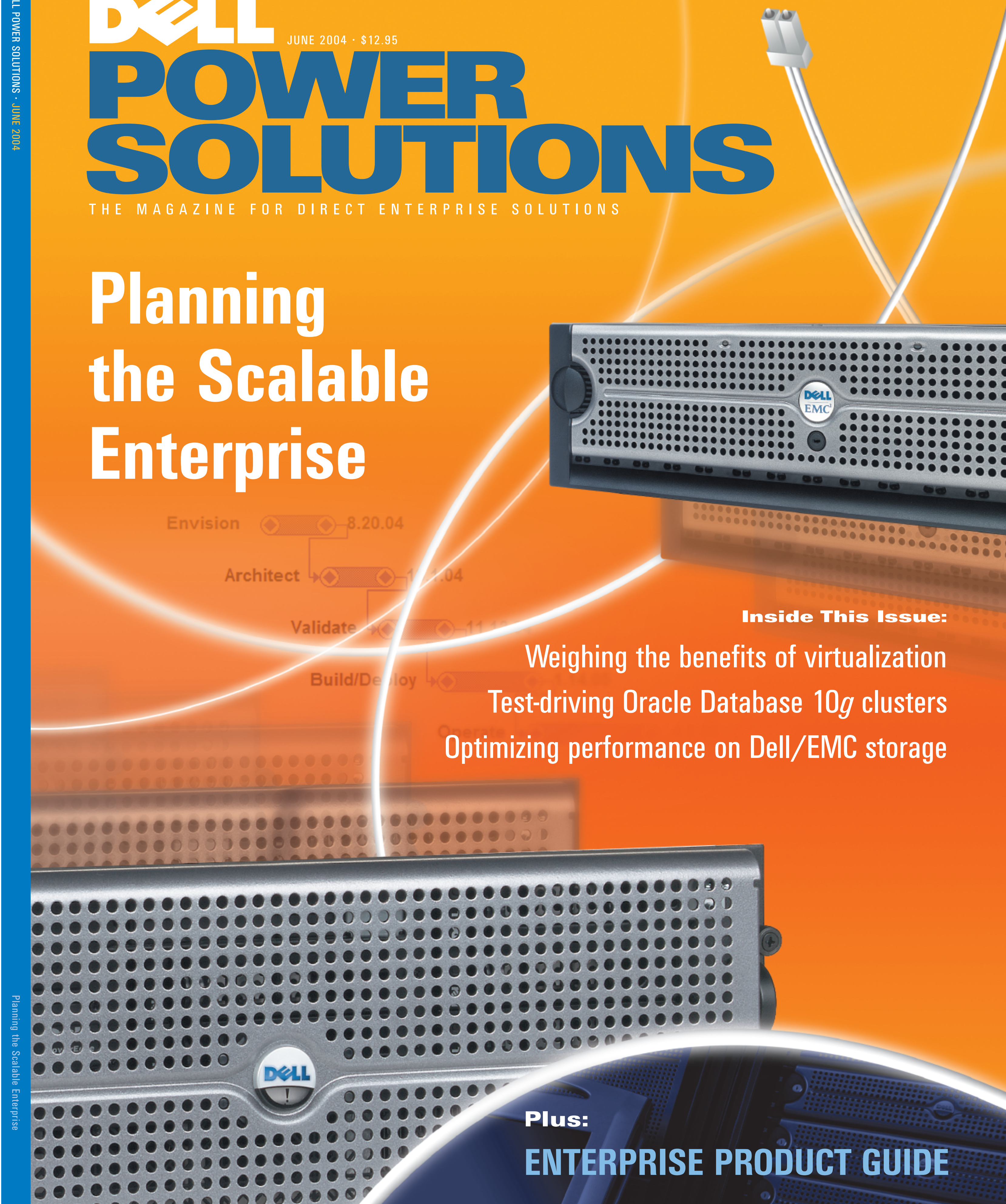
DELL™

JUNE 2004 · $12.95

POWER SOLUTIONS

THE MAGAZINE FOR DIRECT ENTERPRISE SOLUTIONS

Planning the Scalable Enterprise



Inside This Issue:

Weighing the benefits of virtualization

Test-driving Oracle Database 10*g* clusters

Optimizing performance on Dell/EMC storage

Plus:

ENTERPRISE PRODUCT GUIDE

Planning the Scalable Enterprise

Gigabit Ethernet
Multi-Gigabit Ethernet

DELL POWER SOLUTIONS

THE MAGAZINE FOR DIRECT ENTERPRISE SOLUTIONS

JUNE 2004

EDITOR'S COMMENTS

EXECUTIVE INSIGHTS

SCALABLE ENTERPRISE

SYSTEMS MANAGEMENT

SYSTEM ARCHITECTURE

STORAGE

LINUX

HIGH-PERFORMANCE COMPUTING

SECURITY

ADVERTISER INDEX

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, Vicki Van Ausdall
CONTRIBUTING AUTHORS | Tim Abels, Mahmoud Ahmadian, Rizwan Ali, David Ard, Bala Beddhannan, Joel Borellis, Michael E. Brown, John Brunken, Will Claxton, John Coombs, Chris Croteau, Aurelian Dumitru, Clint Espinoza, Diego Esteves, Yung-Chin Fang, Ibrahim Fashho, John Fruehe, Michael L. Gierhart, Richard Golasky, Niroop Gonchikar, Jose Maria Gonzalez, Brian Greene, Sharon Hanson, Jordan Hargrave, Robert Hentosh, John Hull, Dave Jaffe, Ph.D, Peter Lillian, J. Craig Lowery, Ph.D., Paul Luse, Stephan Maahs, Zafar Mahmood, Nathan Martell, Arrian Mehis, Todd Muirhead, Rogelio Noriega, Jason D. Norman, John Page, Shelley Palmer-Fettig, Felipe Payet, Mary Jean Raatz, Paul Rad, Ramesh Radhakrishnan, Ph.D., Ron Rasmussen, Reza Rooholamini, Ph.D., Terry Schroeder, Sudhir Shetty, Scott Stanford, Lisa Stout, Gong Wang, Robert White, Ward Wolfram, Philip Wong, Matthew Wygant

ART

ART DIRECTOR | Mark Mastroianni
DESIGNERS | Phu Tran, Cynthia Webb
ILLUSTRATOR | Cynthia Webb
COVER DESIGN | Phu Tran
CONTRIBUTING PHOTOGRAPHER | Lee Kirgan

ONLINE

WEB PRODUCTION | Brad Klenzendorf

SUBSCRIPTION SERVICES

Subscriptions are free to qualified readers who complete the online subscription form or the subscription reply card in each 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 card by surface mail, or fax the subscription reply card to +1 512.283.0363. For subscription services, please e-mail **us_power_solutions@dell.com**.

ABOUT DELL

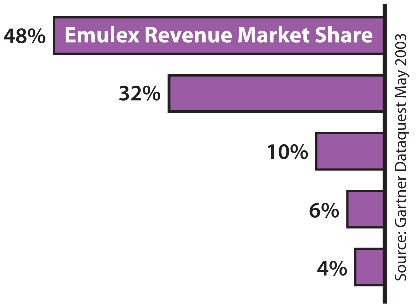
Dell Inc., headquartered in Round Rock, Texas, near Austin, is the world's leading direct computer systems company. Dell is one of the fastest growing among all major computer systems companies worldwide, with approximately 47,800 employees around the globe. Dell uses the direct business model to sell 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.

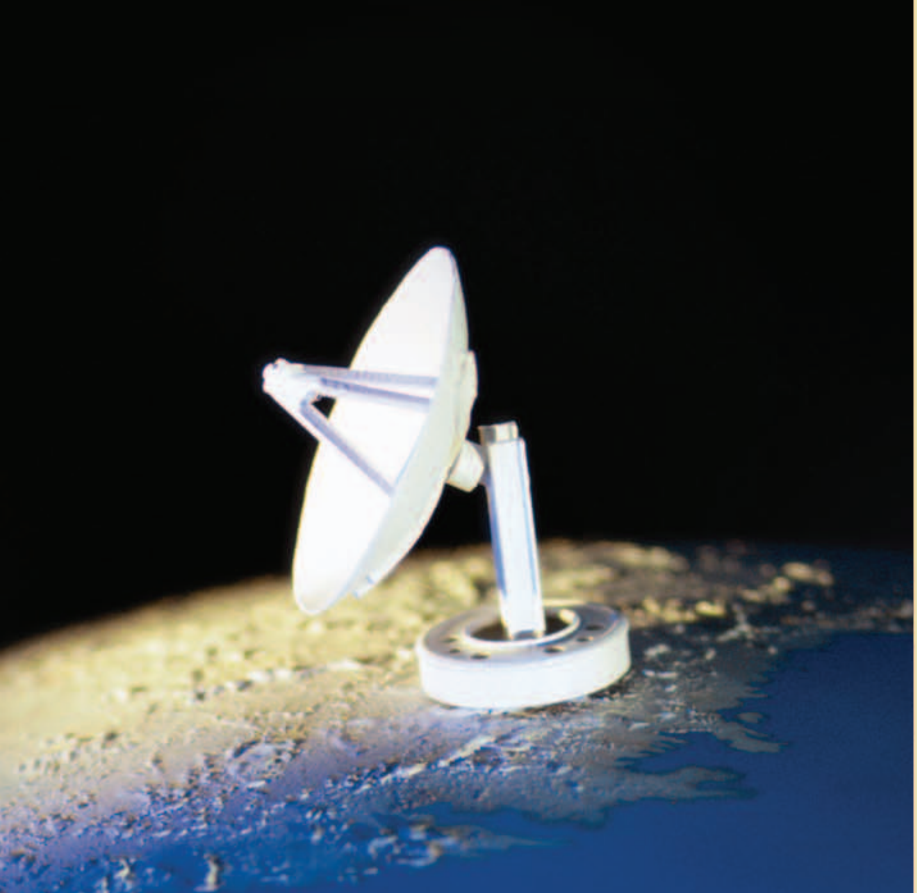
ON THE FRONT COVER | The Dell/EMC CX500 Fibre Channel storage array (top right) and Dell PowerEdge 6650 server (bottom left). Design by Phu Tran.

June 2004

Emulex.
We network storage


48%
Emulex Revenue Market Share
32%
10%
6%
4%
Source: Gartner Dataquest May 2003

Who says **SIZE** doesn’t matter?



The Critical Path to Success

Henry Laurence Gantt died in 1919, just a few short years after he conceptualized the project planning tool that would come to bear his name—the Gantt chart. Unfortunately, Gantt never lived to see his humble critical path method successfully used for planning such massive construction projects as the Hoover Dam and the U.S. Interstate Highway System.

Ratcheting years forward, the Gantt chart has assumed its rightful place as the preeminent planning tool for many IT professionals. At times, the complex web of tasks involved in building the enterprise IT infrastructure may seem as daunting as that faced by the dam and road architects of the past. But help is on the way.

In this issue of *Dell Power Solutions*, the Gantt planning tool remains top of mind as we provide essential technical guidance aimed at helping administrators map out their scalable enterprise strategy. For a quick level-set on the Dell vision of the scalable enterprise turn to page 8, where you'll find an executive overview of three key planning horizons:

- **Simplified operations** enabled by increased standardization and automation
- **Improved resource utilization** enabled by the consolidation and virtualization of server and storage platforms
- **Cost-effective scalability** enabled by the capability to add modular computing and storage capacity as needed

We understand your quest for in-depth technical information before setting plans in motion. To that end, the Scalable Enterprise section of this issue highlights articles on planning for virtualization as an enterprise computing infrastructure, test-driving the new Oracle® Database 10*g* with Real Application Clusters (RAC) technology, and scaling Microsoft® SQL Server™ 2000 software across multiple four-processor servers in real-world environments. In the Storage section, scope out the first of a two-part series on designing and optimizing Dell/EMC storage area network (SAN) configurations.

Still need more? Check out our Systems Management section for the latest on automation using Dell™ Update Packages for the Red Hat® Enterprise Linux® operating system and integrating Windows® Management Instrumentation (WMI) scripting using the Dell OpenManage™ tool set. And, in the launch of our new System Architecture section, read about planning considerations for Intel® Extended Memory 64 Technology (EM64T) on servers and workstations.

As another aid for your planning endeavors, this issue features the Dell Enterprise Product Guide—a handy pull-out reference for Dell PowerEdge™ servers, Dell PowerEdge–Oracle database and application solutions, Dell/EMC and Dell PowerVault™ storage, and Dell PowerConnect™ networking products. Among the listings are the new Dell/EMC AX100 entry-level storage array and the Dell PowerVault 745N network attached storage (NAS) server, which were announced after our March 2004 storage issue went to press.

So be brave. Go ahead and launch your project management software, lay in a few key tasks and milestones, create that Gantt chart—and move one step closer to enabling your own scalable enterprise.

Happy planning!

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

MICROSOFT
2004
IBM
ORACLE
2004
EMC
2004

Planning for the Scalable Enterprise

Paul Gottsegen, vice president of enterprise marketing at Dell Inc., explains how the company's strategy of the scalable enterprise is centered around standardization of core data center elements—to help organizations meet today's business challenges and be prepared for tomorrow.

Organizations have derived substantial benefits from the standardization of servers, storage, operating environments, and applications over the last several years. In particular, price/performance advantages of Intel® architecture–based servers versus proprietary legacy systems[1] have helped enable enterprises to satisfy a growing number of IT processing needs using cost-effective, industry-standard x86-based server architectures and related storage hardware. Based on a robust platform of modular computing components, Dell's enterprise strategy focuses on standardizing core elements of the data center to deliver superior value.

However, along with the price/performance advantages of standards-based systems have come some challenges. Many organizations now must contend with a proliferation of x86-based servers that were deployed with a "one application, one server" mindset. Meanwhile, administrators must address skyrocketing data and related storage requirements while grappling with IT environments that are hard to manage and resources that are underutilized. Challenged to do more with less, IT organizations still must plan for the future and find a way to unleash the underlying horsepower and interoperability that business-critical enterprise applications demand.

What are the key elements of Dell's enterprise strategy?

Our enterprise strategy is, in effect, the scalable enterprise. We are working toward standardization of core elements of the data center to provide superior value. The Dell vision of the scalable enterprise comprises three key components: simplified operations through increased standardization and automation; improved utilization enabled by the consolidation and virtualization of server and storage platforms; and finally, the capability to scale incrementally by adding computing and storage capacity as business needs arise. A significant amount of the core building blocks to realize the scalable enterprise are proven and available today, and we are working to deliver additional elements in the very near future.

What initiatives has Dell taken to help simplify IT operations?

Dell is working with leading systems management software providers and key standards bodies to streamline and automate the provisioning, monitoring, and change management requirements for deploying and operating Intel processor–based servers. Dell believes that organizations should be able to simplify their operating environments and reduce the number of tools required to manage those environments through standardization.

To that end, Dell has led standardization efforts to consolidate systems management capabilities through the integration of Dell™ OpenManage™ software with systems management software from leading providers, including Microsoft® Systems Management Server (SMS) and Altiris® Server Management Suite™ software. Beyond that, the company is working on a Dell OpenManage software development kit (SDK) that will allow other third-party vendors to enable their systems management software to manage Dell servers as well as operating systems and applications. We view this approach as delivering the power of many best-of-breed technologies with the ease of one management tool.

Automating the deployment of servers and storage can help allow organizations to reduce the number of resources required to manage their IT infrastructure, and lay groundwork that enables

[1] For more information about the price/performance benefits of enterprise applications on standards-based systems versus proprietary legacy systems, see "Comparing Oracle9*i* Database Performance on Dell and Sun Servers" by Dave Jaffe, Ph.D., and Todd Muirhead in *Dell Power Solutions*, August 2003.

enterprises to provision and deploy capacity automatically based on predefined usage scenarios. Dell is actively working on this type of dynamic data center model with its alliance partners.[2]

How can enterprises improve resource utilization?

Many of today's Intel x86-based servers are underutilized because they were deployed to run a single application. In fact, average utilization rates of x86-based servers can be as low as 15 percent, according to a recent IDC report.[3] Storage resources are often underutilized as well—particularly when they are attached directly to individual servers. Direct attach storage can lead to an imbalance in storage capacity across the enterprise that is difficult to manage, resulting in some servers being overprovisioned while others are underprovisioned.

Two complementary approaches can help organizations improve resource utilization: consolidation and virtualization. Consolidating disparate server and storage systems into a more centrally managed facility—and aggregating compatible applications onto more powerful servers and centralized storage—can enable organizations to achieve considerable improvements in processor utilization, systems management, and storage capacity.

In addition, the virtualization of servers and storage can help organizations increase processor utilization by running different applications in isolated partitions—or *virtual machines*—on the same physical server, and by provisioning storage capacity across the enterprise storage area network (SAN) as needed. Applications such as VMware® ESX Server™ and VirtualCenter software, and VMware VMotion™ technology, are designed to help organizations move their smaller legacy applications into a consolidated environment and provide IT administrators with the capability to provision server cycles as needed and balance workloads across the enterprise.

This virtualization technology is rapidly approaching enterprise strength. In Dell's engineering labs, recent testing[4] with the latest releases of VMware ESX Server and VMotion virtual machine migration technology demonstrated how virtual machines can be easily relocated from one physical server to another while running heavy production loads—an important prerequisite for the dynamic data center of the future.

How can Dell help organizations scale cost-effectively?

To be competitive, organizations must be able to add modular computing and storage capacity easily and flexibly when needed—and not be required to invest ahead of their needs, as can be the case when implementing larger, legacy midrange and mainframe systems. The Dell commitment to industry-standard servers and storage arrays is designed to help organizations pay as they grow, adding only the incremental capacity they require.

This modular, building-block approach is particularly attractive to enterprises that want to preserve capital for other business needs. It also helps provide excellent scalability because unplanned needs can be fulfilled quickly using powerful, industry-standard server and storage components.

Organizations migrating from more expensive, proprietary systems to cost-effective industry-standard servers and storage can expect substantial benefits in both price and performance. For example, deploying clusters of small, industry-standard servers is a key element of the scalable enterprise strategy, and can help provide redundancy for high availability, reduced hardware costs, and the option to increase processing power simply by inserting another server into the cluster. Dell testing[5] indicates that scaling out with a pair of four-processor industry-standard servers in this manner can produce a formidable 42 percent performance gain over a single eight-processor server.

To be competitive, organizations must be able to add modular computing and storage capacity easily and flexibly when needed.

What paths to the enterprise of the future does Dell envision?

Several paths lead to the enterprise of the future, where servers and storage are dynamically provisioned, or even self-provisioned. Dell's path is one of open standards and integration—enabling enterprises to exercise choice and flexibility when designing their IT infrastructure. Dell recognizes the importance of time to market and the cost of today's IT infrastructure, and realizes that enterprises must pave the way to the dynamic data center of the future in practical phases.

By developing an enterprise strategy based on standardizing core elements of the data center, Dell can enable organizations to build a dynamic IT infrastructure that is designed to provide superior value and scalability well into the future.

[2] For more information about the dynamic data center of the future, see "Optimizing the Virtual Data Center" by J. Craig Lowery, Ph.D., in *Dell Power Solutions*, November 2003; and "Virtualization as an Enterprise Computing Infrastructure" by J. Craig Lowery, Ph.D., in *Dell Power Solutions*, June 2004.

[3] Source: "The Role of Grid Computing in the Coming Innovation Wave" by John Humphreys and Mark Melenovsky, IDC, March 2004.

[4] For more information about VMware virtualization software performance on Dell PowerEdge servers, 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.

[5] Performance gain based on tests of orders performed per minute on two Dell PowerEdge 6650 servers. For more information about scaling out on four-processor industry-standard Dell servers, see "Scalable Enterprise Computing: Testing a Clustered Database on the Dell PowerEdge 6650" by Dave Jaffe, Ph.D., and Todd Muirhead in *Dell Power Solutions*, March 2004.

Virtualization as an Enterprise Computing Infrastructure

As the IT industry continues to refine the notion of using commercial off-the-shelf parts to create enterprise-class data processing systems, the problem of transcending the tight coupling of software stacks to hardware platforms continues to be an obstacle. Hardware virtualization tools, such as VMware® ESX Server™ software and Microsoft® Virtual Server software, represent an evolving genre of products that can dissociate software from hardware, providing several impressive systems management benefits.

BY J. CRAIG LOWERY, PH.D.

In recent years, open standards–based hardware and software products have proliferated. Many enterprise IT administrators have embraced this trend by migrating to standards-based systems in their data centers. However, issues of manageability still arise. More and more, IT administrators are turning to hardware virtualization tools, such as VMware® ESX Server™ software and Microsoft® Virtual Server software, to help manage these standards-based platforms.

Understanding the problem with scaling out

Sheer computing power is inexpensive. In an enterprise data center where demand for computing resources continues to rise, a natural tendency has been to acquire more systems to meet the needs of users. At first, this proliferation was thought to be an economical solution, but it soon became apparent that scalability was not simply a matter of finding rack space and power for new servers.

On the surface, using several inexpensive, standards-based components—such as one- and two-processor Intel® architecture–based servers with 1U form factors—can be an appealing option for the typical enterprise data center. As the need for more processing power arises, administrators increase capacity by adding more of these small components to the aggregate system. Because the components are standards-based, they are largely interchangeable and even disposable; if a system malfunctions, it can be easily replaced. This method, commonly referred to as *scaling out,* has merit but is not without caveats.

Manageability has become the chief obstacle for the commercial off-the-shelf (COTS) approach to enterprise computing. Certainly, many standard systems can be colocated in a data center, loaded with software, and placed into service. However, little time passes before administrators must apply operating system (OS) patches, repurpose a server, perform hardware maintenance, or adjust hardware allocation because of changes in demand. These challenges may foil the scale-out proposition for early adopters, even though great improvements are being made in software change-management tools. In fact, management issues in a scale-out environment can be so problematic that many IT departments have returned to consolidation of hardware resources. Instead of scaling out, they choose to *scale up*—that is, replace their many small, inexpensive systems with a few large, proprietary, and very costly systems in an effort to make the data center easier to manage.

Unfortunately, those who have consolidated into a scale-up configuration are treating a symptom, not the root

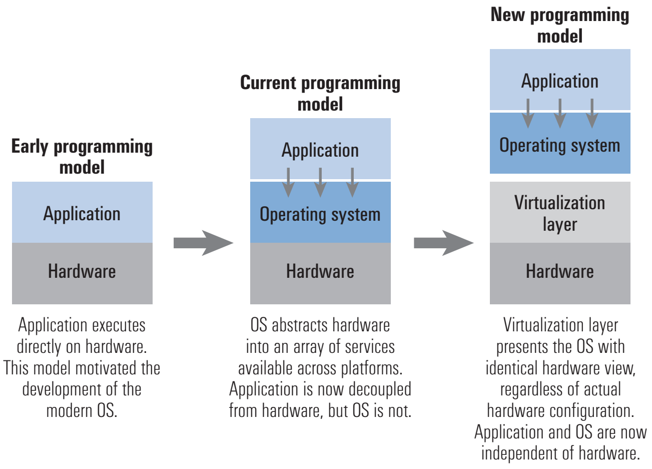

Early programming model
Application
Hardware
Application executes directly on hardware. This model motivated the development of the modern OS.
Current programming model
Application
Operating system
Hardware
OS abstracts hardware into an array of services available across platforms. Application is now decoupled from hardware, but OS is not.
New programming model
Application
Operating system
Virtualization layer
Hardware
Virtualization layer presents the OS with identical hardware view, regardless of actual hardware configuration. Application and OS are now independent of hardware.

Figure 1. Evolving programming models

cause, of management issues in a COTS-based data center. The root cause is not the proliferation of the servers themselves, but the tight coupling of software to hardware. Each management issue mentioned previously—software patching, hardware repurposing, maintenance downtime, and workload balancing—is difficult to solve even with the most advanced systems management tools, largely because software is closely bound to hardware.

In spite of the considerable achievements in hardware standards, significant differences in server systems still exist—usually caused by newer technologies that are still maturing through the standardization curve. Differences in components such as network interfaces, disk subsystems, processor families, and storage configurations can result in an OS configuration that is tailored to the hardware configuration. If an administrator removes the disks from one system and inserts them into another system, chances are the second system will not even start, much less function stably, unless its configuration is identical to that of the first system.

Decoupling the hardware/software stack

Consider that if all server hardware were identical, then a software stack—an OS and its hosted applications—could quickly move from system to system without alteration. In such a world, software and hardware are *decoupled,* and the hardware is truly interchangeable and disposable. Servers can be repurposed within minutes by simply copying, or *cloning,* an archived, pre-patched software stack. Decoupling also enables IT administrators to balance workloads more effectively by reassigning hardware to software dynamically.

However, standard hardware does not imply *identical* hardware. Standards help to manage the technology differences between systems in such a way that their benefits can be easily accessed by all parties. For computer components, standards often take the form of an interface specification. For example, TCP/IP—a standard for network communication—is an interface for accessing a network that takes on many different physical forms. Yet the format and function of TCP/IP datagrams and the protocol are the same, regardless of the physical network.

Virtualization applies the same notion of abstraction to computer hardware by creating a new interface, the *virtual machine,* which becomes the standard interface for software stacks to execute on standardized, but different, underlying physical hardware. This is not a new idea; however, it is being applied at a different point in the combined hardware and software stack (see Figure 1).

Early computers ran application software directly on the hardware—a tedious programming model that motivated the development of the modern OS. Operating systems shield applications from the underlying hardware by abstracting the hardware as an array of managed resources and services. Although applications have more freedom from hardware in this current model, they still depend on the OS for a file system that holds the application's persistent state—that is, the application's configuration and data. The new hardware-virtualization programming model bundles applications and operating systems together (including the file system), allowing them to move as a unit across various physical platforms.

Understanding the basics of virtualization

A virtual machine (VM) is a software-created construct that is functionally equivalent to physical hardware, at least from the perspective of its ability to execute software. A VM has the same features as a physical system: expansion slots, network interfaces, disk drives, and even a BIOS. Like physical systems, VMs can be powered up and powered down. When the VM is powered up, an OS runs on the virtual disk. Also called the VM's *image*, *stack*, or *container*, the virtual disk is actually a large file stored on the underlying physical system. Administrators can transfer a VM from one physical system to another by moving its image, or by storing its image in a shared location accessible by both physical systems.

Figure 2 depicts two options for implementing VMs. Both approaches divide the hardware/software stack into three layers: application software layer, virtualization layer, and physical

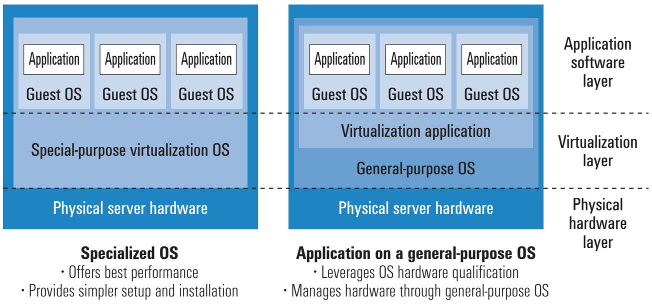

Application
Guest OS
Application
Guest OS
Application
Guest OS
Special-purpose virtualization OS
Physical server hardware
Specialized OS
• Offers best performance
• Provides simpler setup and installation
Application
Guest OS
Application
Guest OS
Application
Guest OS
Virtualization application
General-purpose OS
Physical server hardware
Application on a general-purpose OS
• Leverages OS hardware qualification
• Manages hardware through general-purpose OS
Application software layer
Virtualization layer
Physical hardware layer

Figure 2. Two methods for implementing hardware virtualization

hardware layer. The two implementations differ only in how they implement the middle virtualization layer.

Specialized OS. The virtualization layer is a special-purpose OS that is installed directly onto the physical hardware, as one would install other operating systems. This OS can manage only the resources of the physical system, to share these resources among the VMs that it hosts. It does not, for example, have its own general-purpose file system or support user processes. Because no generic hosting OS exists, this approach offers the better performance of the two alternatives and has fewer setup and installation steps. Because it is not a mainstream OS, however, driver availability may be limited for some physical components. Products such as VMware ESX Server use this approach.

Application on a general-purpose OS. The virtualization layer is the combination of a general-purpose OS, such as a Microsoft Windows® or Linux® OS, and a virtualization application. Because it leverages a preexisting, full-service OS, this approach offers broader compatibility with hardware and can take advantage of existing management infrastructures already created for these environments. However, because the virtualization layer depends on an intervening OS, performance can be noticeably degraded. Products such as Microsoft Virtual Server and VMware GSX Server™ software use this approach.

Realizing the benefits of virtualization

Hardware virtualization can offer enterprise computing environments benefits such as consolidation, normalization, isolation, replication, and relocation.

Consolidation. This practice involves converting many physical servers to VMs and hosting the VMs on a reduced number of physical systems. Consolidation saves space, but that is only one of many areas in which enterprises can benefit significantly from hardware virtualization.

Normalization. This ability shields software from hardware peculiarities and change. For example, consider a driver update in a heterogeneous OS environment including Windows, Linux, and Novell® NetWare® platforms. Without virtualization, three or more driver versions must be identified and installed. However, with virtualization, only one driver must be found and installed at the virtualization layer—guest operating systems and applications continue to use generic drivers with their virtualized generic hardware. Legacy operating systems that are no longer supported by the manufacturer can execute on newer physical hardware, which enables delayed migration of legacy systems.

Isolation. Some applications do not behave well in a shared, multi-programmed environment, either because of conflicting configuration requirements, resource domination, faulty code that can crash the entire system, or other side effects. Virtualization can be used to partition applications into their own private OS containers on a single physical system so that resource allocations can be enforced at the virtualization layer. If an application crashes its own VM, it does not affect other VMs.

Replication. Quick deployment of software stacks becomes a trivial task using a virtualization approach. First, administrators create reference software stacks for various applications—Web servers, mail servers, or file servers, for example—and then copy these images to a read-only archive, or a *gold master library*. Later, when a particular type of server is needed, administrators can clone the image from the archive and place it on a system that has spare capacity, usually within a matter of seconds.

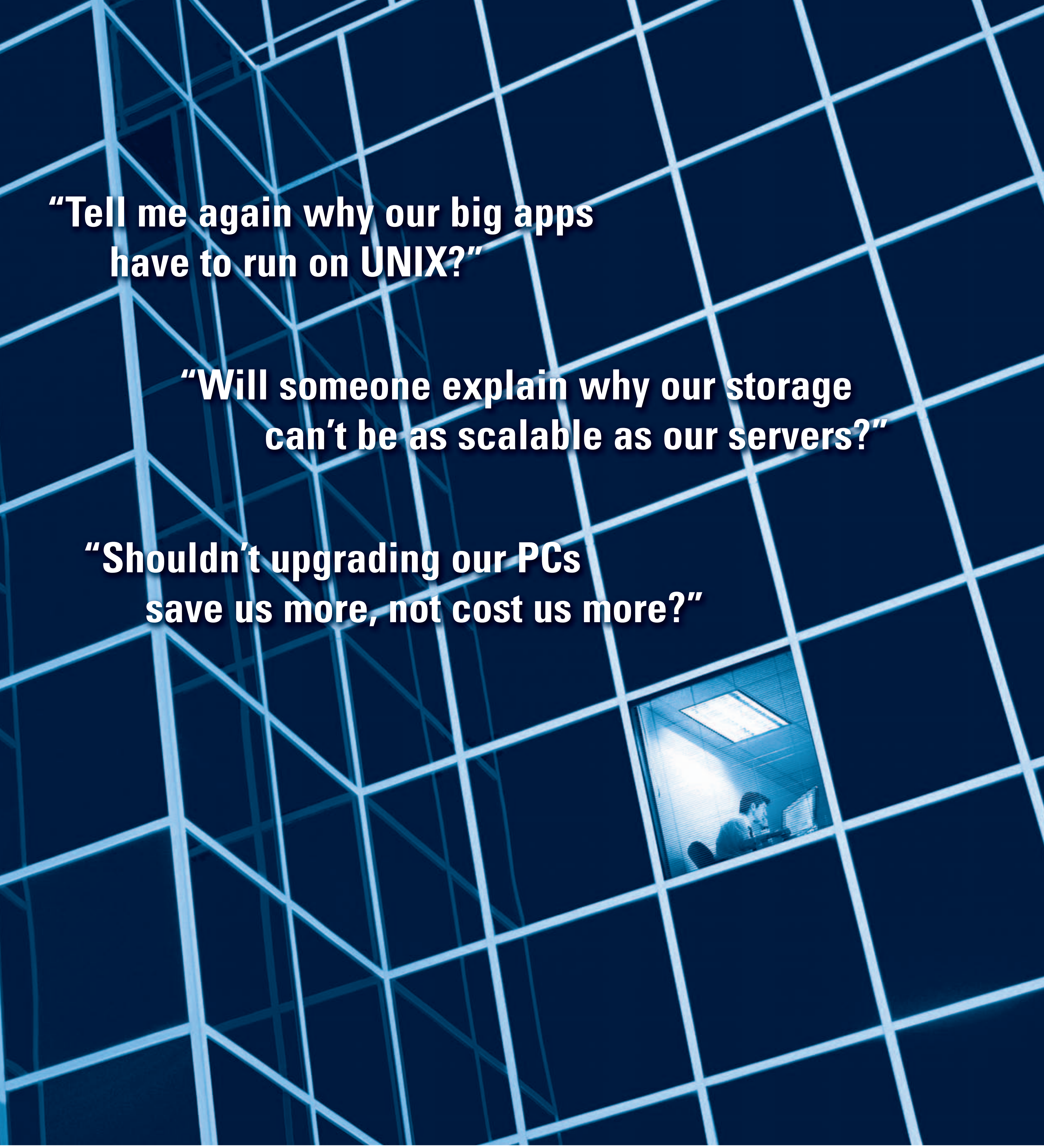
Relocation. For the same reasons that virtualization facilitates replication, the decoupling of hardware and software enables the software stack to move freely between available virtualized platforms. At the very least, administrators can shut down a VM, move its image to another physical system, and then bring the VM back up. More sophisticated mechanisms allow for executing VMs to move in real time, without shutting them down and without affecting service to users. This mobility offers the greatest potential benefit to the scalable enterprise.

Moving mindsets

Hardware virtualization is not new to computing. However, it is a recent development in low-cost Intel architecture–based systems, for which it holds great promise for mitigating the management problems associated with scale-out architectures. The main challenges that virtualization technology poses are, surprisingly, not of a technical nature. Virtualization of Intel processor–based systems is best known as a tool for space consolidation, especially in test and development environments—not in production environments. Consequently, many IT administrators familiar with virtualization technology may have preconceived notions about its applicability in a production environment—and be unaware of the numerous benefits beyond simple consolidation.

Although it does not solve every systems management problem, virtualization is ready to move into production in select large-scale enterprise computing environments. Prime candidates include server farms hosting a diverse mix of lightly to moderately loaded services. Many early adopters of virtualization technology in enterprise environments are reporting great success. While traditional systems management tools are evolving to make it easier to manage non-virtualized platforms, they are also incorporating virtualized platforms into the aggregate data center view. Virtualization has matured to the point where it provides the required stability, and the benefits are extremely compelling. Although some may try to minimize the scope of this impact, most would agree that hardware virtualization has a definite role to fulfill in the scalable enterprise.

J. Craig Lowery, Ph.D. (craig_lowery@dell.com) is a senior engineering development manager in the Dell Product Group—Enterprise Solutions Engineering. 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.



“Tell me again why our big apps have to run on UNIX?”

“Will someone explain why our storage can’t be as scalable as our servers?”

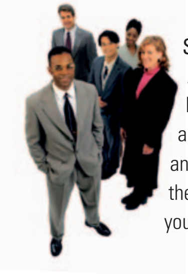
“Shouldn’t upgrading our PCs save us more, not cost us more?”

YOU KNOW BETTER.

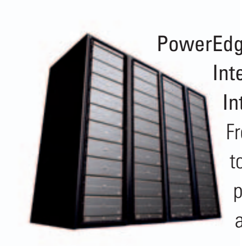
DELL KNO

Build a scalable enterprise with Dell. Get real

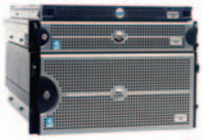
How? Dell technology. Intel® processor-based. Easy-to-manage. Pay



Services
Across the entire IT lifecycle, from planning and design to implementation and asset disposal, Dell offers the services to help reduce your ongoing support costs.



PowerEdge™ Servers with Intel® Xeon™ Processors and Intel® Itanium® 2 Processors
From e-mail applications to CRM demands, Dell servers provide great performance and superior value.



Storage Solutions
Enhance datacenter management by increasing storage productivity as you scale capacity, performance and availability.

intel inside XEON™

WS HOW.

savings today and a better path to the future.

as you grow with award-winning solutions from desktop to datacenter.

Latitude™ Notebooks with Intel® Centrino™ Mobile Technology
Help your people do more in more places than ever.

OptiPlex™ Desktops with Intel® Pentium® 4 Processors and HT Technology
Designed to lower ownership costs while improving productivity and security.

Printers
Dell Laser Printers enhance your workgroup printing capabilities with high-speed and high-volume at a low cost.

Real solutions for real challenges. Easy as DELL™

Contact a Dell Professional for technology built around you.

1-866-239-DELL • www.dell.com/dellknows10
toll free

Managing Energy Resources

in the Virtual Data Center

Virtual data centers (VDCs) hide all hardware and software components behind an abstraction layer, presenting the computing environment as a collection of managed services. This article describes how the VDC differs from traditional data centers with regard to energy management and surveys the environmental benefits of the emerging VDC model.

BY ROBERT WHITE AND TIM ABELS

Traditional data centers rarely manage energy centrally or as a first-class resource. However, the continued increase in data center hosting capacity leads to increased energy consumption—resulting not only in higher economic costs but also in higher environmental costs, such as increased emissions of carbon dioxide (CO_2) and other greenhouse gases. Consequently, the monitoring and conservation of energy has emerged as a challenge to data center design, including goals of reducing power and cooling costs, reducing energy needs, and better managing current energy consumption.

Scaling out data center resources using dense, inexpensive, off-the-shelf components is an appealing approach for meeting large-scale, mixed enterprise-computing requirements. But these largely interchangeable and even disposable servers, storage, and switch resources are often undermanaged, which can generate cost and environmental challenges in a traditional data center.

The virtual data center (VDC)[1] model can be applied to manage energy as a first-class resource and to guarantee energy at predictable costs. The VDC virtualizes all hardware and software resources, presenting the data center as a service view of a single computer. Many recent variations of the VDC model are available, such as utility computing and grid computing,[2] and each can satisfy quality of service (QoS) at guaranteed resource costs.

A fundamental goal of energy resource management across the data center is to minimize environmental impact while maintaining QoS for all users. Significantly reducing energy consumption can yield environmental as well as corresponding cost benefits. The real-time, flexible control that a VDC

[1] For more information, see "Building the Virtual Data Center" by J. Craig Lowery, Ph.D., in *Dell Power Solutions*, February 2003.

[2] For more details and early standardization definitions, visit http://www.dmtf.org and http://www.ggf.org.

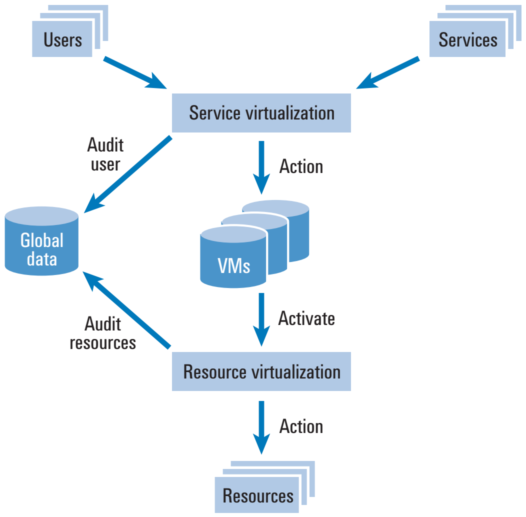


Figure 1. Architecture of a static VDC

can provide may enable the VDC approach to significantly exceed the cost and energy reductions that administrators may achieve from consolidation within traditional data centers. This article[3] describes how the VDC extends the traditional data center model to centrally enable real-time, dynamic control of energy resources.

Applications as virtual machines in a VDC

A VDC implements applications as self-contained virtual machines (VMs), which interface two abstractions (see Figure 1): virtualized resources, including servers, storage, and switches; and virtualized services, including users and user-like services. Without a complete feedback loop for policy-based automation of VMs, the VDC can be thought of as a static control system.

Physically, a VM consists of two files:

- **Software image:** This file contains the application and its operating system (OS).
- **XML-based configuration:** This file addresses hardware, management, users, and services. The XML-based configuration file allows a VM to be self-contained, resulting in a high degree of flexibility because VM software is not tightly coupled to any aspect of the system—it can be paused, copied, backed up, scaled, and so forth without affecting the rest of the system.

VMs are not bound to specific resource instances, so resources can be provisioned and resized dynamically to applications as required, and applications may migrate frequently from one set of resources to another. The resource virtualization layer is the runtime environment for instances from the VM pool and the resource pool.

Similarly, VMs are not bound to specific user and service instances. The service virtualization layer is the runtime environment for instances from the VM pool and the pool of users and services, where services are automated users.

Alternatives to VMs

Figure 2 shows several alternatives to VMs. Nonpartitioned alternatives, such as workload management, lack the isolation needed to ensure that a misbehaving application does not waste resources or crash the OS and thus all its hosted applications.

Partitioning at other layers presents application limitations. For example, software partitioning, in which applications run in separate workspaces on a single OS, is too transient for an isolating abstraction layer. Software partitioning is currently impractical for general-purpose computing because it must be revised for each application programming interface (API) change in the host OS. In general, hardware partitioning is more expensive than software partitioning and lacks the dynamic management of VMs.

The virtual data center model can be applied to manage energy as a first-class resource and to guarantee energy at predictable costs.

Dynamic VDC for automated system monitoring and adjustments

Without a feedback loop, a static VDC is as limited as a temperature system with no central thermostat for continuous and

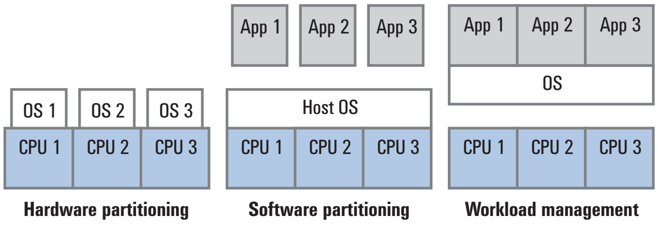


Figure 2. Partitioned and nonpartitioned alternatives to VMs

[3] Copyright ©2004 IEEE. This article, printed with permission from the IEEE organization, summarizes the paper "Optimizing Energy Resource Management in the Virtual Data Center" presented by Robert White and Tim Abels at the 2004 IEEE International Symposium on Electronics and the Environment. To purchase a copy of the complete paper from IEEE, visit http://www.iseesummit.org.

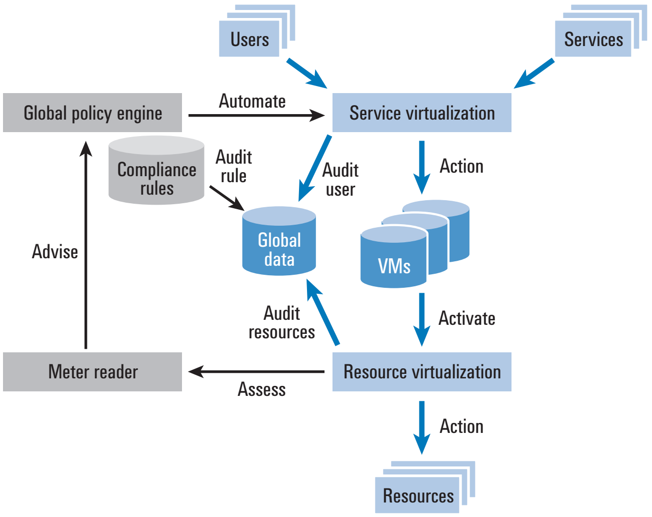


Figure 3. Architecture of a dynamic VDC

adaptive automation. A centralized, dynamic VDC extends a static VDC by incorporating a complete feedback loop that enables policy-based automation (see Figure 3). In a dynamic VDC, a meter reader monitors the virtualization runtime instead of the physical resources, because many false assumptions can occur when monitoring the hardware directly. For example, a resource may not exist physically or globally—or long enough for the system to discover or manage it. The system is updated with new users and services, which initiate additional monitoring loops in the dynamic control system. These dynamic extensions (indicated by the black feedback loop in Figure 3) are the only globally scoped management system and can supplement multiple static VDCs, as shown in Figure 4.

VMs are not bound to specific resource instances, so resources can be provisioned and resized dynamically to applications as required and applications may migrate frequently from one set of resources to another.

Environmental and cost benefits of VDCs

A static VDC can provide environmental and cost reductions at multiple levels: across the entire data center, across resources, and within resources (see Figure 5). Each energy-reducing feature provided by a static VDC has a corresponding dynamic variant, which leverages the control system monitoring and automatic adjustments of the dynamic VDC.

Enabling environmental control systems

Adding environmental data to a dynamic VDC can help to guarantee power, cooling, and waste policies. Such data includes hardware attributes such as physical location, environmental impact, current containment, and enumeration.

To dynamically comply with changing environmental policies, a VDC requires several conditions, including:

- **Global control system:** This includes a thermostat and electrical monitoring with automated correction and assumes a trusted, central authority—in contrast to multiplexing in a grid environment.
- **Global audit:** This archives static configuration updates of resource usage, user actions, and policies—for example, logging the time and initiator of a rule change that exceeded the resource operating redline limit of 25°C.
- **No backdoors:** Unlike traditional data centers, a VDC operates using virtualization layers. Each virtualization layer in a VDC provides full separation of resources using a single, authoritative writer—thereby avoiding contention among multiple writers.
- **Homogeneous resources:** These allow cost and power consumption to scale proportionally to service throughput.
- **Containment:** This involves dynamically migrating VMs to enable load-balanced cooling or power reduction at each level of granularity (from processor to resource to rack).

Toward an environmentally sound data center

Thirty years ago, resource virtualization drastically changed single-node operating systems. Data center administrators must now prepare for similar fundamental changes across their operations.

The VDC can help administrators significantly reduce many of the environmental challenges they face in the traditional data center.

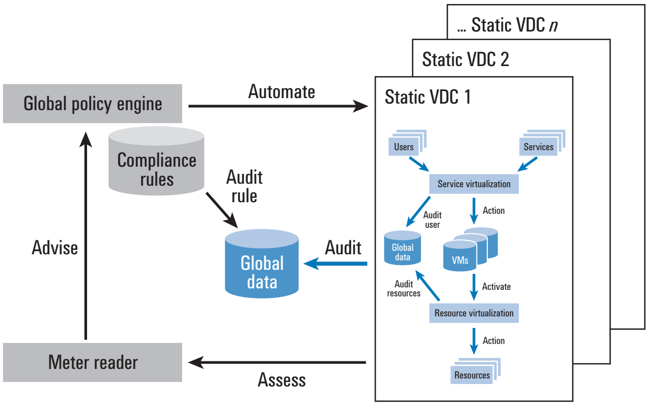


Figure 4. Dynamic extensions to multiple static VDCs

	Static energy reduction	Dynamic control extension*
Across entire data center	• Density: Enables more servers in fewer facilities and consumes less floor space, cabling, and racks than a conventional data center	• Resource repurposing across the data center: Removes most physical boundaries
	• Uniformity: Allows fewer resource configurations,** with more instances on fewer support and management systems than a conventional data center	• Application prioritization: Reassigns resources from lower-priority applications
	• Audit: Provides static data model for complete archive	• Dynamic mapping: Stores information for resources that are temporary, non-unique, and non-global
Across resources	• Consolidation: Allows more VM runtime instances on fewer resources than a conventional data center	• Dynamic consolidation: Matches stateful, running VMs within and across resources
	• Isolation: Provides more VM and life-cycle categories on fewer resources (development, test, staging, and production) than a conventional data center***	• Dynamic VM promotion: Dynamically reassigns VMs in various stages of the software life cycle, from development through test, staging, and production
	• Utilization: Pools excess capacity for higher utilization per resource and fewer hot spares than a conventional data center	• Resource prioritization: Prioritizes queues of resources by service level
	• Autonomic healing: Extends resource life cycles proactively with work-arounds and scheduled component replacements	• High availability: Restarts failed resources or refreshes aging resources with no downtime after offloading VMs
	• Scaling up: Results in more VMs per resource (64-bit, multicore) than a conventional data center	• Scaling out: Results in more VMs dynamically allocated across all resources
Within resources	• Resource power: Controls power-up, power-down, and Wake on LAN	• Load balancing, power-up, and power-down group nodes and group support systems: Detects cyclic power-up and power-down opportunities
	• Low power states of resources: Controls power management technologies such as Advanced Configuration and Power Interface (ACPI)	• Graceful power degradation for groups and prioritized group members

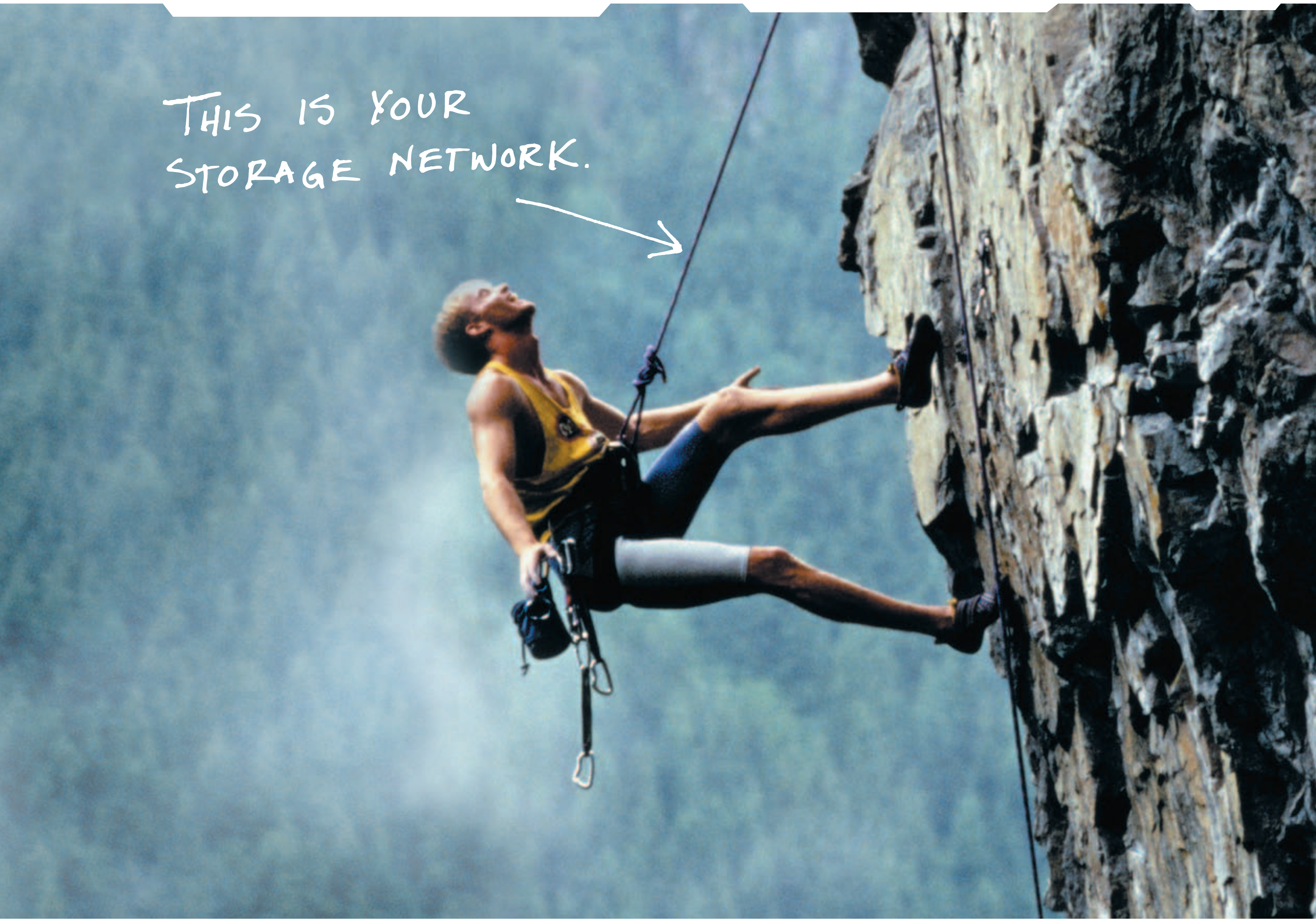
** Features occur in real time; no rebooting is required. ** Virtualization enables legacy applications on current platforms. *** Uniform infrastructure helps reduce preproduction cycle time.*

Figure 5. Static and dynamic energy-reducing features of the VDC

In addition, implementing a VDC can help lower costs, centralize management, and enable dynamic control of the data center. Dynamic VDCs can provide similar benefits in the management of other resources, such as clients; additional resource attributes, such as security, data integrity, or business continuity; and VDC variants, such as clusters and grids.

Robert White (robert_white@dell.com) is an environmental engineer in the environmental affairs department at Dell.

Tim Abels (tim_abels@dell.com) is a senior software architect currently developing scalable enterprise computing systems at Dell. Tim has an M.S. in Computer Science from Purdue University.


THIS IS YOUR
STORAGE NETWORK.

Using Virtual Machines

to Simulate Complex IT Environments

Technical and business solutions designed for today's complex enterprise networks require in-depth testing and validation before release. Real-life scenarios must be simulated during the test cycle—the larger the application scale, the more complex the test environment. This article suggests a method for addressing many of the needs IT administrators encounter when creating complex test environments that comprise multiple networks, servers, and clients. The proposed approach uses hardware virtualization technology to help reduce the costs, resources, and setup time associated with enterprise software testing.

BY AURELIAN DUMITRU AND J. CRAIG LOWERY, PH.D.

Many software products are designed to function as part of a large, distributed network infrastructure. To help ensure that they will perform as expected, such applications must be tested in complex IT environments that simulate real-life scenarios. This article examines the test environment for a product that interfaces with Microsoft® Active Directory® directory service. In this scenario, Dell engineers scrutinized both the reliability of the product interface—that is, its ability to discover a domain controller—and scalability features such as schema changes and functionality in a multidomain, or *forest*, environment.

The product test bed comprised the following:

- One root primary and one backup domain controller
- One Domain Name System (DNS) server
- Several organizational units, each with a few user groups
- A subdomain (such as a departmental domain) serving several remote user groups

More advanced testing required that multiple domain trees run side by side, and that the product function appropriately when trusts between the domain trees were created. The test plan also called for evaluating the product in mixed-performance environments such as networks comprising local area networks (LANs) and wide area networks (WANs).

The scope of this test plan necessitated a large, complex network that could approximate a production environment. Active Directory infrastructures are usually created with tens or even hundreds of computer systems connected in a multiple-segment, multiple-subnet configuration. These environments are typically found only in medium and large corporations and are generally too expensive to set up for testing purposes because of the large number of physical systems required—including servers, client workstations, network routers, and switches. Aside from the cost of the equipment itself, the time to install many different systems can be prohibitive. Hardware virtualization technology can help test engineers simulate complex IT environments by significantly reducing the cost of physical hardware and the time and resources needed to deploy it.

Understanding server virtualization

At the heart of the virtualization concept is the idea that a single physical server can appear to be multiple physical servers, called virtual machines (VMs), as shown in Figure 1. The physical server hardware layer at the bottom of the figure comprises a standard Intel® architecture–based system such as a Dell™ PowerEdge™ server. The virtualization layer comprises software that creates the VMs by multiplexing the physical resources of the underlying server. For example, the physical processor is time-shared, the memory is partitioned, and network traffic is interleaved across the VMs.

Within the virtualization layer, an interesting extension to networking is possible: Virtual networks enable administrators to create multiple subnetworks of VMs, completely contained within one physical system. The virtualization layer can provide gateway services between virtual networks, or between a virtual network and a network external to the physical server. Dell engineers made extensive use of the virtual networking feature when simulating the complex IT environment described in this article.

The manner in which physical disk storage is shared is particularly important. To understand why, first consider that an operating system (OS) installed directly on a physical server maps the blocks of physical disk space into a file system. If all physical disk blocks were serialized into a single structure, the structure could exist as a single file in a larger, encompassing file system. The virtualization layer creates virtual disks in this manner. Each VM has a large file—a virtual disk—that functions as a block-mapped device.

Virtualization layers can be implemented either as special-purpose operating systems, as with VMware® ESX Server™ software, or as applications running on a general-purpose OS. VMware GSX Server™ software and Microsoft Virtual Server software are examples of the latter. In either case, administrators can install software applications on a VM in the same way as on a physical server. Each OS installed on a VM, or *guest OS*, can contain one or more applications. Furthermore, a mixture of Microsoft Windows®, Linux®, and Novell® NetWare® operating systems can reside on a single physical server. The file containing a VM's guest OS and applications is called an image, or *software stack,* and sometimes simply a VM (although technically, a VM is the virtualized hardware and not the software stack that it executes).

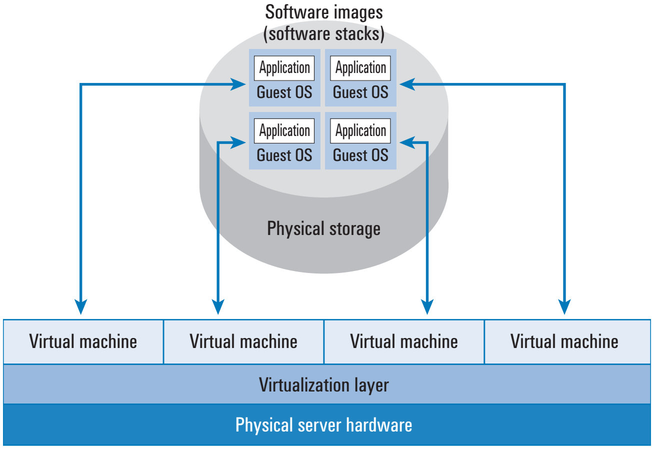


Figure 1. Hardware virtualization architecture

Hardware virtualization technology can help test engineers simulate complex IT environments by significantly reducing the cost of physical hardware and the time and resources needed to deploy it.

Hardware virtualization offers many important benefits—such as consolidation, normalization, isolation, replication, and relocation—that affect not only the creation of complex test environments but many other applications as well. For example, hardware virtualization can enable organizations to reduce the total number of physical systems through *consolidation*. The software stack on each VM "sees" the same virtual hardware regardless of the actual underlying physical hardware because of the *normalization* that the virtualization layer provides. VMs do not interfere with each other and interact only through intended network communications because of the *isolation* that the virtualization layer imposes on each VM. A software stack for which multiple instances will be needed can be cloned by copying the virtual disk's file, allowing for quick *replication*. Finally, because virtualized hardware is the same across all physical systems, VMs can benefit from easy *relocation*—the capability to move a VM from physical server to physical server as needed, either by moving the virtual disk file or by placing the virtual disk file on network shared storage.

Creating the virtual test environment

When Dell engineers built the test environment for the scenario described in this article in January 2004, they relied heavily on three key virtualization benefits: consolidation, isolation, and replication. These features can help reduce the setup time and resources required for large-scale testing.

Consolidation. One inhibitor to creating the physical test environment for a large-scale enterprise application was the sheer number of physical systems required. Although theoretically all the systems could be consolidated onto one large physical server using

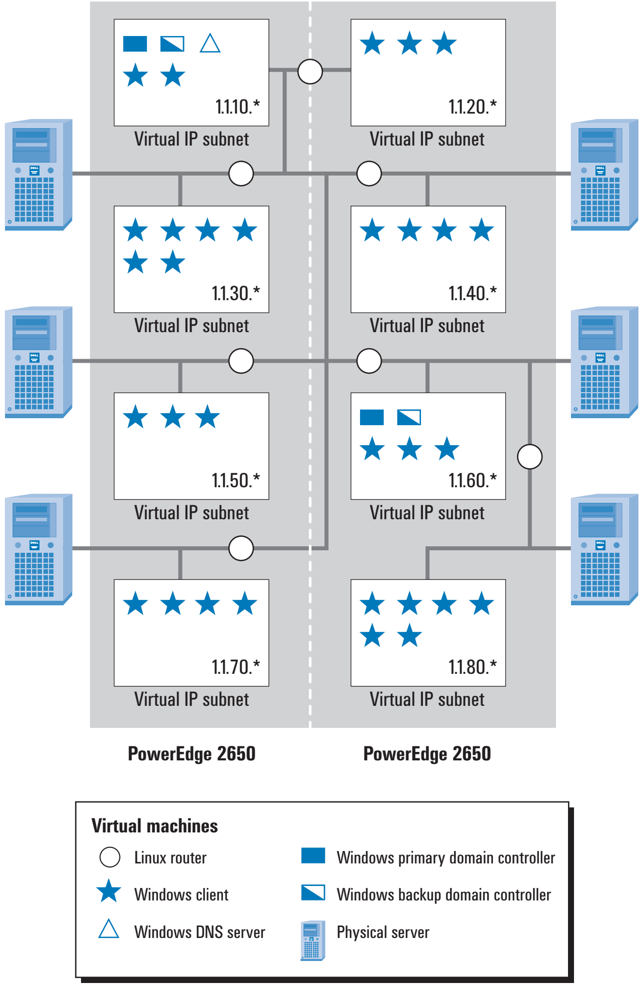


Figure 2. Virtualized network test environment for Active Directory client

virtualization technology, some limitations exist. In all cases, the number of VMs that can be hosted depends on the physical capacity of the underlying server and the aggregate workload characterization of the VMs. For this particular test case, engineers determined that between 10 and 20 VMs could easily coexist on a single PowerEdge 2650 server. And this scenario presented an additional constraint: The product being tested is a physical hardware component and must be installed in a nonvirtualized environment. Therefore, the systems on which the product under test was installed had to be physical systems. However, the rest of the test bed, comprising domain controllers, client workstations, and network routers, did not have the same physical dependency and could be consolidated onto two physical systems.

Isolation. Because virtualization enables multiple, potentially different operating systems to run simultaneously on the same physical server without side effects, the many types of systems found in a heterogeneous network environment can be created virtually. The domain controllers for this test environment as well as client workstations were hosted by virtualized Microsoft Windows 2000 Server systems, and the network routers were virtualized Linux routers. An important aspect of this test environment is that the fact that these systems are virtualized is transparent to other servers, both virtual and physical. For all practical purposes—such as communicating with other virtual and physical systems—VMs have the appearance and functionality of physical systems.

Replication. For this scenario, test engineers needed to construct many instances of domain controllers, client workstations, and network routers. However, they needed to create only one "reference" software stack for each of these categories. Once built, each reference stack was cloned simply by copying and then personalizing the file with a configuration specifying details such as virtual machine ID and network name. Dell engineers used this method to create as many instances of each component as needed.

Figure 2 shows the implementation details for the virtual environment used to test the Active Directory client product. Physical servers on which the product was installed are indicated along the right and left sides of the diagram. The large box in the center shows all of the VMs, in this case distributed across two Dell PowerEdge 2650 servers—each with dual Intel Xeon™ processors, 8 GB of memory, six physical network interfaces, and 138 GB of hard disk space—running VMware ESX Server 2.0.1 software.

The virtual IP subnets contained various VMs that are depicted in Figure 2 as stars, triangles, and rectangles. Six of the eight subnets (1.1.30.* through 1.1.80.*) were connected to both virtual and physical systems. The routers, represented by circles between the virtual subnets, were implemented as VMs running the Linux OS configured as a software router. Although not as fast as the hardware equivalent, a software router—regardless of the implementation, virtual or physical—offers an important benefit: control over the latency of network communication. This software router configurability played an important role in simulating the communication latencies found in LAN and WAN environments.

Evaluating the virtualized Active Directory test environment

The test plan required that engineers simulate an Active Directory domain (including subdomains) with up to 25 users. Figure 3 presents the relevant information for evaluating the impact of virtualization in creating the Active Directory test environment. Comparing an all-physical setup with the hybrid physical-virtual setup used in this test environment shows that virtualization can help reduce deployment efforts and associated costs, and simplify changing test configurations.[1]

[1] 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.

THE SDLT 320.

BECAUSE YOU BELIEVE IN BACKING UP, NOT STARTING OVER.

So do we. Chances are your data is backed up on many of the over 100 million DLTtape™ cartridges shipped to date. Luckily for you, the SDLT 320 won't leave any of your mission-critical data behind. It was designed to provide you with backward-read compatibility to your existing DLTtape IV™ cartridges while providing you with the power, performance and reliability your current applications demand. Using the power of DLTtape Technology, the SDLT 320 boasts a monstrous 160 GB of native capacity and a blistering 16 MB/s native transfer rate. With high performance and ultra-reliability, it's no wonder why DLTtape Technology is the choice of 98% of FORTUNE 500® businesses and is endorsed by leading systems, software and channel partners. To learn more, go to dell.com/SDLT320.

OPTIONS ARE A BEAUTIFUL THING.™

Comparison criteria	All-physical setup	Virtual-physical setup
Initial setup time and associated costs	With all equipment at hand, the physical setup can take anywhere between 24 and 40 hours.	VMs can be easily cloned, which enabled engineers to grow the test configuration from one to 25 clients in about two hours. Each VM is essentially a file that can be manipulated through standard file-copy commands. Each time a VM is cloned through this method, an extra step is required to personalize the new VM. Besides saving time on wiring, copying and personalizing each VM takes only about five minutes as opposed to running a standard installation (at least 20 minutes for Windows 2000 Server, for example). Overall, it should take fewer than 16 hours to deploy hardware to be tested; install ESX Server software; create master copies of the main components: domain controller, DNS server, Dynamic Host Configuration Protocol (DHCP) server, client, and router; clone the VMs as necessary; and check the functionality of the entire system.
Ease of switching test configurations	Switching between test configurations may require rewiring some or all of the systems. Depending on the specific configuration needs, it could take from two hours to a day or more to rewire, change IP addresses, and so on. For example, changing the authentication mode from in-domain to cross-domain may require moving the component under test to a different subnet.	Virtual networks enable network configuration changes to be accomplished in a matter of minutes by bringing up the ESX Server Web console, identifying the VMs that require network configuration changes, and effecting the changes. Depending on the specific configuration needs, new virtual network adapters may need to be created and assigned to specific VMs.
Equipment	Implementing a physical setup comparable to the virtual-physical test environment described in this article would require procuring at least 10 servers, 25 workstations, 8 routers, LAN cables, and so on.	Two PowerEdge 2650 servers can host all the servers, workstations, and routers for the virtual-physical test environment described in this article.
Support effort	The larger the physical test infrastructure, the more time and effort it requires for support and maintenance.	Less hardware requires less support and maintenance. Moreover, if a VM goes down, it can be cloned immediately, reducing overall downtime significantly.

Figure 3. Evaluating virtualization benefits in achieving the base requirement

Figure 4 highlights the benefits of rapid replication in meeting the second test requirement: to create multiple domains. The findings of this test case indicate that cloning allows the virtual infrastructure to be quickly extended, creating additional domains simply by cloning existing domain infrastructure.

As demonstrated by the scenario described in this article, hardware virtualization offers many advantages for building complex test environments. By lowering initial and subsequent setup times, reducing the need for hardware, and decreasing the support effort for large test beds, virtualization can help minimize testing costs and favorably affect schedules—potentially helping to improve time to market for technical and business software products.

Beyond the creation of test environments, hardware virtualization offers many other potential enterprise benefits. For example, reducing the number of physical servers through consolidation can reduce capital and operating expenses as well as ongoing maintenance costs. In addition, managing VM workloads dynamically across the data center can help increase uptime and enable IT organizations to respond quickly and flexibly to business-critical computing demands across the enterprise.[2]

J. Craig Lowery, Ph.D. (craig_lowery@dell.com) is a senior engineering development manager in the Dell Product Group—Enterprise Solutions Engineering. 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.

Comparison criteria	All-physical setup	Physical-virtual setup
Initial setup time and associated costs	The same amount of effort must be spent to create each new domain. Nothing can be reused.	Cloning VMs enabled engineers to create each new domain component in a matter of minutes without incurring the time and expense of physical reconfiguration.
Equipment	The more domains that need to be deployed, the more servers are needed.	Two PowerEdge 2650 servers per domain satisfied the requirement more cost-effectively than a rack of server equipment.

Figure 4. Evaluating virtualization benefits in achieving the scalability requirement

Aurelian Dumitru (aurelian_dumitru@dell.com) is a senior software engineer with the Custom Solutions Engineering team at Dell, where he works to deploy and customize systems management solutions for medium to large enterprises. Before joining the Custom Solutions Engineering team, he was lead engineer for the Remote Management Delivery team. Aurelian has 12 years of experience in hardware, software, and system design and integration and has two patents approved. He has an M.S.E.E. degree from the Technical University of Iasi, Romania.

FOR MORE INFORMATION

Dell and VMware ESX Server:
http://www.dell.com/vmware

[2] For more information, see "Virtualization as an Enterprise Computing Infrastructure" by J. Craig Lowery, Ph.D., in *Dell Power Solutions*, June 2004.

Evaluating Price/Performance of VMware ESX Server on Dell PowerEdge Servers

This article examines the price/performance advantage that enterprises can derive from running VMware® server virtualization software on two four-processor Dell™ PowerEdge™ 6650 servers compared to a single eight-processor IBM® eServer® xSeries® 445 server. In this comparison, Dell engineers demonstrated that the Dell configuration can offer 27 percent better price/performance than the IBM server.

BY TODD MUIRHEAD; DAVE JAFFE, PH.D.; AND FELIPE PAYET

Server virtualization enabled by VMware® ESX Server™ software running on two four-processor Dell™ servers can offer significant benefits when compared to a similar deployment on an eight-processor or larger server. These benefits include risk mitigation, expansion flexibility, and operational flexibility.[1] This article focuses on the price/performance of server virtualization deployments by comparing two four-processor Dell PowerEdge™ 6650 servers to a similarly configured, eight-processor IBM® eServer® xSeries® 445 server—with both configurations running the same VMware virtualization software stack.

The tests performed for this study, in February 2004, simulated an enterprise test-and-development environment in which a 1 GB Microsoft® SQL Server™ database was captured in a virtual machine (VM) and cloned to provide multiple, identical environments for application development and stress-testing. Sixteen VMs were created and then stress-tested on the eight-processor IBM server. Two four-processor PowerEdge 6650 servers, each running eight VMs, were stress-tested in a similar manner. The two PowerEdge 6650 servers, at a total price of $91,604 (including the cost of the VMware ESX Server software), could handle 9.2 percent more orders per minute than the IBM eServer xSeries 445 server, which was priced at $106,442 (also including VMware ESX Server software). As a result, the two PowerEdge servers provided a 27 percent price/performance advantage over the IBM server in this study.[2]

Setting up the test configurations

Two Dell PowerEdge 6650 servers and an IBM eServer xSeries 445 server were configured for the test comparison. Each PowerEdge 6650 used four Intel® Xeon™ processors MP at 2.8 GHz, with 16 GB of memory. The IBM eServer xSeries 445 server used eight Intel Xeon processors MP at 2.8 GHz, with 32 GB of memory.

[1] 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.

[2] U.S. prices for the Dell PowerEdge 6650 servers and the IBM eServer xSeries 445 server are cited from the Dell and IBM online stores, respectively, as of May 7, 2004. These prices also include the cost of VMware ESX Server software for each configuration.

	Each Dell PowerEdge 6650	IBM eServer xSeries 445
Virtualization software	VMware ESX Server 2.0.1	VMware ESX Server 2.0.1
Price of virtualization software	$3,750 per two-CPU license for a total price per server of $7,500	$3,750 per two-CPU license for a total price per server of $15,000
CPU	Four Intel Xeon processors MP at 2.8 GHz with 2 MB level 3 (L3) cache	Eight Intel Xeon processors MP at 2.8 GHz with 2 MB L3 cache
Memory	16 GB	32 GB
Internal disks	Two 18 GB disks	Two 36 GB disks
NICs	Two 10/100/1000 Mbps (internal) Two Intel PRO/1000XT Gigabit Ethernet	Two 10/100/1000 Mbps (internal) Two Intel PRO/1000XT Gigabit Ethernet
Disk controller	PowerEdge RAID Controller 3, Dual Channel (PERC3/DC)	IBM ServeRAID™ Fibre Channel
HBA	QLogic 2340	QLogic 2340
Height	4U (7 inches)	4U (7 inches)
Price of server	$38,302 per server; $76,604 for two servers	$91,442
Total price (servers and VMware software)*	$45,802 per server; $91,604 for two servers	$106,442

**U.S. prices for the Dell PowerEdge 6650 servers and the IBM eServer xSeries 445 server are from the Dell and IBM online stores, respectively, as of May 7, 2004. Each system was configured as shown in Figure 1. Prices include the cost of VMware ESX Server software for each configuration. Neither the IBM price nor the Dell price includes the cost of Fibre Channel HBAs in the server, because HBAs are considered part of the external storage configuration.*

Figure 1. Configuration and prices for the two Dell PowerEdge 6650 servers and the one IBM eServer xSeries 445 server used in the VMware ESX Server test

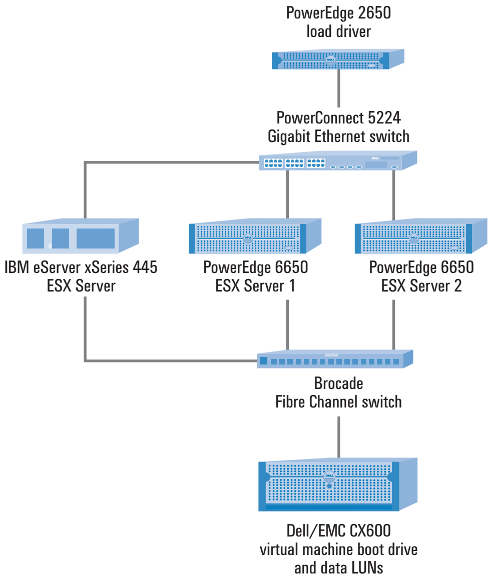


Figure 2. Server and storage configuration for test environment

Thus, the two Dell servers and the single IBM server were configured using the same number of processors and the same amount of memory.

VMware ESX Server software allows administrators to specify how all the Gigabit[3] Ethernet network interface controllers (NICs) on each system are used. In the test configurations, each system included one ESX Server service console, which was used to administer and configure the ESX Server software. The test team dedicated one on-board NIC to the ESX Server service console on each system. In addition, one Intel PRO/1000XT Gigabit Ethernet NIC was dedicated to the VMs on each PowerEdge 6650; on the IBM server, two Intel PRO/1000XT Gigabit NICs were dedicated to the VMs, thus keeping the number of VMs per NIC constant throughout the test. Figure 1 shows the configuration details for each server.

The PowerEdge 6650 servers and the IBM eServer xSeries 445 server were attached to a storage area network (SAN) through QLogic® 2340 Fibre Channel host bus adapters (HBAs). The SAN consisted of a Dell/EMC CX600 Fibre Channel controller and one external DAE2 disk array enclosure with ten 73 GB drives (see Figures 2 and 3).[4] The Dell test team created two 5-disk RAID-5 logical storage units (LUNs) and assigned them to the ESX Server farm. When VMware VirtualCenter and VMotion™ software are used to enable the movement of VMs from one physical server to another, the storage used by the VMs must be visible to all ESX Server hosts. So the test team assigned the two LUNs used by the VMs to all three servers. In addition, the test team used a Dell PowerEdge 2650 server to drive a load against the Microsoft SQL Server databases that were installed in VMs on the three servers.[5]

Creating virtual machines

The Dell test team created one VM to act as the master and then cloned it to create the other VMs. The team created the master VM with 512 MB of RAM, one CPU, a 10 GB hard drive on the SAN, and Microsoft Windows Server™ 2003, Enterprise Edition, as the guest operating system (see Figure 4). Half the VMs were assigned to one LUN

Controller	One Dell/EMC CX600
Disk enclosure	One Dell/EMC DAE2
Disks	Ten 73 GB 10,000 rpm drives
LUNs	Two 5-disk RAID-5 LUNs for VMs
Software	EMC® Navisphere® Manager and Access Logix™

Figure 3. Dell/EMC storage configuration for test environment

[3] 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.

[4] The price of the SAN was not included in the calculation of the IBM and Dell server prices.

[5] The price of the PowerEdge 2650 server was not included in the calculation of hardware and virtual software costs for the two test configurations.

Get there.

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.

27% Better Price/Performance. Fast Business Response.

Four-processor Dell PowerEdge™ 6650 servers running VMware ESX Server™ software can deliver 27% better price/performance on virtualized workloads compared to more expensive, proprietary eight-processor and larger systems.* Plus, a virtualized IT infrastructure on multiple Dell PowerEdge servers 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®

* Results based on a Dell-devised Microsoft® SQL Server™ virtual machine performance test running a 1 GB database. Tests were performed by Dell Labs in February 2004 on two Dell PowerEdge 6650 servers running VMware ESX Server 2.0.1 software, each configured with four Intel® Xeon™ processors MP at 2.8 GHz with 2 MB level 3 (L3) cache, 16 GB of RAM, and two 18 GB hard drives versus an IBM® eServer® xSeries® 445 eight-processor server running VMware ESX Server 2.0.1 software and configured with eight Intel Xeon processors MP at 2.8 GHz with 2 MB L3 cache, 32 GB of RAM, and two 36 GB hard drives. For this test, performance was measured based on number of orders per minute. Actual performance will vary based on configuration, usage, and manufacturing availability. Prices for each system based on list prices from each company's respective Web site as of May 7, 2004. System prices include the hardware cost of each server as well as the software cost of VMware ESX Server, but do not include the price of the storage area network that was attached to all three servers. For further information, see the Dell white paper, "The Scalable Enterprise: Price/Performance Benefits of VMware ESX Server on the Dell PowerEdge 6650 vs. the IBM xSeries 445" available at http://www.dell.com/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.

Operating system	Microsoft Windows Server 2003, Enterprise Edition
Number of CPUs	One
RAM	512 MB
Hard disk size	10 GB

Figure 4. VMware VM settings for test environment

	Two Dell PowerEdge 6650 servers	One IBM eServer xSeries 445 server
Number of VMs	16 (8 per server)	16
Orders per minute (opm)	31,665	28,984
CPU utilization	91%	91%
Dell performance advantage	9.2%	n/a
Configuration price	$91,604	$106,442
Price/performance ($/opm—lower is better)	2.893	3.672
Dell price/performance advantage	27%	n/a

Figure 5. VMware ESX Server test results

and the other half to the other LUN. Similarly, the VMs were assigned to networks in such a way that, on the IBM eServer xSeries 445, half the VMs used one of the Intel NICs and the other half used the other Intel NIC; and on each Dell PowerEdge 6650, all the VMs used a single Intel NIC. This configuration enabled the test team to load balance the VMs across the hardware.

Running the test

Each VM in this test ran a 1 GB Microsoft SQL Server–based application that simulates a DVD store.[6,7] The database on each VM contained 2 million customers; 900,000 orders; and 100,000 titles. The team used the PowerEdge 2650 server to drive an order-entry workload against each VM. Figure 5 shows the results of the test. The 16 VM database servers could handle 9.2 percent more orders per minute when running on the two Dell PowerEdge 6650 four-processor servers than when running on the IBM eServer xSeries 445 eight-processor server. Factoring in the list configuration prices, the IBM server's price/performance was 27 percent higher than the total price/performance of the two PowerEdge 6650 servers.

Server virtualization enabled by VMware ESX Server software running on multiple four-processor Dell servers can offer significant benefits, including risk mitigation, expansion flexibility, and operational flexibility.

Achieving the scalable enterprise

Running identical VMware virtualization software stacks on two Dell PowerEdge 6650 servers and on an IBM eServer xSeries 445 server, and utilizing the Dell-developed database VM performance test described in this study, the Dell systems achieved a price/performance advantage of 27 percent. Consistent with the Dell Scalable Enterprise strategy, this excellent price/performance is an example of the value that Dell systems can provide by leveraging industry-standard components.

Todd Muirhead (todd_muirhead@dell.com) 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. (dave_jaffe@dell.com) 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.

Felipe Payet (felipe_payet@dell.com) manages the Dell and VMware relationship within the Software Alliance Team of the Dell Enterprise Server Group. Previously, he worked in various product management, business development, and emerging technology marketing roles at Dell, Intel, and several startups. Felipe has a B.A. in Economics from Yale University and an M.B.A. from the Sloan School of Management at the Massachusetts Institute of Technology (M.I.T).

FOR MORE INFORMATION

Dell and VMware:
http://www.dell.com/vmware

[6] For more information on this DVD store application, 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. This article describes testing in which a larger (100 GB) version of this application was used to demonstrate scalability features of VMware software on Dell PowerEdge servers.

[7] The use of Microsoft SQL Server in this test does not indicate that Microsoft has tested or certified SQL Server installations on VMware virtualization software. Furthermore, as described in Microsoft Knowledge Base article 273508 (http://support.microsoft.com/default.aspx?scid=kb;[LN];273508), Microsoft does not support issues that occur in Microsoft operating systems or programs running on a VM until the same issue can be reproduced outside the VM environment.

Scaling Out with Oracle RAC 10*g* on Dell Clusters

Oracle® Database 10*g* is designed to enable enterprise grid computing, which comprises clusters of industry-standard servers and modular storage working together to balance workloads by shifting resources on demand. In this way, enterprise grids may provide increased hardware utilization, stability, and scalability while helping to lower total cost of ownership. This article—the result of a joint project between Dell and Oracle engineering teams—explains how organizations can effectively manage and distribute computing power across data centers by using Oracle Real Application Clusters (RAC) 10*g* and Oracle Enterprise Manager 10*g* on Dell™ servers and storage.

BY PAUL RAD; ZAFAR MAHMOOD; IBRAHIM FASHHO; REZA ROOHOLAMINI, PH.D.; AND JOEL BORELLIS

Within large organizations, the segregation of database servers into disparate islands often can lead to inefficient use of available computational and storage resources. Oracle® Database 10*g* helps address this problem by providing an adaptive software infrastructure that enables enterprise grid computing—an approach that efficiently uses clusters of low-cost, industry-standard servers and modular storage to balance workloads and provide capacity on demand.

Throughout the Oracle Database 10*g* development process, Dell and its partners participated in the Oracle Database 10*g* prerelease testing program. The Dell team worked with Oracle engineers to perform solution and interoperability testing on Dell™ PowerEdge™ servers and Dell/EMC storage products to help ensure compatibility with the new database. This article—the result of a joint project between Dell and Oracle engineering teams—explains how enterprises can use Oracle Database 10*g* and Oracle Enterprise Manager 10*g* on industry-standard Dell PowerEdge servers to cost-effectively manage and distribute data processing and storage across data centers.

Enabling a cluster-based Oracle environment

Oracle Real Application Clusters (RAC) 10*g*, an option to Oracle Database 10g that allows the database to run business applications on clusters, can provide the foundation for an enterprise grid environment. Oracle RAC is a cluster database designed with a shared-cache architecture. It runs on multiple servers that are attached to shared storage.

An Oracle RAC database appears as a single standard Oracle database to end users, but administrators can use similar maintenance tools and practices for a single Oracle database on the entire cluster. All standard backup and recovery operations, including the use of Oracle Recovery Manager, work transparently with Oracle RAC. All SQL operations are also identical for both single and cluster configurations.

In the releases of Oracle Database 10*g* and Oracle RAC 10*g*, Oracle has introduced several product features, product enhancements, and manageability improvements that can help reduce database administration effort and provide increased administrative flexibility.

Enhanced cluster availability

With Oracle Database 10*g*, Oracle introduced its own clusterware, called Cluster Ready Services (CRS). CRS performs the node-monitoring functions and supports the high-availability features of Oracle RAC. The integration of CRS and Oracle RAC provides high availability—not only for Oracle instances in the cluster, but also for supporting applications. To support high availability for the diverse applications that run within the RAC environment, CRS defines resources whose availabilities are managed by the clusterware. The resources that CRS manages comprise processes and other entities associated with a RAC database, including management and availability functions. These resources are automatically created and defined to CRS either during installation or afterward through standard user-interface tools. Resources that CRS automatically manages include the Global Services Daemon (GSD), the Oracle Notification Service (ONS) daemon, virtual IP (VIP) addresses, listeners, databases, instances, and services. CRS is the backbone of the high-availability features that are now inherent in RAC.

Simplified storage management

Automatic Storage Management[1] (ASM) helps simplify the administration of Oracle database files. ASM allows administrators to manage disk groups instead of individual database files. Administrators can define a particular disk group as the default disk group for a database.

High availability and reliable recovery

Oracle Database 10*g* extends Oracle Flashback Technology, which was introduced in the Oracle9*i*™ database. This technology includes the Flashback Database feature, which helps administrators quickly recover the entire Oracle database to a previous state from a past checkpoint. In addition to performing flashback operations at the database level, administrators are enabled to recover an entire table using the Flashback Table feature. The Flashback Recovery feature allows the database to recover tables that administrators have inadvertently dropped.

Enterprises can use Oracle Database 10*g* and Oracle Enterprise Manager 10*g* on industry-standard Dell PowerEdge servers to cost-effectively manage and distribute data processing and storage across data centers.

Oracle Database 10*g* Enterprise Edition extends Oracle Data Guard by including the Real Time Apply feature, which enables standby databases to be closely synchronized with the primary database. Real Time Apply provides many benefits, including quick switchover and failover operations, instantly up-to-date results after an administrator changes a physical standby database to read only, up-to-date reporting from a logical standby database, and the ability to use large log files for site disaster recovery.[2]

Efficient data propagation

For most enterprise applications, data must be shared as it is created or changed across databases rather than being occasionally updated in bulk. To address this need, Oracle Database 10*g* includes the Oracle Streams feature, which propagates data between databases and nodes to keep two or more copies in sync as updates are applied. It also provides a unified framework for information sharing—combining message queuing, replication, events, data-warehouse loading, notifications, and publish-subscribe operations.

Oracle Database 10*g* has a heterogeneous, transportable tablespace feature that lets administrators quickly move a tablespace across Oracle databases. A combination of the Oracle Streams and transportable tablespace features allows administrators to issue a single command that can ship a tablespace from one database to another database, reformat the tablespace if the second database is in a different Endian format (byte order), mount this tablespace into the second database, and start syncing the tablespace with changes in the first database.

[1] For more information, see "Enabling a Highly Scalable and Available Storage Environment Using Oracle Automatic Storage Management" by Zafar Mahmood, Joel Borellis, Mahmoud Ahmadian, and Paul Rad in *Dell Power Solutions*, June 2004.

[2] For more information, see "Optimizing Disaster Recovery Using Oracle Data Guard on Dell PowerEdge Servers" by Paul Rad, Zafar Mahmood, Ibrahim Fashho, Raymond Dutcher, Lawrence To, and Ashish Ray in *Dell Power Solutions*, March 2004.

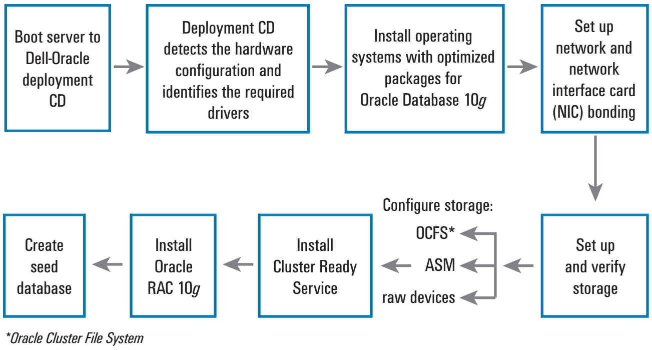

Boot server to Dell-Oracle deployment CD
Deployment CD detects the hardware configuration and identifies the required drivers
Install operating systems with optimized packages for Oracle Database 10*g*
Set up network and network interface card (NIC) bonding
Set up and verify storage
Configure storage:
OCFS*
ASM
raw devices
Install Cluster Ready Service
Install Oracle RAC 10*g*
Create seed database
*Oracle Cluster File System

Figure 1. Oracle Database 10*g* installation using Dell-Oracle deployment CDs

Installing Oracle Database 10*g* on Dell PowerEdge servers

Dell and Oracle have compatibility-tested several Dell server-storage configurations with Oracle Database 10*g*. Supported servers include the PowerEdge 6650, PowerEdge 6600, PowerEdge 4600, PowerEdge 2650, PowerEdge 2600, and PowerEdge 1750 in the following configurations:[3]

- **Single node:** Node uses local storage or direct attach storage
- **Two nodes (SCSI cluster):** Each node connects to a Dell PowerVault™ 220S SCSI enclosure for shared storage through a Dell PowerEdge Expandable RAID Controller 3, Dual Channel (PERC 3/DC) or PERC 4/DC
- **Up to eight nodes (Fibre Channel Cluster):** All nodes and shared Fibre Channel storage (Dell/EMC CX200, CX300, CX400, CX500, CX600, and CX700 storage systems) connect through a McDATA® or Brocade® Fibre Channel switch

Figure 1 shows the procedures for installing Oracle Database 10*g* on the PowerEdge servers using the Dell-Oracle deployment CDs. These CDs provide utilities and functions for setting up a PowerEdge server to run Oracle Database 10g and Oracle RAC 10*g*. The deployment CDs are designed to deliver the most current Dell-optimized drivers and to configure the operating system environment—such as partitioning and RPM™ (Red Hat® Package Manager) packages—for running Oracle database applications on PowerEdge servers.

Managing the Oracle environment

Because the scale-out computing paradigm requires scalable management tools across the entire solution stack, tools and techniques that depend on proprietary protocols to manage individual systems can be ineffective. One solution is to enable technology providers to automate the management of their individual components and then integrate these components into comprehensive management offerings through the use of open management standards.

Dell OpenManage™ Server Administrator and Oracle Enterprise Manager 10*g* help simplify the management of an Oracle 10*g* enterprise environment on Dell PowerEdge servers by using open management standards such as Simple Network Management Protocol (SNMP), as shown in Figure 2.

Hardware management through Server Administrator

Dell OpenManage Server Administrator provides administrators with comprehensive, one-to-one hardware management capabilities within the data center. Server Administrator features include proactive monitoring of server health, diagnostics for troubleshooting, alerts and notifications, and remote access.

Software management through Enterprise Manager

Oracle Enterprise Manager 10*g* is a management application for administering, monitoring, and tuning the Oracle environment. Oracle designed Enterprise Manager to improve the manageability of Oracle RAC, enhancing it to help make RAC as manageable as any single database instance. Enterprise Manager 10*g* delivers a single-system image of RAC databases, providing consolidated screens for managing and monitoring individual cluster components.

Oracle Enterprise Manager 10*g* provides two different management frameworks—Grid Control and Database Control. RAC is supported in both modes. Enterprise Manager 10*g* Database Control is configured within the same ORACLE_HOME directory of the database target and can be used to manage only one database at a time. Grid Control can be used to manage multiple databases, Oracle Application Server 10*g*, and other target types in the enterprise across different ORACLE_HOME directories.

Enterprise Manager 10*g* Grid Control offers cluster-level management, using the standard SNMP protocol to pull content from

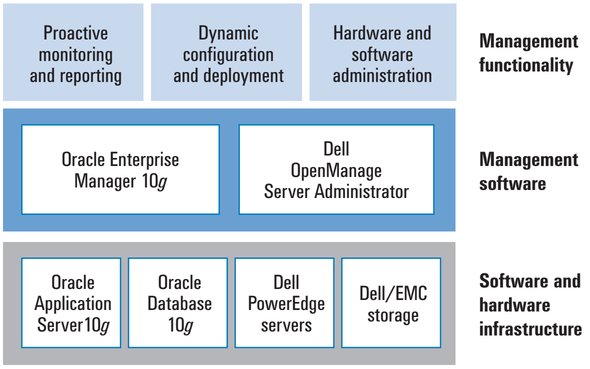
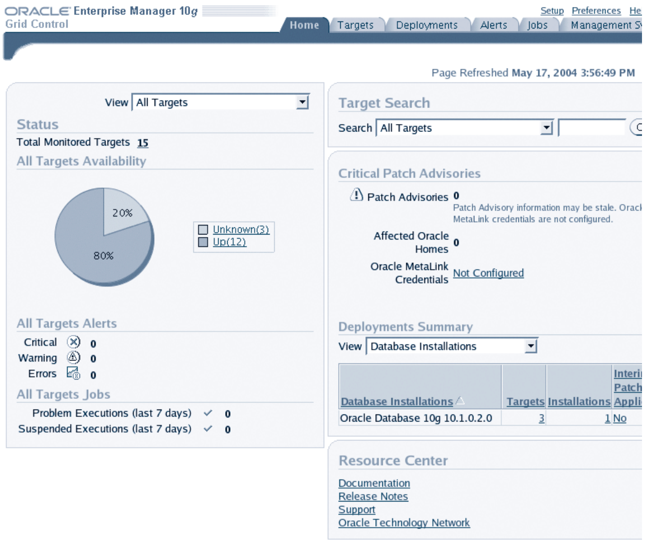

Proactive monitoring and reporting
Dynamic configuration and deployment
Hardware and software administration
Management functionality
Oracle Enterprise Manager 10*g*
Dell OpenManage Server Administrator
Management software
Oracle Application Server10*g*
Oracle Database 10*g*
Dell PowerEdge servers
Dell/EMC storage
Software and hardware infrastructure

Figure 2. Oracle Enterprise Manager 10*g* and Dell OpenManage Server Administrator provide comprehensive management capabilities

[3] To review the latest supported configurations of Oracle9*i* and Oracle Database 10*g* on Dell PowerEdge servers, visit http://www.dell.com/oracle.



Page Refreshed **May 17, 2004 3:56:49 PM**

View: All Targets

Status

Total Monitored Targets 15

All Targets Availability

☐ Unknown(3)
☐ Up(12)

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

Critical Patch Advisories

Patch Advisories 0
Patch Advisory information may be stale. Orac
MetaLink credentials are not configured.

Affected Oracle Homes 0

Oracle MetaLink Credentials Not Configured

Deployments Summary

View: Database Installations

Database Installations	Targets	Installations	Interi Patch Appli
Oracle Database 10g 10.1.0.2.0	3	1	No

Resource Center

Documentation
Release Notes
Support
Oracle Technology Network

Figure 3. Oracle Enterprise Manager 10*g* Grid Control: Cluster Home page

Oracle management agents running on each server. Discovery of a cluster is performed similarly to that of regular nodes. The application discovers the cluster and its related targets, including nodes, listeners, RAC databases, RAC instances, and other targets. Administrators can also proactively update the cluster configuration when nodes are added or deleted. Credentials can be set at the cluster and cluster database level, ensuring single sign-on to the RAC environment.

Exploring Oracle Enterprise Manager 10*g*

As shown by the Cluster Home page in Figure 3, the look-and-feel of Enterprise Manager 10*g* has changed significantly from previous releases. On this page, Enterprise Manager represents a single-system image for a RAC cluster as a composite target composed of nodes and cluster databases. RAC-specific information on this page includes a summary of the number of critical, warning, and error cluster alerts, as well as the number of problem executions and suspended executions of cluster jobs in the past seven days. Administrators can click the number associated with each alert to see a detailed list of the cluster database–related alerts and when they occurred. Also, a separate alerts section for the entire cluster centralizes alert reporting on hosts across all the nodes in the cluster.

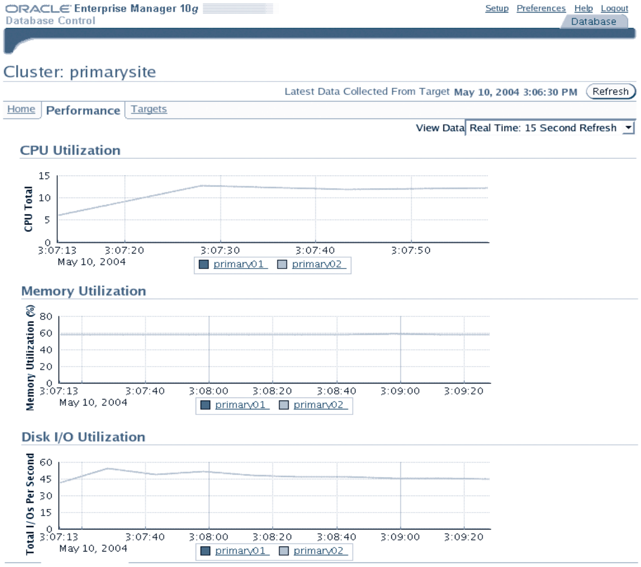
In addition, an overall summary of the cluster is provided on the Cluster Home page. The page shows the current status and cluster availability over the past 24 hours. A cluster is deemed to be up if at least one cluster node is running. The cluster is down if all nodes are down. In the configuration area, administrators can see the clusterware version, along with the hardware and operating system of the cluster.

Enterprise Manager 10*g* also provides a page that summarizes various performance metrics of a cluster. Figure 4 shows an example of the Cluster Performance page, which displays an overview of CPU, memory, and disk I/O utilization for each node in the cluster.

Centrally managing the Oracle environment using Enterprise Manager 10*g* Grid Control

In Oracle Enterprise Manager 10*g* Grid Control, Oracle provides a single, central console for managing the Oracle environment. Enterprise Manager also offers complete access to management information. Its HTML-based console interface allows an administrator to manage the Oracle environment remotely. Figure 5 shows the architecture of a Web application that can be monitored using Enterprise Manager 10*g* Grid Control. The framework consists of multiple components, such as:

- Oracle Management Service, a Java™ 2 Platform, Enterprise Edition (J2EE™) Web application, working with Management Repository, a centralized database for management information about all targets and applications managed by Enterprise Manager
- Oracle Enterprise Manager 10*g* Grid Control Management Agents, which reside on each monitored target, to communicate information to Oracle Management Service
- A four-node Oracle RAC 10*g* cluster, which provides the data for the application
- Middle-tier components, such as Oracle Web Cache, which provide performance improvements
- Oracle Application Server components, such as Oracle Application Server Containers for J2EE or Oracle HTTP Server, to develop custom applications



Cluster: primarysite

Latest Data Collected From Target **May 10, 2004 3:06:30 PM**

View Data: Real Time: 15 Second Refresh

Figure 4. Oracle Enterprise Manager 10*g* Database Control: Cluster Performance page

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

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:

- **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.

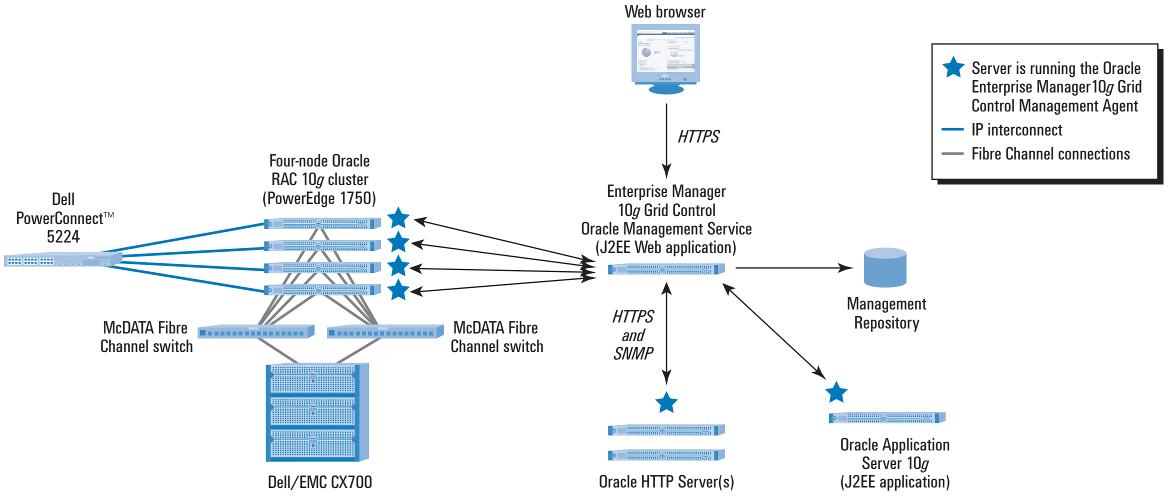

Web browser
HTTPS
Server is running the Oracle Enterprise Manager10g Grid Control Management Agent
IP interconnect
Fibre Channel connections
Dell PowerConnect™ 5224
Four-node Oracle RAC 10g cluster (PowerEdge 1750)
Enterprise Manager 10g Grid Control Oracle Management Service (J2EE Web application)
Management Repository
McDATA Fibre Channel switch
McDATA Fibre Channel switch
HTTPS and SNMP
Dell/EMC CX700
Oracle HTTP Server(s)
Oracle Application Server 10g (J2EE application)

Figure 5. Architecture of a Web application managed by Enterprise Manager 10*g* Grid Control

Enterprise Manager provides tools for administrators to manage each of these components. Creating a Web Application target in Enterprise Manager allows administrators to combine the individual components of a Web application into a single managed target.

Enabling scalable management of a scale-out environment

As enterprises seek cost-effective options for scaling out their IT infrastructures, many are turning toward enterprise grids. Dell produces entry-level, midrange, and high-end clusters built from industry-standard servers and modular storage that work together to help improve scalability in an enterprise grid system. Oracle Database 10*g*, which is designed to enable enterprise grids, can effectively distribute computing power across data centers, managing workloads by adding or removing nodes based on variable business processing needs. This capability allows data center managers to easily scale out in small increments using two- and four-processor servers—a cost-effective method for achieving performance and reliability objectives.

A scale-out infrastructure requires scalable management tools. Dell and Oracle have developed such a solution, using Oracle Enterprise Manager 10*g* and Dell OpenManage Server Administrator to help provide simple, standards-based management of an Oracle enterprise database environment running on Dell servers and storage.

Paul Rad (paul_rad@dell.com) is a senior software engineer in the Dell Database and Application Engineering Department of the Dell Product Group. Paul's current interests are in the areas of operating systems, database systems, clustering, storage technologies, and virtualization. Paul has master's degrees in both Computer Engineering and Computer Science from The University of Texas at San Antonio.

Zafar Mahmood (zafar_mahmood@dell.com) is a software engineer in the Dell Database and Application Engineering Department of the Dell Product Group. He has been involved in database performance optimization, database systems, and database clustering solutions for more than six years. Zafar has an M.S. in Electrical Engineering with a specialization in Computer Communications from the City University of New York.

Ibrahim Fashho (ibrahim_fashho@dell.com) is the development manager for the Database and Application Solutions Group at Dell. His responsibilities include delivering Dell/Oracle solutions based on PowerEdge servers, PowerVault storage, Dell/EMC storage, and the appropriate interconnects. Ibrahim has a B.S. in Electrical and Computer Engineering from The University of Texas at Austin.

Reza Rooholamini, Ph.D. (reza_rooholamini@dell.com) is the director of the Enterprise Solutions Engineering Group at Dell, which develops Linux® and cluster products. He has a B.S. in Electrical Engineering from the University of Illinois at Urbana, an M.S. in Electrical Engineering and an M.S. in Computer Science from the University of Wisconsin, and a Ph.D. in Computer Science/Engineering from the University of Minnesota. Reza has over 30 publications in areas of his research interest, including distributed systems, multimedia systems, high-performance computing, storage systems, high-availability clustering, and interconnects.

Joel Borellis (joel.borellis@oracle.com) is a technology director for Oracle, working exclusively with Dell for the past two years. He is in the Oracle Server Technologies Organization, where he works closely with the Dell Database and Application Engineering Department to build Dell/Oracle solutions.

FOR MORE INFORMATION

Dell and Oracle solutions:
http://www.dell.com/oracle

Oracle Technology Network:
http://otn.oracle.com

Open management standards:
http://www.dmtf.org/standards

Enabling a Highly Scalable and Available Storage Environment Using Oracle

Automatic Storage Management

As databases grow larger, requiring more disk drives and cluster nodes, traditional storage management techniques can become less efficient, more complex, and prone to human error. Automatic Storage Management (ASM), a volume manager for Oracle® Database 10*g*, helps simplify and automate many of the difficult storage management tasks surrounding large Oracle databases. This article introduces the concept of ASM and presents techniques for implementing Oracle Database 10*g* databases on Dell/EMC Fibre Channel and Dell™ PowerVault™ SCSI storage arrays.

BY ZAFAR MAHMOOD, JOEL BORELLIS, MAHMOUD AHMADIAN, AND PAUL RAD

As database storage requirements approach thousands of disk drives, or tens of nodes in a cluster, traditional techniques for file management can stop working. Traditional alternatives for storing database files—such as Oracle® Cluster File System, third-party volume managers, and raw devices—do not always scale efficiently. In addition, such approaches require independent effort on every node within a cluster and thus become too prone to human error. With so many disks, tasks such as storage provisioning, disk initialization, file system expansion, and manual load balancing can become complex and cost-prohibitive.

Database administrators (DBAs) need tools that can help increase their productivity and automate many manual tasks. To address these needs, Oracle has introduced Automatic Storage Management (ASM), a new feature in Oracle Database 10*g* that helps simplify the routine management of physical storage media. This article describes how ASM can solve many of the practical file management problems inherent in large Oracle databases and how ASM functionality can be applied to Dell/EMC Fibre Channel and Dell™ PowerVault™ SCSI storage arrays.

Introducing Oracle Automatic Storage Management

ASM is an Oracle database file system and volume manager that is built into the Oracle Database 10*g* kernel. It provides both a graphical user interface (GUI) through the Oracle Enterprise Manager 10*g* Grid Control management application[1] and a simple SQL-based interface that allows DBAs to use familiar commands such as `CREATE`, `ALTER`, and `DROP`.

[1] For more information, see "Scaling out with Oracle RAC 10*g* on Dell Clusters" by Paul Rad; Zafar Mahmood; Ibrahim Fashho; Reza Rooholamini, Ph.D.; and Joel Borellis in *Dell Power Solutions*, June 2004.

ASM enables storage management of all Oracle files—including data files, log files, control files, archive logs, and Recovery Manager backup sets—across thousands of disk drives with minimal management overhead. ASM also provides storage management for multiple nodes in an Oracle Real Application Clusters (RAC) database environment, as well as for single-node symmetric multiprocessing (SMP) servers.

Using ASM, DBAs can dynamically scale out their Oracle database storage without affecting database availability. ASM also automatically load balances I/O across all available disks defined in the storage environment, helping to prevent hotspots and to maximize performance despite rapidly changing data-usage patterns. Additionally, ASM helps reduce disk fragmentation such that data relocation is unnecessary for reclaiming unused space. ASM facilitates automatic online disk space reorganization for the incremental addition or removal of storage capacity.

DBAs can use ASM to maintain redundant copies of data for increased fault tolerance at the database level, or they can configure ASM on top of hardware-based RAID technologies such as those available in Dell/EMC and Dell PowerVault storage arrays (see Figure 1). Moreover, ASM automatically tunes I/O operations based on performance characteristics for specific classes of data—such as redo, data, control, archive, and temporary data files—without requiring human interaction on a per-file basis.

Enabling an ideal database storage environment

ASM helps address the critical requirements of an optimal database storage environment, which should provide the following features:

- **Dynamic provisioning:** The storage environment should dynamically scale the storage without affecting database availability.
- **I/O balancing:** The storage environment should load balance the I/O across available I/O channels and spindles using an optimal striping methodology, such as hardware RAID implemented at the storage-array level and EMC® PowerPath® path-management software configured to load balance the I/O across host bus adapters (HBAs). If the number of I/O channels or disk spindles increases, the storage environment should be able to load balance the I/O across the new channels and spindles as needed with minimal intervention.
- **High availability:** The storage environment should provide fault tolerance and high availability in case of disk or I/O channel failures using optimal mirroring techniques, such as hardware RAID implemented at the storage-array level and EMC PowerPath software configured to load balance the I/O across HBAs.
- **High performance—application I/O:** The storage environment should be optimized according to application I/O behavior and needs. For example, Oracle RAC 10*g* databases require direct I/O that bypasses the file-system buffer cache.
- **High performance—direct I/O:** Current mechanisms to provide direct I/O support for Oracle database files are either too complex, as with raw devices, or not well integrated with Oracle database releases. An ideal storage mechanism should be able to provide high direct-I/O performance without compromising manageability or robustness.
- **High performance—asynchronous I/O:** The storage environment should be able to perform kernel mode asynchronous I/O. A file system–based storage environment requires special configuration files, device drivers, or kernel parameter tweaks to enable asynchronous I/O—adding further complexity to the existing storage environment.

Dynamic provisioning and storage scalability

ASM virtualizes database storage into disk groups, which serve as repositories for Oracle database files. A disk group is a logical grouping of several individual disks in a storage array or of several RAID logical storage units (LUNs). ASM enables DBAs to manage a small set of disk groups, and ASM automates the placement of the database files within those disk groups.

Adding or removing disks or LUNs to and from a disk group is a dynamic process that does not affect database availability. Figure 2 shows an example in which storage is added to an existing disk group, groupA, on the fly. Without shutting down the database,

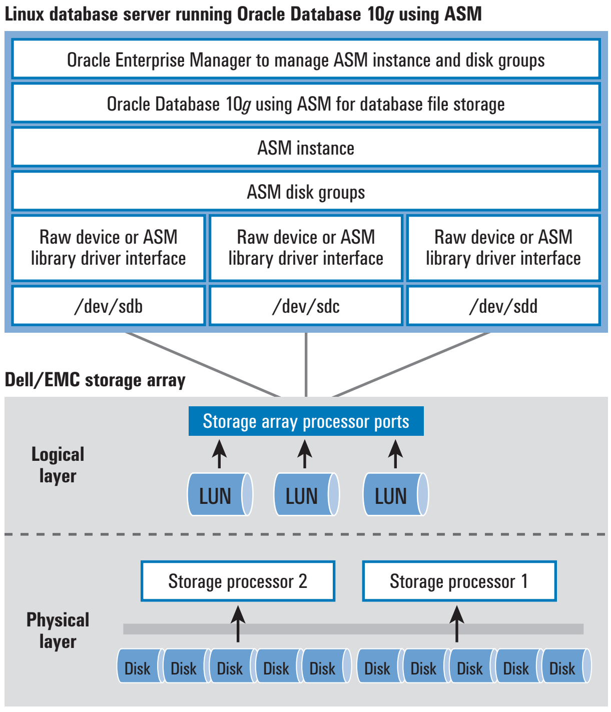

Linux database server running Oracle Database 10g using ASM
Oracle Enterprise Manager to manage ASM instance and disk groups
Oracle Database 10g using ASM for database file storage
ASM instance
ASM disk groups
Raw device or ASM library driver interface
Raw device or ASM library driver interface
Raw device or ASM library driver interface
/dev/sdb
/dev/sdc
/dev/sdd
Dell/EMC storage array
Storage array processor ports
Logical layer
LUN
LUN
LUN
Storage processor 2
Storage processor 1
Physical layer
Disk
Disk
Disk
Disk
Disk
Disk
Disk
Disk
Disk
Disk

Figure 1. Basic architecture of ASM

DELL™

ASM automatically formats the new disks and rebalances the disk group by striping data across all available disks.

High storage availability

Disk groups have three redundancy options for fault tolerance and availability.

External redundancy. When ASM incorporates external storage LUNs that already provide fault tolerance using RAID techniques, DBAs can configure disk groups to use the external redundancy of the storage array. DBAs should use external redundancy when creating disk groups on Dell/EMC Fibre Channel storage arrays or Dell PowerVault SCSI storage arrays, which have hardware-based RAID provided by Dell PowerEdge Expandable RAID Controller (PERC) cards. See "Considerations when using external redundancy" for storage array configurations that can help achieve high performance and data availability when using external redundancy.

Figure 3 shows a command that creates a disk group, groupA, with three RAID LUNs that are configured on a Dell/EMC Fibre Channel storage array.

Normal redundancy. When the external storage array does not use RAID, DBAs can configure disk groups as a two-way mirror. Drives in a disk group can be divided into failure groups, which are disk collections that can become unavailable through failure of an associated component, such as disks or controllers. To achieve high availability, the disks of each failure group within a given disk group should reside on separate controllers.

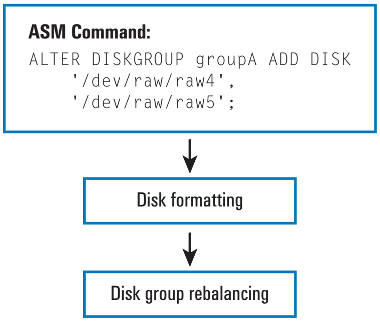
In an environment with multiple storage arrays, DBAs can implement failure groups by creating a single disk group across all arrays. A failure group within the disk group corresponds to disks from just one array. This usage protects against failure of a storage array.

The sample command shown in Figure 4 uses two disk controllers: controller1 and controller2. Each controller is connected

CONSIDERATIONS WHEN USING EXTERNAL REDUNDANCY

ASM does not use failure groups with the external redundancy option. Instead, it depends on the redundancy provided by the intelligent storage array. Storage arrays hide the physical boundaries of disks required for proper placement of data. Therefore, for optimal performance and data availability, the storage arrays should be created with disks that are identical in size, speed, and data-transfer capacity. For example, a RAID-0+1 array of four 146 GB disks yields a capacity of 292 GB, which can be subdivided into two logical disks: one for the database and the other for the flash recovery area. Both logical disks can then be put in the same ASM disk group. The inner portion of this logical disk should be used as the flash recovery area, and the outer portion, which exhibits better performance, should be used as the database area.

A logical disk that resides on the outer portion of the drives in the array performs better than one located elsewhere on the drive because the outer portion of a drive has a higher rotational speed and more sectors per track; logical disks corresponding to inner tracks have a slower transfer rate. Therefore, for better performance, the database area should be the logical disk that is located on the outer portion of the array. Logical disks with different performance characteristics should be put in separate disk groups.



ASM Command:

```
ALTER DISKGROUP groupA ADD DISK
    '/dev/raw/raw4',
    '/dev/raw/raw5';
```

Figure 2. Adding storage to a disk group on the fly

```
CREATE DISKGROUP groupA EXTERNAL REDUNDANCY DISK
'/dev/raw/raw1',
'/dev/raw/raw2',
'/dev/raw/raw3';
```

Figure 3. Creating a disk group with external redundancy

```
CREATE DISKGROUP groupA NORMAL REDUNDANCY
FAILGROUP controller1 DISK

'/dev/raw/raw1',
'/dev/raw/raw2',
'/dev/raw/raw3'

FAILGROUP controller2 DISK

'/dev/raw/raw4',
'/dev/raw/raw5',
'/dev/raw/raw6';
```

Figure 4. Creating a disk group with normal redundancy

to an equal number and size of disks. When the disk group is created, the normal redundancy option is specified along with the FAILGROUP keyword. Each failure group consists of an independent disk controller and disks that ASM uses for a two-way mirror. When the Oracle database creates or updates files on a normal-redundancy disk group, it automatically mirrors data between both failure groups to provide high availability if a disk or controller fails.

> ASM automates most of the manual steps required by the other approaches, so that DBAs do not need to make complex provisioning decisions and are freed for other critical tasks.

High redundancy. The high-redundancy option provides three-way mirroring for a disk group. This option requires that three sets of disks of the same size reside on three separate controllers. When the Oracle database creates or updates files on a high-redundancy disk group, it automatically mirrors data across three failure groups to provide a higher degree of protection than the normal-redundancy option.

High storage performance

ASM provides a built-in mechanism for implementing the Oracle stripe and mirror everything (SAME) methodology, in which all types of data are striped and mirrored across all available drives. The SAME approach can result in an I/O load that is evenly distributed and balanced across all disks within the disk group.

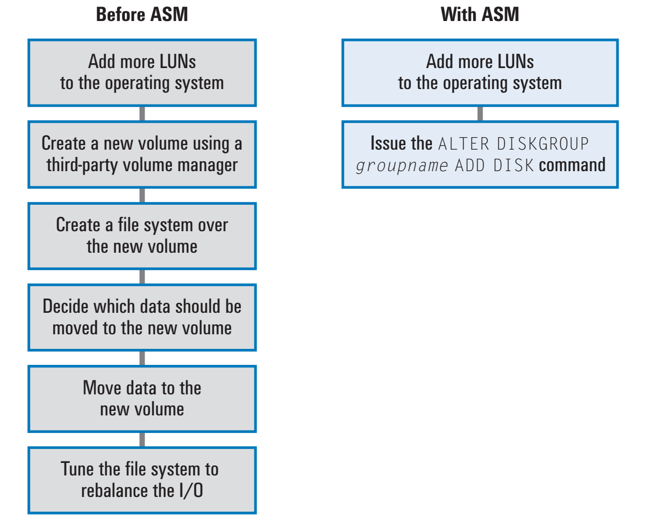

Before ASM
Add more LUNs to the operating system
Create a new volume using a third-party volume manager
Create a file system over the new volume
Decide which data should be moved to the new volume
Move data to the new volume
Tune the file system to rebalance the I/O
With ASM
Add more LUNs to the operating system
Issue the ALTER DISKGROUP groupname ADD DISK command

Figure 5. Scaling up storage for Oracle Database 10*g*

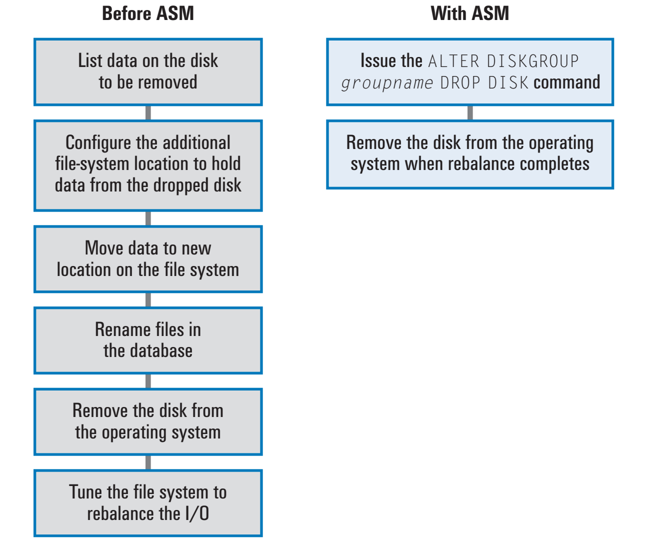

Before ASM
List data on the disk to be removed
Configure the additional file-system location to hold data from the dropped disk
Move data to new location on the file system
Rename files in the database
Remove the disk from the operating system
Tune the file system to rebalance the I/O
With ASM
Issue the ALTER DISKGROUP groupname DROP DISK command
Remove the disk from the operating system when rebalance completes

Figure 6. Scaling down storage for Oracle Database 10*g*

ASM also can be used to create disk groups across multiple storage array–based RAID LUNs. This technique, called *plaiding*, results in double striping: striping at the storage-array level and striping at the host level.

In a disk group with external redundancy, the SAME approach should be used at the storage-array level. All LUNs created in the Dell/EMC storage array should be configured as RAID-10. If the storage array's RAID is not being used, then disk groups should be created using normal or high redundancy across multiple I/O controllers, which may result in optimal I/O performance for Oracle Database 10*g*.

ASM also uses direct I/O to access Oracle database files; file systems lack this capability without tweaking extra kernel parameters or relinking Oracle binaries with asynchronous I/O (AIO) libraries. In an Oracle RAC environment, direct I/O is required to avoid corrupting database files, because the file-system buffer cache is local to each node and other nodes in the cluster do not have access to it.

Moreover, ASM provides kernel asynchronous I/O by default without requiring special parameter settings. Using asynchronous—or *nonblocking*—I/O can improve performance over serial I/O: asynchronous I/O requires fewer I/O processes to handle multiple I/O requests because pending I/Os do not need to wait in the queue.

Simplifying storage provisioning

Compared to other storage options such as raw devices, Oracle Cluster File System, and third-party volume managers, ASM can greatly simplify the process of scaling storage, as shown in Figures 5 and 6. ASM automates most of the manual steps required by the

CREATING AN ORACLE DATABASE 10*g* DATABASE USING ASM

This sidebar describes the basic procedure for configuring and using ASM with Oracle Database 10*g* databases running on the Red Hat® Enterprise Linux® operating system. In this example, all three LUNs—/dev/sdb, /dev/sdc, and /dev/sdd—have been configured as RAID-10 according to the Oracle recommendation for using the SAME methodology. Because RAID is provided at the storage-array level, disk groups are created using the external-redundancy option.

1. Verify that the external storage devices are visible on the Linux host:

   ```
   cat /proc/partitions
   ```

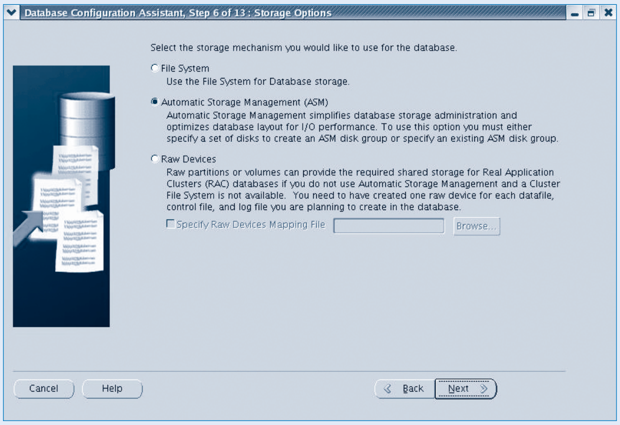
   The configured storage LUNs on the Dell/EMC or Dell PowerVault storage array should appear as follows:

   ```
   /dev/sdb
   /dev/sdc
   /dev/sdd
   ```

   Partitions do not need to be created on the storage devices because ASM can incorporate whole disks when creating or expanding a disk group.

2. ASM on Linux hosts requires that the external storage devices either be bound to raw character devices or be configured using the ASM library driver. To help easily identify the raw devices, rename them using ASM-friendly names:

   ```
   mv /dev/raw/raw1     /dev/raw/ASM1
   mv /dev/raw/raw2     /dev/raw/ASM2
   mv /dev/raw/raw3     /dev/raw/ASM3
   ```



Database Configuration Assistant, Step 6 of 13 : Storage Options

Select the storage mechanism you would like to use for the database.

○ File System
Use the File System for Database storage.

● Automatic Storage Management (ASM)
Automatic Storage Management simplifies database storage administration and optimizes database layout for I/O performance. To use this option you must either specify a set of disks to create an ASM disk group or specify an existing ASM disk group.

○ Raw Devices
Raw partitions or volumes can provide the required shared storage for Real Application Clusters (RAC) databases if you do not use Automatic Storage Management and a Cluster File System is not available. You need to have created one raw device for each datafile, control file, and log file you are planning to create in the database.

□ Specify Raw Devices Mapping File

Figure A. Selecting ASM as the storage mechanism

3. Make user *oracle* of group *dba* the owner of the candidate disks:

   ```
   chown oracle.dba /dev/raw/ASM*
   ```

4. Bind the storage devices to the raw character devices:

   ```
   raw /dev/raw/ASM1    /dev/sdb
   raw /dev/raw/ASM2    /dev/sdc
   raw /dev/raw/ASM3    /dev/sdd
   ```

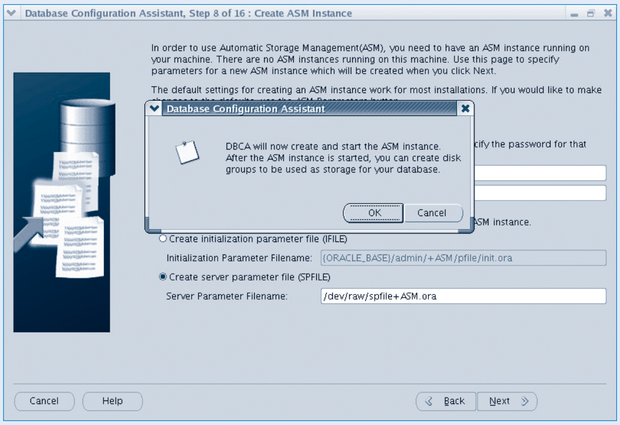
5. To make the binding consistent on each reboot, edit /etc/sysconfig/rawdevices to add these entries:

   ```
   /dev/raw/ASM1        /dev/sdb
   /dev/raw/ASM2        /dev/sdc
   /dev/raw/ASM3        /dev/sdd
   ```

   At this point, three storage devices are available for ASM.

6. Create an Oracle Database 10*g* database using the Oracle Database Configuration Assistant (DBCA). DBCA automatically performs the following operations:

   - If ASM is chosen as the database file-storage method (see Figure A), DBCA automatically creates and starts an ASM management instance (see Figure B).
   - DBCA incorporates and formats the specified storage devices, creates disk groups as specified by the DBA (see Figure C), and prepares them for the Oracle Database 10*g* database.
   - DBCA creates the database using disk groups selected by the DBA as data file storage, as shown in Figure D.



Database Configuration Assistant, Step 8 of 16 : Create ASM Instance

In order to use Automatic Storage Management(ASM), you need to have an ASM instance running on your machine. There are no ASM instances running on this machine. Use this page to specify parameters for a new ASM instance which will be created when you click Next.

The default settings for creating an ASM instance work for most installations. If you would like to make

Database Configuration Assistant

DBCA will now create and start the ASM instance. After the ASM instance is started, you can create disk groups to be used as storage for your database.

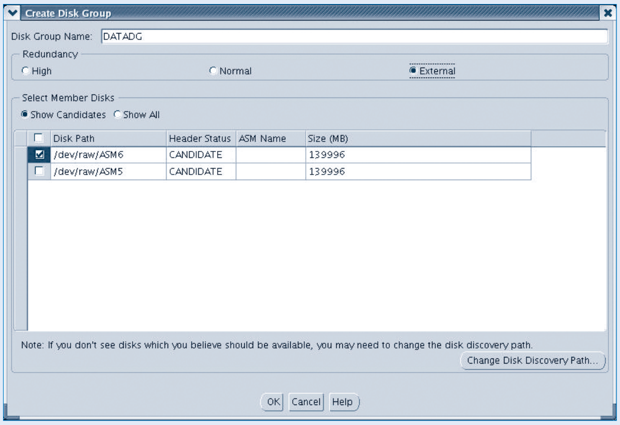
○ Create initialization parameter file (IFILE)

Initialization Parameter Filename: {ORACLE_BASE}/admin/+ASM/pfile/init.ora

● Create server parameter file (SPFILE)

Server Parameter Filename: /dev/raw/spfile+ASM.ora

Figure B. Creating and starting an ASM instance



Disk Group Name: DATADG

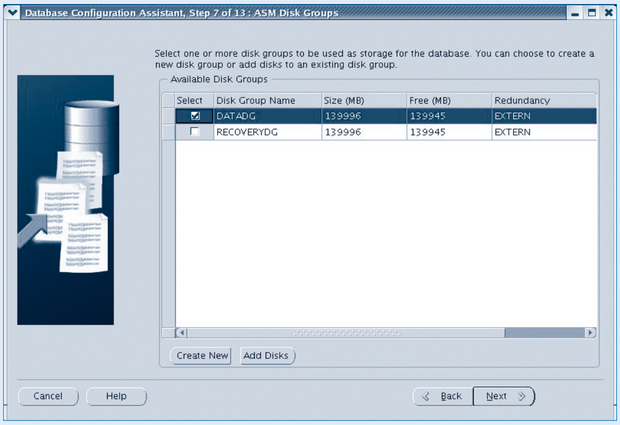
Redundancy: ○ High ○ Normal ● External

Select Member Disks: ● Show Candidates ○ Show All

□	Disk Path	Header Status	ASM Name	Size (MB)
☑	/dev/raw/ASM6	CANDIDATE		139996
□	/dev/raw/ASM5	CANDIDATE		139996

Note: If you don't see disks which you believe should be available, you may need to change the disk discovery path.

Figure C. Selecting member disks of a disk group


Database Configuration Assistant, Step 7 of 13 : ASM Disk Groups
Create New
Add Disks
Cancel
Help
Back
Next

Select one or more disk groups to be used as storage for the database. You can choose to create a new disk group or add disks to an existing disk group.

Available Disk Groups

Select	Disk Group Name	Size (MB)	Free (MB)	Redundancy
☑	DATADG	139996	139945	EXTERN
□	RECOVERYDG	139996	139945	EXTERN

Figure D. Selecting disk groups for database storage

other approaches so that DBAs do not need to make complex provisioning decisions and are freed for other critical tasks.

ASM enables storage management of all Oracle files across thousands of disk drives with minimal management overhead.

Providing efficient management of a large database environment

Building a scalable storage environment for databases has been a challenge for DBAs, especially in a RAC environment in which multiple nodes access the database. The management of database storage using existing storage management techniques, which are detached from the databases, becomes very complicated and impractical when database size grows into the terabyte range.

Oracle Database 10*g* can support databases with sizes ranging into multiple exabytes,[2] because it can support a single data file that contains 4×10^9 Oracle blocks; previously, the limit was 4×10^6 Oracle blocks. To simplify the management of these large databases, Oracle tightly integrated ASM with the database instead of relying on storage management based on the operating system. ASM enables the Oracle database to automatically provision the required storage to the database files as needed from a pool of disk groups that may be expanded, shrunk, or rebalanced for optimal I/O on the fly—freeing DBAs from many of the complex, routine tasks of storage management.

Zafar Mahmood (zafar_mahmood@dell.com) is a software engineer in the Dell Database and Application Engineering Department of the Dell Product Group. He has been involved in database performance optimization, database systems, and database clustering solutions for more than six years. Zafar has an M.S. in Electrical Engineering with a specialization in Computer Communications from the City University of New York.

Joel Borellis (joel.borellis@oracle.com) is a technology director for Oracle, working exclusively with Dell for the past two years. He is in the Oracle Server Technologies Organization, where he works on Oracle and Dell solutions in the areas of development, engineering, and sales and marketing.

Mahmoud Ahmadian (mahmoud_ahmadian@dell.com) is a development engineer in the Dell Database and Application Engineering Department of the Dell Product Group. He has an M.S. in Computer Science from the University of Houston.

Paul Rad (paul_rad@dell.com) is a senior software engineer in the Dell Database and Application Engineering Department of the Dell Product Group. Paul's current interests are in the area of operating systems, database systems, clustering, storage technologies, and virtualization. Paul has master's degrees in both Computer Engineering and Computer Science from The University of Texas at San Antonio.

FOR MORE INFORMATION

Oracle Database 10*g*:
http://www.oracle.com/database

Oracle Automatic Storage Management:
http://otn.oracle.com/obe/obe10gdb/manage/asm/asm.htm

[2] One exabyte = 2^{60} (1,152,921,504,606,846,976) bytes, or 1,024 petabytes.

Running Oracle E-Business Suite

on Dell PowerEdge Servers

Oracle® E-Business Suite 11*i* running on Dell™ PowerEdge™ servers and Dell/EMC storage can scale out to support thousands of concurrent users. This article describes a Dell and Oracle project designed to demonstrate that enterprise applications such as E-Business Suite can run effectively on industry-standard servers.

BY JOHN PAGE

Oracle® E-Business Suite 11*i* is a complete set of applications that can automate business processes such as supply chain management and customer service operations. The suite is widely deployed in several industries, including banking and manufacturing. Oracle E-Business Suite and other enterprise-class software have traditionally been associated with proprietary RISC-based systems rather than commodity servers based on industry standards. This scenario is now changing as more organizations deploy mission-critical software on standards-based servers like the Dell™ PowerEdge™ 2650 and Dell PowerEdge 6650 servers.

However, for standards-based servers to meet the needs of the enterprise, they must be able to support large numbers of concurrently connected users. It is not unusual for some businesses to have thousands of users working on the system at the same time. Therefore, if an enterprise application such as Oracle E-Business Suite is to be deployed on standards-based servers, that application must be capable of scaling out. This means that an organization can run the application on a number of relatively small two- or four-processor servers and add servers in parallel as business requirements grow.

To determine how well Oracle E-Business Suite scales out on Dell PowerEdge servers, Dell and Oracle performed a joint project in August 2003 at the Dell Application Solution Centre (ASC) in Limerick, Ireland. The purpose of the project was to simulate a real-world implementation of Oracle E-Business Suite, and specify the additional hardware components required to scale out.

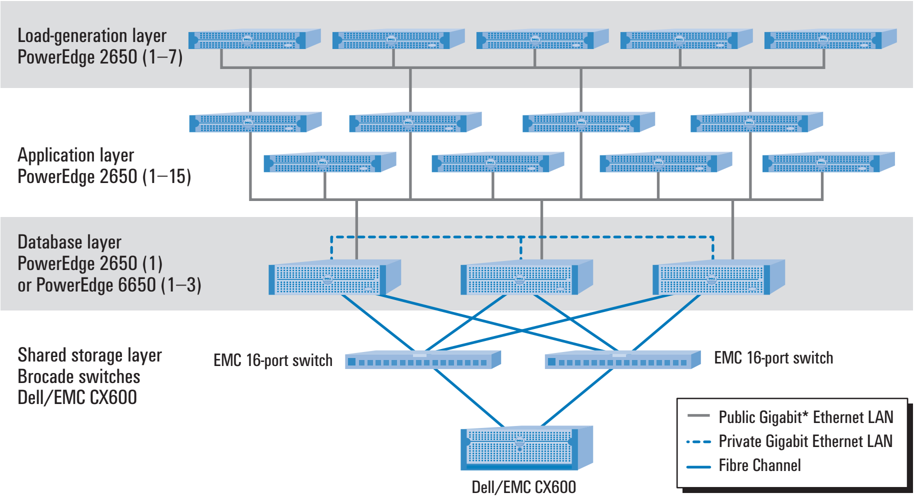
Key components of the test environment

The test environment consisted of four distinct layers: the load-generation, application, database, and storage layers.

Load-generation layer. The Mercury Interactive LoadRunner® version 7.51 tool was used to simulate user activity. The project team loaded this software on Dell PowerEdge 2650 servers running Microsoft® Windows® 2000 Advanced Server Service Pack 3 (SP3). Engineers used between one and seven load-generator systems during the test, depending on the number of users being simulated.

Application layer. The application software was Oracle E-Business Suite 11*i* version 11.5.8. This software was loaded onto PowerEdge 2650 servers running Red Hat® Enterprise Linux® AS 2.1 with kernel 2.4.9e25. The project team used between 1 and 15 application servers during the test—again, depending on the number of users being simulated.

Database layer. The applications that comprise Oracle E-Business Suite 11*i* use the Oracle9*i*™ database for data storage. The database software for the project was Oracle9*i* release 2 (9.2.0.3). Initially, the database was run on a



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

Figure 1. The test environment

PowerEdge 2650 server. As the number of users increased, engineers replaced this server with a PowerEdge 6650. As user numbers increased further, the project team employed up to three PowerEdge 6650 servers, configuring the servers into a cluster using Oracle9*i* Real Application Clusters (RAC) technology. The operating system (OS) for the database servers was Red Hat Enterprise Linux AS 2.1 with kernel 2.4.9e25.

Storage layer. The shared storage resided on a Dell/EMC CX600 storage array, and the database itself was stored on a 70 GB logical storage unit (LUN) that was part of a six-disk RAID-10 set. The database was populated with data from the Oracle Vision data set—a standard set of sample data that Oracle uses for projects such as this one.

Figure 1 shows the test environment. The precise hardware and software specifications for each layer are provided in Figure 2.

	Load-generation layer	Application layer	Database layer	Storage layer
Hardware	PowerEdge 2650	PowerEdge 2650	PowerEdge 2650 and PowerEdge 6650	Dell/EMC CX600 Two Brocade® Fibre Channel switches
Software	Mercury Interactive LoadRunner 7.51	Oracle E-Business Suite 11*i* version 11.5.8	Oracle9*i* release 2 (9.2.0.3) Oracle9*i* RAC	N/A
OS	Microsoft Windows 2000 Advanced Server	Red Hat Enterprise Linux AS 2.1	Red Hat Enterprise Linux AS 2.1	N/A
Processors	Two Intel® Xeon™ processors at 2.8 GHz with 512 KB level 2 (L2) cache, Hyper-Threading Technology enabled	Two Intel Xeon processors at 2.8 GHz with 512 KB L2 cache, Hyper-Threading Technology enabled	Four Intel Xeon processors at 2.0 GHz with 512 KB L2 cache, Hyper-Threading Technology enabled	N/A
Memory	6 GB	6 GB	6 GB (PowerEdge 2650) and 8 GB or 16 GB (PowerEdge 6650)	2 GB per service pack
RAID controller	PowerEdge Expandable RAID Controller 3, Dual Channel integrated (PERC 3/Di)	PERC 3/Di	PERC 3, Dual Channel (DC)	N/A
Disk storage	Two 36 GB RAID-1 (OS) Three 36 GB RAID-5 (Linux swap partition)	Two 36 GB RAID-1 (OS) Three 36 GB RAID-5 (Linux swap partition)	Two 36 GB RAID-1 (OS) Three 36 GB RAID-5 (Linux swap partition)	Six 73 GB RAID-10 (database)
Network interface cards (NICs)	Two Broadcom integrated NICs	Two Broadcom integrated NICs	Two Broadcom integrated NICs	N/A
Storage area network (SAN) connectivity	N/A	N/A	Two QLogic 2340 host bus adapters (HBAs)	N/A

Figure 2. Hardware and software specifications

Description of workload characteristics

The test scripts that simulated user activity were specifically written for this project. The scripts were designed to produce within the application a realistic simulation of typical user activity, helping ensure that the results achieved would characterize real-world situations.

The workload included transactions from five core modules in Oracle E-Business Suite 11*i*: Accounts Receivable (AR), Fixed Assets (FA), Inventory (Inv), Order Entry (OE), and Purchasing (PO).

The workload also included Pricing (QP) transactions that provided cross-functional testing of the OE and Inv modules. The workload mix consisted of 22 business transactions made up of 16 online transaction processing (OLTP) transactions and 6 transactions that submitted concurrent requests (see Figure 3).

The sample Vision database was populated with thousands of rows of data, yielding a total of approximately 50 GB of data. The project scripts represented the activity of 40 different users; each user executed a transaction a particular number of times to produce an overall total number of transactions. This process is described in Figure 3 as follows:

- **Base user count:** The number of users running a given transaction
- **Iterations per user:** The number of times a particular user executes this transaction
- **Transaction total:** The base user count multiplied by iterations per user

As engineers increased the load being applied to the test environment, users were added in groups of 40. Because each group of 40 users had the profile described in Figure 3, the overall characteristic of the workload remained constant regardless of the number of users added to the test. In other words, the same ratio of AR, FA, Inv, OE, PO, and QP transactions was applied regardless of the total number of users.

Data collection and analysis

The project team conducted each test cycle by applying a user load to the configuration under test, measuring the results, increasing the user load, measuring again, and continuing in this manner until the hardware was demonstrably unable to handle the user load. This approach enabled engineers to establish the maximum possible number of concurrent users for each hardware configuration.

Engineers gathered data using the Linux `sar` command and the Oracle statspack utility. Mercury Interactive LoadRunner verified that the run completed with no failures and also gathered the following data during each test:

Transaction		Base user count	Iterations per user	Transaction total
Accounts Receivable (AR)	AR enter customer	2	10	20
	AR enter invoice	3	10	30
	AR enter invoice credit memo	1	11	11
	AR enter invoice adjustment	1	9	9
	AR enter receipt	1	10	10
	AR to GL* transfer request	1	4	4
	AR print invoice submittal	1	4	4
	AR print statement submittal	1	4	4
	AR total	11	62	92
Fixed Assets (FA)	FA assets inquiry	4	10	40
	FA total	4	10	40
Inventory (Inv)	Inv create an item	4	12	48
	Inv view an item	3	11	33
	Inv total	7	23	81
Order Entry (OE)	OE insert an order	3	14	42
	OE view an order	2	13	26
	OE total	5	27	68
Purchasing (PO)	PO create a supplier	1	10	10
	PO create a purchase order	2	12	24
	PO view a purchase order	2	10	20
	PO create a requisition	2	9	18
	PO view a requisition	2	9	18
	PO total	9	50	90
Pricing (QP)	QP price list setup	1	11	11
	QP adjust price list	1	9	9
	QP add items to price list	1	10	10
	QP copy a price list	1	10	10
	QP total	4	40	40
Totals		40	212	411

**General Ledger*

Figure 3. Workload details

- Overall average response time
- The 90th percentile response time
- Total number of transactions completed

The 90th percentile response time was the time by which 90 percent of all transactions had completed. If nine out of ten transactions completed within one second, and the tenth took eight seconds, the 90th percentile response time would still be one second.

At the end of each run, the project team analyzed the data and decided whether the workload had passed or failed by establishing whether the hardware had been able to support the workload. Before a workload was deemed to have passed, it had to meet the following criteria; if a workload failed on any one of these criteria, it was deemed to have failed overall:

- The 90th percentile response time did not exceed two seconds
- Overall average response time was less than the 90th percentile response time (to ensure that the values of the excluded 10 percent of transactions were not excessive)
- All business transactions completed successfully

Test cycles and results

The tests were carried out in seven distinct cycles. The goal of each cycle was to establish the maximum number of users that could be supported for one particular hardware configuration. For example, in cycle 2, the project team set out to establish how many users could be supported when the application software and the database software were each running on a separate PowerEdge 2650. In subsequent cycles, the project team added progressively more hardware until, in cycle 7, the configuration had reached 15 application servers, seven load generators, and three database servers. The maximum number of users supported and the hardware configuration tested for each cycle are shown in Figure 4. Note that the maximum user counts are always divisible by 40 because users were added in groups of 40 as the workload increased.

Figure 4 shows that the average response times and the 90th percentile response times for all cycles are comparable with the response times measured from a 40-user workload. (This 40-user workload is described as the *reference workload* in Figure 4.) For example, cycle 7 processed almost a hundred times as many transactions as the reference workload. Despite this, the response times were only marginally greater and were well within the acceptable parameters that defined a successful test.

Of all the cycles, cycle 2 was of particular importance because it established the maximum number of Oracle E-Business Suite users that could be supported on a single PowerEdge 2650 server.

	Topology			Results			
Cycle number	PowerEdge server(s) used in database layer	PowerEdge server(s) used in application layer	Maximum users	70 percent of maximum users	Transactions per hour	Overall average response time	90th percentile response time
Reference workload	One PowerEdge 2650 server	N/A	40	28	280	0.47	0.52
1	One PowerEdge 2650 server	N/A	160	112	1,138	0.58	0.73
2	One PowerEdge 2650 server	One PowerEdge 2650 server	280	200	2,024	0.50	0.63
3	One PowerEdge 2650 server	Two PowerEdge 2650 servers	560	400	3,790	0.61	0.84
4	One PowerEdge 6650 server	Three PowerEdge 2650 servers	840	600	5,268	0.45	0.51
5	One PowerEdge 6650 server	Five PowerEdge 2650 servers	1,400	1,000	9,778	0.47	0.56
6	Two PowerEdge 6650 servers	Ten PowerEdge 2650 servers	2,800	2,000	19,168	0.56	0.69
7	Three PowerEdge 6650 servers	Fifteen PowerEdge 2650 servers	3,920	2,750	26,910	0.66	0.83

Figure 4. Overall results

This enabled engineers to determine the minimum number of servers required in the application layer to support a particular number of users, apart from any database considerations. At the end of cycle 2, the test team had identified that a single PowerEdge 2650 server could support a maximum of 280 concurrent users.

Figure 4 contains two user columns: maximum users and 70 percent of maximum users. The maximum users for each cycle is precisely that: the absolute maximum number of users that a particular configuration will support. However, the project team's recommendation is to allow some leeway and consider 70 percent of the maximum as the number of users that a particular configuration will support.

With this guideline in mind, the following are key findings from the Dell and Oracle joint project:

- In cycle 1, a single PowerEdge 2650 server running both E-Business Suite 11*i* and the Oracle9*i* database supported a maximum of 160 users during the test.
- In cycle 2, a single PowerEdge 2650 server running E-Business Suite 11*i* supported a maximum of 280 users when the database was running on another server.
- In cycle 3, a single PowerEdge 2650 server running the Oracle9*i* database supported a total of two application servers with 280 users each, giving a maximum of 560 users for a configuration that uses a PowerEdge 2650 with 6 GB of memory to run the database software.
- In cycle 4, by switching the database to a PowerEdge 6650 server with 8 GB of memory instead of a PowerEdge 2650 server, the maximum number of supported users rose to 840 using three application servers supporting 280 users each.
- In cycle 5, by increasing the memory in the PowerEdge 6650 server from 8 GB to 16 GB, a maximum of 1,400 users was supported using five application servers supporting 280 users each.
- In cycle 6, by adding a second PowerEdge 6650 server into a RAC cluster along with the first PowerEdge 6650 server, the maximum number of supported users doubled to 2,800—a scalability factor of 1.0.
- In cycle 7, by adding a third PowerEdge 6650 server to the RAC cluster, the maximum number of supported users increased to 3,920—a scalability factor of 0.8.
- The user counts reported for all cycles are maximum user counts. Dell recommends using the “70 percent” column in Figure 4 as the number of users actually supported. This will help guarantee that a particular configuration will be able to support transient increases in load without the need to purchase additional hardware.

Scale-out capabilities for Oracle E-Business Suite on Dell hardware

The results of this Dell and Oracle project strongly support a scale-out strategy for Oracle E-Business Suite 11*i* deployed on standards-based servers. Having established that a single PowerEdge 2650 server can support a maximum of 280 concurrent users in the application layer, the test team was easily able to support more users simply by adding more servers. Ten servers in the application layer supported a maximum of 2,800 users. Beyond this, the scalability factor dropped off somewhat, but engineers were still able to support a maximum of 3,920 users with 15 application servers and three database nodes deployed on the back end. These findings indicate that an enterprise application that had previously run on large, proprietary servers could now run extremely well on several smaller, industry-standard servers such as Dell PowerEdge servers.

John Page (john_page@dell.com) is a systems consultant at the Dell Application Solution Centre in Limerick, Ireland. He has an M.S. in Engineering Science and a B.A. in Electrical Engineering from University College Cork and is a Microsoft Certified Systems Administrator (MCSA).

Building Distributed Microsoft SQL Server 2000 Database Applications on Dell PowerEdge 6650 Servers

Building out data centers using many smaller servers based on four or fewer CPUs rather than a few large servers based on eight or more CPUs can offer cost, redundancy, and ease-of-expansion advantages. To demonstrate how Microsoft® SQL Server™ software can benefit from running on multiple servers, a team of Dell engineers built a 100 GB online store application to run on two Dell™ PowerEdge™ 6650 servers, each with four processors—one system received orders and the other processed financial reports against the order data. This article describes the online store application and the data replication features of SQL Server that connected the database instances on each server.

BY DAVE JAFFE, PH.D.; TODD MUIRHEAD; AND WILL CLAXTON

Microsoft® SQL Server™ 2000 relational database management software is widely deployed in worldwide enterprises for online transaction processing (OLTP), online analytical processing, and data mining. Highly scalable, reliable, easy to deploy, and self-tuning, SQL Server is used for demanding, mission-critical applications. Whether used as the database engine behind such Microsoft products as Commerce Server and Content Management Server, or accessed by custom applications created with the Microsoft Visual Studio® .NET development system, SQL Server can provide a robust environment to manage corporate data.

Dell best practices advocate building enterprise data centers by sharing the workload among farms of small servers that have four or fewer CPUs, rather than concentrating the workload on larger servers with eight or more CPUs. *Scaling out* the data center using smaller servers can offer cost, fault-tolerance, and ease-of-expansion advantages over larger, *scaled-up* servers. SQL Server applications can be designed to run on multiple servers through the use of high-speed data replication technology.[1]

[1] For more information on Microsoft SQL Server replication technology, see “SQL Server 2000 Replication Overview” at http://www.microsoft.com/sql/evaluation/features/replication.asp.

In December 2003, a team of Dell engineers used SQL Server replication technology to build a 100 GB online DVD store application that ran across two Dell™ PowerEdge™ 6650 servers, each with four processors. One PowerEdge 6650 server, which was driven by a Web application (not modeled in this test), collected incoming orders. A second PowerEdge 6650 server generated financial reports from the order data. On a scheduled basis, the updated order data from the first server was replicated to the second server. By running the order-entry and report workloads on separate servers, with each server tuned for its particular workload, Dell engineers achieved performance scaling and redundancy.[2] This article describes the SQL Server test performed by the Dell team, including factors such as the database application, the replication method used to update the second server, and the performance impact of the replication.

Building the database application

To demonstrate SQL Server running on multiple instances, a 100 GB online DVD store was implemented as two replicated SQL Server databases. One SQL Server instance handled the entry of new orders and replicated changes on a scheduled basis to the second SQL Server instance, which was used for generating financial reports. The DVD store (DS) database comprised a set of data tables organized according to a certain schema, along with a set of stored procedures that did the actual work of managing the data in the database as orders were entered and reports were requested. The database back end was designed to be driven from a Web-based middle tier, but because the focus of the Dell test was on the database servers, the back-end stored procedures were driven directly by custom C programs to simulate a Web-based middle tier.

The database schema

The DVD store was composed of four main tables and one additional table (see Figure 1). The Customers table was prepopulated with 200 million customers, including 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, starting in January 2003 and ending in September 2003. 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. For the full DVD store database build script, visit *Dell Power Solutions* online at http://www.dell.com/magazines_extras.

The stored procedures

The DVD store database was managed through seven stored procedures. The first two were used during the login phase. For returning customers, the Login procedure retrieved the customer's information, in particular the CUSTOMERID. For new customers, the New_customer procedure created a new row in the Customers table containing the customer's data.

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 1. Database schema for online DVD store

Following the login phase, the customer might search for a DVD by category, actor, or title. These database functions were implemented by the Browse_by_category, Browse_by_actor, and Browse_by_title procedures, respectively. Finally, after the customer completed the selections, the Purchase procedure was called to complete the transaction. Additionally, the Rollup_by_category procedure calculated total sales by DVD category for the previous month, quarter, and half-year periods. For the stored procedures, visit *Dell Power Solutions* online at http://www.dell.com/magazines_extras.

The driver applications

Separate multithreaded driver programs were written to model the OLTP order-entry workload and the report request workload.

Online transaction processing. 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 customer think times or key times were factored in, the database connections remained full, simulating what happens in a real multitiered application—a few connections are pooled and shared among 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.

Each thread of the OLTP driver 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 driver measured order rates and the average response time to complete each order. Several tunable parameters were

[2] This is just one example of spreading a SQL Server workload across multiple servers. For an overview of different scale-out choices, see *The Definitive Guide to Scaling Out SQL Server* by Don Jones at http://www.dell.com/sql/ebook.

used to control the application, as described in Figure 2.

Reports. The report request driver program was similar to the OLTP driver in that each thread connected to the database and started making stored procedure calls. Each thread made repeated calls to the Rollup_by_category stored procedure, which calculated total sales by DVD category for the previous month, quarter, and half year, until every report for all 16 categories was completed. In each test, eight simultaneous reports were run.

Replicating the SQL Server databases

Microsoft SQL Server 2000 provides several configuration options when setting up replication, but they all use the same basic model of publishers, distributors, and subscribers. The test environment simulated the requirement to run reports and accept new transactions against the same database. Replication allowed two synchronized copies of the same data. Through the use of two PowerEdge systems running SQL Server, connected by a replication mechanism, all new transactions occurred in one instance of SQL Server and the reports were run on the other instance.

Replication types and terminology

SQL Server uses the model of publishers, distributors, and subscribers to replicate data, and it provides a set of wizards to help administrators set up each of these components. The *publisher* database is the source for data to be replicated. The *distributor* service pushes out the data from the publisher for replication. The *subscriber* database receives the data from the distributor. Multiple subscribers can exist, allowing for multiple replicated copies of the same data. In the test environment described in this article, the server accepting the new transactions was configured as the publisher and the distributor of all tables and stored procedures for the DVD store database.

Parameter	Description	Value(s) used in test
n_threads	Number of simultaneous connections to the database	1
warmup_time	Warm-up time before statistics are kept	1 minute
run_time	Runtime during which statistics are kept	Varied
pct_returning	Percent of customers who are returning	95 percent
pct_new	Percent of customers who are new	5 percent
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 2. OLTP driver parameters

Scaling out the data center using smaller servers can offer cost, fault-tolerance, and ease-of-expansion advantages over larger, scaled-up servers.

Additionally, SQL Server offers different types of replication, which are defined by how often the updates to the database are published to the subscribing database and how those updates are sent. Using *snapshot publication* for SQL Server replication means that a complete and updated snapshot of the data is periodically sent to the subscriber. A *merge publication* means that changes made to both the publisher and subscriber copies of the database are merged periodically. *Transactional publication* allows for data from the publisher to be sent out as incremental changes on a scheduled basis.

Configuration for replicating the DVD store

SQL Server 2000 was installed on two Dell PowerEdge 6650 servers. Then, the initial DVD store database was loaded with the same data on both systems. The objective was to run reports on one node and accept new transactions on the other node. To achieve this objective, the Dell engineers set up transactional replication whereby the publisher database accepted new orders and replicated updates nightly to the subscriber database, which ran reports. By moving the report generation to a second server, the test team ensured that running the reports would have no effect on the active transactions occurring on the publisher database. An option was specified during the initial setup of the transactional replication that the subscriber database already had the data and schema. Specifying this option allowed the replication to begin with new orders because both databases had already been pre-loaded with exactly the same data to begin testing.

A daily transactional replication between the two SQL Server instances was set for 12:05 A.M. so that all new transactions from the day before would be replicated to the reports server. This replication allowed orders and reports to be scaled across the two servers with minimal impact on the performance of either server.

Observing SQL Server in action

To demonstrate how SQL Server replication can be used to run the online DVD store application on two servers, Dell engineers started a simulated workload on the SQL_orders server of approximately 233 orders per minute (about 10 million per month) using the order-entry driver. After accumulating roughly one day's worth of orders (about 335,000), with orders still coming in at the same rate, the test team manually initiated replication from the SQL_orders server (the publisher) to the SQL_reports server (the subscriber), thereby simulating the daily replication.

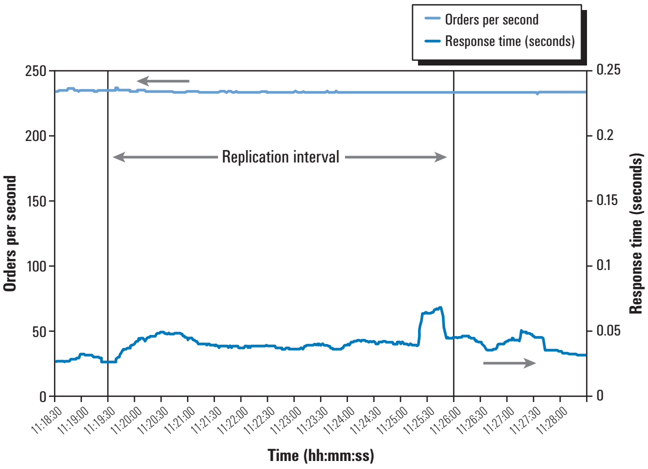


Figure 3. Performance of one Dell PowerEdge 6650 server during SQL Server replication

Figure 3 shows the results of the SQL Server replication. The response time—the total response time for all phases of the order, including login, browse, and purchase—rose slightly during the replication period but still averaged under 0.1 second, with a few individual orders as high as 0.5 second. The order rate was essentially flat. Thus, the replication occurred in real time, took about seven minutes, and had very little effect on the performance of the order-entry system as experienced by the customer.

Scaling out the enterprise with replication technology

Data replication is one method among many for building a Microsoft SQL Server database application across multiple servers. Using an online DVD store application, the Dell test team demonstrated how to replicate SQL Server data, with orders being received on one Dell PowerEdge 6650 server and financial reports being generated by a second PowerEdge 6650 server that had a copy of the same data. New orders were replicated nightly from the order-entry server to the reports server. The test results showed that replication had a minimal impact on the ability of the order-entry server to receive orders, while continuing to accept orders at a rate of 10 million per month.

Dave Jaffe, Ph.D. (dave_jaffe@dell.com) 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.

Todd Muirhead (todd_muirhead@dell.com) is an engineering consultant on the Dell Technology Showcase team. He specializes in storage area networks 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.

Will Claxton (will_claxton@dell.com) is the Dell alliance manager for enterprise applications, focusing on oversight and strategic direction for Microsoft applications. He has a bachelor's degree from Texas A&M University and a graduate degree from The University of Texas at Austin.

FOR MORE INFORMATION

Dell and Microsoft SQL Server 2000:
http://www1.us.dell.com/content/topics/global.aspx/alliances/en/microsoft_sql?c=us&cs=555&l=en&s=biz

Building a Highly Scalable Database Platform

Using Microsoft SQL Server 2000 Enterprise Edition (64-bit)

Microsoft® SQL Server™ 2000 Enterprise Edition (64-bit) is designed to offer improvements in memory availability and parallel-processing performance compared with 32-bit SQL Server software. This article describes the capabilities of Microsoft SQL Server 2000 Enterprise Edition in the 64-bit Intel® architecture environment—highlighting differences from the 32-bit environment and discussing some of the applications and usage scenarios that can benefit from a SQL Server platform optimized for the 64-bit environment. Additionally, guidance is provided on identifying potential applications and setting appropriate expectations for deployment and performance.

BY WILL CLAXTON

To make timely and informed business decisions in dynamic environments, organizations store and analyze massive amounts of business data. Always a critical part of the IT infrastructure, the database is at the center of three converging IT trends: growth in application size and complexity, development of high-end database management system (DBMS) capabilities, and consolidation of data center servers.

Because of their size, some applications may be reaching the limits of the 32-bit platform, specifically regarding number of processors and addressable memory. At the same time, high-end DBMS capabilities have become a business necessity as organizations gather and analyze data from numerous databases, and serve that data to growing numbers of business users. Further, many organizations are moving to consolidate servers to simplify critical data center operations. Consolidation can reduce management complexity and cost while reducing physical space requirements in the data center.

In response to these trends, organizations can take three general paths:

- **Scale up without 64-bit Intel architecture (IA-64) migration:** Upgrade the existing 32-bit server environment to include new and more powerful Intel® Xeon™ processors with expanded cache capability. Upgrading 32-bit servers also enables administrators to deploy storage with improved disk performance, which can help alleviate the disk I/O bottleneck in database applications.

- **Scale up with IA-64 migration:** Replace 32-bit servers with 64-bit servers designed to provide applications with greater memory addressability and storage capability.
- **Scale out:** Restructure the database using a variety of methods to provide the required performance.[1]

This article explores the reasons why and circumstances under which organizations may consider 64-bit database technology using Microsoft® SQL Server™ 2000 Enterprise Edition software as well as how IA-64 architecture can help remove certain performance bottlenecks cost-effectively.

IA-64 components in SQL Server 2000

Microsoft SQL Server 2000 Enterprise Edition (64-bit) components include:

- **Database server:** Core database functionality
- **Server agent:** Alerts and management
- **Analysis server:** Online analytical processing (OLAP) and data mining

The preceding IA-64 components are code-compatible with the 32-bit version of SQL Server 2000, allowing administrators to integrate an IA-64 server with other SQL Server 2000 database servers.

Capabilities of IA-64 architecture

The enhanced scalability and performance capabilities of SQL Server 2000 (64-bit) is enabled by several features of the 64-bit Intel Itanium® architecture. Key benefits are as follows.

Memory addressability. Generally, 32-bit systems can address only a 4 GB address space. See the section "Comparison of SQL Server 2000 (32-bit) with extension technology versus 64-bit IA-64 architecture" in this article for more information about options available with Address Windowing Extensions (AWE) and Physical Address Extension (PAE) on 32-bit platforms. The Microsoft Windows Server™ 2003 operating system running on Intel Itanium IA-64 architecture supports up to 1,024 TB of physical memory and 512 GB of addressable memory.

Parallel-processing support. Intel Itanium chips include several features that are designed to enhance parallel-processing performance compared to 32-bit Intel chips.[2] The Intel Itanium 2 chip offers a wider system bus, more registers,[3] and Explicitly Parallel Instruction Computing (EPIC) technology, which is designed to enable a processor to execute up to six instructions simultaneously. Such performance improvements in parallel processing can help enhance many SQL Server 2000 parallel operations, including parallel query resolutions, index builds, backup and restore operations, data loads, and maintenance operations.

Because of their size, some applications may be reaching the limits of the 32-bit platform, specifically regarding number of processors and addressable memory.

Enhanced bus architecture. The bus architecture on current IA-64 chip sets can provide greater throughput than in 32-bit environments.[4] More data can be passed to the cache and processors quickly—an enhancement somewhat analogous to the improvement that broadband offers over dial-up connections.

Comparison of SQL Server 2000 (32-bit) with extension technology versus 64-bit IA-64 architecture

Microsoft SQL Server 2000 Enterprise Edition (32-bit) uses the Microsoft Windows® 2000 Server AWE and PAE application programming interfaces (APIs) to support very large amounts of physical memory in applications and the operating system, respectively. For some applications, using AWE to enhance SQL Server 2000 (32-bit) may be a viable alternative to upgrading to the IA-64 platform.

Standard 32-bit systems can map 4 GB of memory at most, limiting the addressable memory space for Windows 2000 systems to 4 GB. With 2 GB reserved for the operating system, only 2 GB of memory remain for the application—in this case, SQL Server 2000. Administrators can increase the amount of addressable application memory to 3 GB by setting the `/3GB` switch in the Windows boot.ini file. However, slight performance degradations may occur when using AWE and PAE in this way, and not all aspects of SQL Server 2000 (and many other applications) can take advantage of these extensions.

In contrast to the 32-bit limitations, SQL Server 2000 (64-bit) makes extended memory available to all database processes and operations. When using the 64-bit version of SQL Server 2000 on Itanium 2–based hardware like the Dell™ PowerEdge™ 7250 server, a SQL Server instance can address up to 512 GB,[5] which is the current maximum memory supported by Windows Server 2003. (The theoretical addressable limit is 18 exabytes.[6]) This memory is available to all components of SQL Server and to all operations within

[1] For more details on the scale-out option, see *The Definitive Guide to Scaling Out SQL Server* by Don Jones at http://www.dell.com/sql/ebook.

[2] For more information, visit http://www.microsoft.com/technet/prodtechnol/sql/2000/evaluate/64btdwc4.mspx.

[3, 5] For more information, visit http://www.microsoft.com/sql/64bit/productinfo/overview.asp.

[4] For more information, visit http://www.intel.com/ebusiness/pdf/prod/itanium/wp022404.pdf, http://www.intel.com/design/chipsets/e8870/index.htm, and http://www.intel.com/products/server/processors/server/xeon/index.htm?iid=ipp_browse+featureprocess_xeon&.

[6] One exabyte = 2^{60} (1,152,921,504,606,846,976) bytes, or 1,024 petabytes.

the database engine. As a result, SQL Server 2000 (64-bit) is designed to enhance performance of a wide range of memory-intensive database applications.

Usage scenarios for SQL Server 2000 (64-bit)

Although SQL Server 2000 (64-bit) is designed to offer significant scalability and performance for many applications—including SAP®, PeopleSoft®, Siebel®, and other applications that require frequent disk caching—not every application will benefit from the 64-bit version of SQL Server 2000. This section is designed to help organizations determine whether it is more appropriate to use existing SQL Server instances or develop new applications based on the IA-64 architecture.

The improved memory and parallel-processing capabilities of SQL Server 2000 (64-bit) compared with SQL Server 2000 (32-bit) are compelling in several usage scenarios, including:

- **Enhancing performance for memory-constrained relational database applications:** Helping to alleviate memory constraints allows a larger percentage of the database—or possibly the entire database—to reside in memory.
- **Creating or accelerating large OLAP systems with rapid response-time requirements:** Fast databases provide decision makers with quick access to simplified views of complex data.
- **Consolidating multiple Windows-based databases and applications onto fewer, larger systems:** By hosting multiple databases on a single IA-64 platform, organizations can simplify management, improve storage utilization, and generally improve operational efficiency.
- **Scaling up current applications that are experiencing significant growth:** Migrating existing database servers that are outgrowing their current platform does not affect the other tiers of multi-tiered applications.
- **Replacing UNIX systems and applications:** The IA-64 platform offers a powerful alternative to UNIX® systems for high-end database servers.

The following section discusses factors to consider when evaluating specific applications for the IA-64 platform.

Relational database performance factors

Memory-intensive SQL Server relational database workloads are good candidates for SQL Server 2000 (64-bit). Many SQL Server resources are restricted to a 3 GB limit in the 32-bit environment, resulting in systems that are starved for memory. Such systems degrade performance because applications wait for resources and experience delays while processors compile stored procedures that have been evicted from cache. These systems may also experience excess disk activity when writing objects such as hash tables—which cannot fit into the available memory—to disk. Moving to SQL Server 2000 (64-bit) can help improve the performance of applications experiencing the following memory-related performance problems:

- **Recompilation of stored procedures that were evicted from memory:** The IA-64 environment provides a large plan cache for high-volume transaction applications using large numbers of stored procedures. This helps reduce the need to compile stored procedures that have been evicted from memory—reducing CPU utilization and reducing query latency.
- **Resource semaphore waits associated with queries awaiting memory grants:** Multiple queries utilizing large-scale hash joins—especially those executing against a data warehouse and spanning large data sets—can be adversely affected by resource semaphore waits.

Other relational database operations that can benefit from extended memory include:

- **Index creation:** In a 32-bit environment, index creation, including full-text indexing, is restricted to a 3 GB workspace.
- **Complex queries:** Operations that use sorting, large hash joins, or hash aggregates to construct complex queries can benefit from extended memory. Memory-intensive hash joins are very efficient, but when memory is under pressure, these queries may be removed from cache in favor of slower memory-conserving query plans.
- **Active stored procedures:** Benefits can be achieved for large numbers of active stored procedures through improved cache capacity. The IA-64 architecture can help substantially reduce overall CPU utilization and latency by helping to eliminate the need to evict procedures from cache and compile repeatedly.
- **Server cursors:** As database objects that applications can use to manipulate data sets, server cursors can more readily be kept in memory, helping to improve performance because memory access is faster than hard disk access.

Many I/O-intensive applications can potentially benefit from loading a larger working data set into memory, which is possible in the IA-64 environment. Although applications requiring more memory than the 64 GB limit supported by AWE are rare, certain applications or workloads can benefit from the speed of extracting database pages from extremely large cache memory instead of the disk subsystem.

Additionally, the improved in-processor parallelism capabilities of IA-64 chip sets as compared to 32-bit chips benefits SQL Server 2000 (64-bit) in situations where context switching degrades performance. Administrators can configure SQL Server 2000 to use fibers instead of threads for more efficient parallel operations. Using fibers is helpful in three cases: when most CPUs consistently run at or near capacity, when the application executes across many CPUs, and when a high level of context switching occurs.

Analysis Services considerations

SQL Server 2000 Analysis Services requires that all dimensions for all *cubes*—OLAP data structures that offer richer information by providing a 3-D view of a database—be held in memory simultaneously. This is true in both 32- and 64-bit environments; however, 64-bit environments can hold much larger cubes in memory. Because Analysis Services cannot take advantage of the memory extensions of AWE, its memory is limited to 3 GB in a 32-bit environment, even if more memory is actually available.

The additional memory available to IA-64 platforms gives Analysis Services the capability to support very large dimensions or numerous large dimensions. Consider a 64-bit environment for OLAP applications that require:

- Very large dimensions—SQL Server 2000 (64-bit) has demonstrated support for dimensions consisting of more than 50 million members[7]
- A large number of sizeable dimensions
- Large memory use of process buffers
- Very large cubes—significant performance benefits for very large cubes can be achieved through the use of the file-system cache, helping to reduce the need for physical disk access for base-cube or aggregate data during queries; this can benefit Analysis Services deployments that have extensive fact partitions and a large number of aggregates, even if the dimensions themselves fit into the memory limits of the 32-bit environment
- Fast cube-processing requirements—memory availability helps reduce the need for writing to temporary files on the disk subsystem
- A large number of concurrent users

Before adopting an IA-64 environment for data analysis, administrators must ensure that the following components are available:

- **OLE DB providers:** IA-64 object linking and embedding database (OLE DB) providers are necessary for all data sources used to populate a database in the IA-64 environment. SQL Server 2000 (64-bit) includes an OLE DB provider for accessing SQL Server. When using other data sources from other database vendors, administrators should verify the availability of a suitable OLE DB. Alternatively, administrators can configure a 32-bit server utilizing Data Transformation Services (DTS) to pump data from other OLE DB sources to the IA-64 SQL target.
- **User-defined functions (UDFs):** Analysis Services UDFs or other components written in the Microsoft Visual Basic® 6 development system are not supported in the IA-64 environment. Administrators must verify whether these components exist and, if necessary, rewrite them using C++ and compile them using the IA-64 software development kit (SDK).

Alternatives to IA-64 architecture

Although an IA-64-based system can offer significant performance increases, it is unrealistic to assume that a 64-bit system will automatically double the performance of a 32-bit system. In some situations, upgrading to IA-64 may not be the best alternative. Instead, for example, administrators may achieve a greater improvement in application performance by increasing the number of CPUs or the speed of the CPUs.

For many SQL Server workloads, the 2 to 3 GB of virtual address space available with a 32-bit platform is sufficient. If a workload performs well using 2 to 3 GB of memory (or when using AWE) and does not require scaling up beyond four sockets, the benefits of fast CPUs in a 32-bit architecture may outweigh the advantages of moving to an IA-64 platform.

Many operations within SQL Server 2000 can benefit from the fast CPU speeds currently available on the modern Intel Xeon architectures, which now have clock speeds exceeding 3 GHz, than from the memory and scalability benefits of the IA-64 environment, which is currently limited to 1.5 GHz. Examples include single-threaded query response times on systems that are not busy—or aggregations, hash joins, string comparisons, and other common operations that have adequate memory on 32-bit systems and reside comfortably on two- to four-socket servers. These applications may be better served by using the fastest 32-bit chip sets, and administrators should be aware of the CPU speed trade-off in such circumstances.

Many operations within SQL Server 2000 can benefit from the fast CPU speeds currently available on the modern Intel Xeon architectures.

Administrators must also consider the implications of using very large amounts of memory on the IA-64 platform. For example, a system configured with very large amounts of memory could take a long time to shut down because the system checkpoint must flush a significant amount of data. As a result,

[7] For more information, visit http://www.microsoft.com/technet/prodtechnol/sql/2000/evaluate/64btdwc4.mspx.

IT staff may choose to perform system checkpoints at small intervals, such as every minute. This can be achieved through the recovery-interval server configuration option.

Limitations of SQL Server 2000 (64-bit)

When deciding on an IA-64 platform and SQL Server 2000 (64-bit), administrators should also evaluate potential platform limitations. At press time, they were as follows:

- The IA-64 platform is still a relatively new architecture.
- SQL Server tools such as Enterprise Manager, Query Analyzer, and SQL Profiler are not yet unavailable on IA-64. Best practice is to run these from 32-bit SQL clients.
- No support exists for the execution of DTS packages. DTS packages can be saved on an IA-64 instance but not executed on an IA-64 instance. Best practice is to provide the 32-bit server hosting DTS with a high-speed connection (Gigabit[8] Ethernet) to the IA-64 server that is the source or target of the transformations in the DTS packages, as well as to other data sources.
- No Microsoft-provided Oracle® database or IBM® DB2® drivers for the IA-64 platform exist, and third-party database vendor support is limited for drivers on IA-64-based Windows platforms. This restricts the capability to define linked-server connections from IA-64 instances to non-Microsoft databases.
- Microsoft Operations Manager (MOM) is not currently supported on the IA-64 platform, and availability of third-party systems management tools is limited. Best practice is to capture the performance counters and events generated by Windows and SQL Server 2000 (64-bit) over the network to a 32-bit instance, and use the tools that support a 32-bit instance.
- Device drivers for I/O, storage area networks (SANs), and other components in the environment may not yet be available for an IA-64 platform.
- The Microsoft .NET framework for IA-64 is not yet available.

Advantages of SQL Server 2000 (64-bit)

SQL Server 2000 (64-bit) addresses the need to provide a highly scalable database platform for memory intensive, performance-critical applications. The IA-64 version of SQL Server 2000 provides massively scalable performance for large, complex queries through:

- Large memory addressing
- Nearly unlimited virtual memory
- Support for up to 64 processors in symmetric multiprocessing (SMP) systems
- Enhanced parallelism

While many SQL Server workloads perform well in a 32-bit environment, the additional memory and processors available to the IA-64 environment are valuable in several situations, including:

- Scale-up scenarios requiring 16-processor or larger SMP servers
- Workloads with large-scale sorting, hash joins, and query memory such as complex relational data warehouse queries
- Analysis Services applications with very large dimensions or large volumes of data that can leverage file system cache
- Other applications that may be memory-constrained in a 32-bit environment

SQL Server 2000 (64-bit) is optimized for the Intel Itanium 2 processor, and uses advanced memory addressing capabilities for essential resources such as buffer pools, caches, and sort heaps—helping reduce the need to perform multiple I/O operations to bring data in and out of memory from disk. Great processing capacity without the penalties of I/O latency can provide a mechanism to achieve new levels of application scalability. Using Itanium 2–based servers with large amounts of memory, SQL Server 2000 (64-bit) is designed to load and process multigigabyte databases in significantly less time than that required in a 32-bit environment.

Moreover, the 64-bit version of SQL Server 2000 can achieve performance and scalability gains while maintaining integration with existing products and applications and offering a simple migration path. SQL Server 2000 (64-bit) can integrate easily into a database server cluster with 32-bit databases. Organizations can use 32-bit application servers connecting to IA-64-based database servers, phasing in the IA-64 technology as required.

Will Claxton (will_claxton@dell.com) is the Dell alliance manager for enterprise applications, focusing on oversight and strategic direction for Microsoft applications. Will has a bachelor's degree from Texas A&M University and a graduate degree from The University of Texas at Austin.

FOR MORE INFORMATION

SQL Server 2000 Enterprise Edition (64-bit):
http://www.microsoft.com/sql/64bit/default.asp

[8] 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.

Deploying Dell Update Packages

for Red Hat Enterprise Linux

Dell™ Update Packages for the Red Hat® Enterprise Linux® AS operating system can simplify installation of Dell PowerEdge™ servers and help enterprises lower administrative costs.

BY BALA BEDDHANNAN, JOHN BRUNKEN, AND JASON D. NORMAN

Managing system firmware on production servers can be troublesome for a busy system administrator or IT staff. If the number of managed nodes is large, as with a cluster, or a server has high uptime requirements because of its role in the enterprise, the time required to perform updates may hinder their execution. For Dell™ PowerEdge™ servers running the Red Hat® Enterprise Linux® AS 2.1 operating system, updating system firmware has been simplified with the release of Dell Update Packages for Red Hat Linux. Dell Update Packages are self-contained system firmware updates that provide their own inventory, validation, and application logic. They may be used for single-instance execution or for mass-deployment through various software distribution utilities. This article examines the features provided by version 1.1 Dell Update Packages and explores possibilities for their use. These features include interactive mode and non-interactive mode; tools available within the Red Hat Linux distribution; and distributed shell, a third-party GNU General Public License (GPL) tool for updating servers sequentially or simultaneously.

Deploying Dell Update Packages in Linux environments

Version 1.1 Dell Update Packages support the PowerEdge servers and system firmware listed in Figures 1 and 2. Although the number of supported Linux kernels is narrow in scope, Dell Update Packages are designed to provide the flexibility to support additional kernels through a command-line option. See the *Dell Update Packages User's Guide* at http://support.dell.com for details.

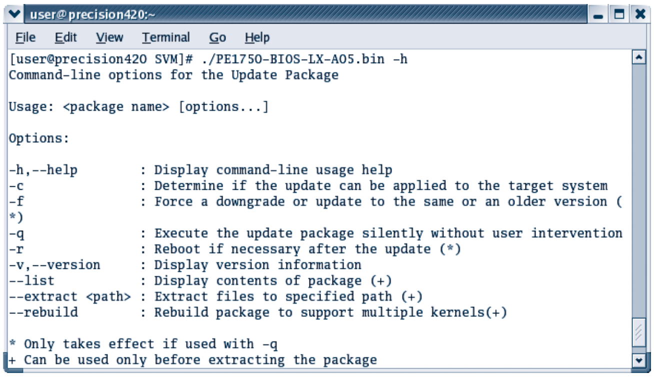
Red Hat Linux is a rich environment for system administration, providing a wide variety of management methods. Among these, remote shells help allow administrators to interactively and securely perform management activities on an unlimited number of systems from the convenience of a single workstation. Shell scripting provides a flexible and reliable means of executing commands, allowing system updates to be scheduled or triggered by a specific event. Scripts written in a native or portable scripting language, such as

		Embedded Server Management (ESM)	
Model	BIOS	ESM3	ESM4 Backplane
PowerEdge 1650	Yes	Yes	No
PowerEdge 1750	Yes	Yes	No
PowerEdge 2600	Yes	Yes	No
PowerEdge 2650	Yes	No	Yes
PowerEdge 4600	Yes	Yes	No
PowerEdge 6600	Yes	Yes	No
PowerEdge 6650	Yes	Yes	No

Figure 1. Dell PowerEdge servers and firmware components supported by Dell Update Packages

Linux kernel	Kernel source
2.4.9-e.3	Red Hat Enterprise Linux AS 2.1 Native
2.4.9-e.12	Red Hat Errata
2.4.9-e.27	Red Hat Errata
Others	See the *Dell Update Packages User's Guide* at http://support.dell.com for more information

Figure 2. Red Hat Linux kernels supported by Dell Update Packages



```
[user@precision420 SVM]# ./PE1750-BIOS-LX-A05.bin -h
Command-line options for the Update Package

Usage: <package name> [options...]

Options:

-h,--help        : Display command-line usage help
-c               : Determine if the update can be applied to the target system
-f               : Force a downgrade or update to the same or an older version (
*)
-q               : Execute the update package silently without user intervention
-r               : Reboot if necessary after the update (*)
-v,--version     : Display version information
--list           : Display contents of package (+)
--extract <path> : Extract files to specified path (+)
--rebuild        : Rebuild package to support multiple kernels(+)

* Only takes effect if used with -q
+ Can be used only before extracting the package
```

Figure 3. Command-line options for Dell Update Packages

bash or Perl, can be tailored to the specific requirements of a deployment. In addition, third-party products can further enhance this functionality and help suit the needs of the IT infrastructure.

Updating packages in interactive mode

When in interactive mode, an update package may be executed from a local console attached to the target server or from an administrator's workstation by using a remote shell, such as ssh. Once a package has been delivered to the target system, administrators can execute the package in its simplest form by typing "./*packagename*.bin" from the working directory, where *packagename* is the model and component-specific name of the update package. This syntax assumes that the .bin file has the appropriate permissions set and that a user account with root privileges executes the update. Once underway, the update package will echo its progress and status to the console (`stdout`) and log its activity to /var/log/messages. In interactive mode, the user must provide feedback as the utility executes, such as answering "yes" or "no." Figure 3 illustrates the command-line options that appear on the console when the `-help` switch is used.

Updating packages in non-interactive mode using native Red Hat tools and CLI options

Deploying Dell Update Packages in non-interactive, or *quiet,* mode is desirable when an update task must be delegated, repeated over time, or scheduled for execution at a specific time. Dell Update Packages are useful in an interactive manner, but the command-line interface (CLI) options and exit codes available in non-interactive mode can enhance the effectiveness of these packages. The CLI options and exit codes enable system administrators to create scripts that use test constructs to control the iteration of package deployment and to track deployment results in a repeatable manner suited to the storage and security infrastructure.

The standard terminal shell in Red Hat Linux is bash, one of several shells included in a typical Red Hat Linux installation. Bash is programmable and features an interpreted language that can be used interactively or in a scripted, automated manner. By using built-in and user-defined functions and calling external programs, bash can be programmed to securely distribute an update package and to execute it on one or more systems.[1]

Figure 4 itemizes key areas for a productive deployment script. Although not comprehensive, this figure addresses common execution issues and offers solutions from the Red Hat Enterprise Linux distribution (RHEL) and Dell Update Packages. Refer to the *Dell Update Packages for Red Hat Linux User's Guide* at http://support.dell.com and the man pages[2] in Red Hat Linux for further details on implementing these suggestions.

The shell-scripted deployment method provides portability and distribution compatibility, and is well suited for environments that restrict the use or installation of third-party software. However, the success of a production-level script development will depend on the complexity of the script, variances in the system infrastructure, and the administrator's level of expertise with scripting. Administrators

Script stage	Subroutine	Available options
Target system enumeration	Target system compatibility	**Dell Update Packages:** `-c` executes only the dependency checks for the update package—that is, user privilege, compatible target system and OS, and firmware interdependencies; `--rebuild` rebuilds an update package to enable its use with a kernel other than the default Dell-supported kernels
Distribution	Update package delivery	**RHEL:** rcp, scp, ftp, sftp, rdist, rsync
	Administrative account	**RHEL:** Root-level administrative account
	Authentication method	**RHEL:** ssh agents, trusted host authentication, plain-text key authentication
	File permissions	**RHEL:** chmod, chown, chgrp
Application	Deployment script or update package scheduling	**RHEL:** at, cron, batch
	Update package application	**Dell Update Packages:** `-q` executes without user interaction (required for non-interactive mode); `-f` forces downgrading to an older firmware version or reapplying an update of the same version, if necessary
	Target system availability	**Dell Update Packages:** `-r` provides the option to reboot a server immediately after an update; package precedence allows the ESM* update to precede the BIOS update without rebooting if a target server is to receive both types of firmware updates (it requires only a single reboot for both, thus avoiding an extra boot cycle)
Verification	Script iteration	**Dell Update Packages:** CLI exit codes indicate a problem if an exit code value is not zero (non-zero codes can be enumerated to identify a specific error) **RHEL:** Application exit codes report the success of each script iteration
	Log review	**RHEL:** grep, sendmail® or other mail transport, tee, wc, manual logs review

**Embedded Server Management*

Figure 4. Planning considerations for a scripted deployment

[1] For more information about bash programming, see "Introduction to bash" by the GNU Project (http://www.gnu.org/software/bash/bash.html), as well as "Bash Programming—Introduction How-To" (http://www.tldp.org/HOWTO/Bash-Prog-Intro-HOWTO.html) and "Advanced Bash-Scripting Guide" (http://www.tldp.org/LDP/abs/html) by the Linux Documentation Project.

[2] Linux man pages are available on the Web at http://www.tldp.org/docs.html#man or from the command prompt in bash by typing `man command`.

who have the option of using third-party tools and who wish to avoid the time required to code and test a script may prefer the alternative deployment method described in the next section.

Updating packages with dsh—the third-party application approach

Distributed shell, also known as dancer's shell or dsh, was originally designed for clustered environments but now supports non-clustered networks.[3] The dsh program allows administrators to execute a binary or any standard shell command on any number of machines at once. Execution can occur in series or in parallel, and machines can be assigned to groups based on any given criteria. This capability enables administrators to group similar machines or similar kernels across the enterprise, based on package requirements. The dsh program is available in two forms—a GNU C package and a Perl version—but both offer the same general capabilities.

The dsh program can use Remote Shell (rsh) or Secure Shell (ssh) for connection and command execution. Although rsh is the default, ssh is recommended for security reasons. Once compiled, dsh requires only a few changes to its configuration file to allow communication through ssh. The dsh program connects to each target system as the user; if dsh is executed as root, connections to target machines are made as root. Logins through dsh are not interactive, so any PATH statements needed beyond the default shell environment (/etc/bashrc on Linux systems running bash) must be added to the dsh command line. Dell Update Packages do not require additional PATH statements outside the standard bashrc environment.

For a mass deployment of Dell Update Packages using dsh, administrators should take the following steps:

1. Designate one system—a workstation or server—to be the controller. Install and configure dsh on this system, changing the configuration to allow for ssh.
2. Create a common mount point on the network using any method—Network File System (NFS) and Samba work well.
3. Generate ssh keys (using `ssh-keygen -t` *`type of key`* `-b` *`size`*) on each target server and on the controller. If ssh agents are being used, proceed to step 5.
4. Copy the contents of $HOME/.ssh/*id_keytype*.pub from each server to a file on the common mount point (for example, copy id_rsa.pub to the /mnt/common/authorized_keys file). Once the authorized_keys file is populated, copy it to the $HOME/.ssh directory of each target server, as well as to the $HOME/.ssh directory of the controller. This will permit ssh connections to occur without administrators having to supply authorization credentials to each system.
5. Once ssh is configured to operate without prompting for login credentials, populate the dsh machines.list file with the IP addresses or hostnames of the target servers. This file is white space–delimited; one entry per line is recommended.
6. With machines.list populated, dsh can be used to execute a command on every system in the machines.list file at once. The man pages for dsh include detailed instructions on setting up groups, enabling administrators to execute commands against a common group of servers.
7. Copy Dell Update Package(s) to the common mount point.
8. Issue the `dsh` command to execute the package on every target server in the machines.list file. The command should resemble the following:

   ```
   dsh -a -c "sh /mnt/common/PE1650-BIOS-LX-A05.bin
   -q -r"
   ```

   where `-a` tells dsh to execute the command on all target servers, `-c` tells dsh to execute concurrently, and the line in quotes is the command to be run on the target servers.

The dsh program is readily scriptable, enabling administrators to write scripts easily—and in the language of their choice—to prompt for variables and execute the `dsh` command. A framework can be created for future package names and locations, prompting for each of these variables. The script also can be used to build the machines.list file based on input from administrators, allowing for greater flexibility. Finally, command-line options for the package, and any notifications, can be added to the script. This type of fully featured script requires minimal effort and offers administrators great flexibility and ease of use during each update cycle.

Simplifying systems management in Linux environments

Dell Update Packages are available for select Dell PowerEdge servers running Red Hat Enterprise Linux AS as well as for PowerEdge servers running Microsoft® Windows® operating systems.[4] By deploying Dell Update Packages, IT administrators can simplify the management of PowerEdge servers and help achieve high performance and availability in their data centers.

Bala Beddhannan (bala_beddhannan@dell.com) is a test engineer and advisor in the Dell Enterprise System Test group. He has a B.E. from Anna University, India, and an M.S. in Interdisciplinary Engineering from Texas A&M University.

John Brunken (john_brunken@dell.com) is a senior system test technician in the Dell Enterprise Server Group. He has an Associate of Science degree in Electronic Technology from Austin Community College and has 14 years of experience in the IT industry.

Jason D. Norman (jason_d_norman@dell.com) is a senior system test technician in the Dell Enterprise Server Group. He has 11 years of experience in the IT industry.

[3] For more information on dsh, please visit the dsh author's Web site at http://www.netfort.gr.jp/~dancer/software/dsh.html.en or the Perl dsh project at http://sourceforge.net/projects/dsh.

[4] For information about Windows-based systems, see "An Introduction to Dell Update Packages" by Karl Friedrich and Sandeep Karandikar in *Dell Power Solutions,* August 2003.

Maintaining Network Uptime Using

Serial Console Port Management

As enterprises increasingly rely on the network infrastructure to convey vital information and communication, IT managers face a growing imperative to keep networks free from downtime. Console port management tools such as the Equinox™ CCM Console Manager can help reduce management complexity and maximize uptime. By combining in-band and out-of-band management features, the Equinox CCM Console Manager is designed to provide secure, reliable local and remote access to and management of serially managed network devices such as servers, routers, hubs, storage area networks, firewalls, and switches.

BY LISA STOUT AND RON RASMUSSEN

The need for continuous access to business information has increased the reliance of enterprises on networks. Stable, well-functioning networks help organizations to improve customer service, increase supply-chain efficiency, and bring products to market faster. High-availability networks help provide the infrastructure for optimal revenue streams and uninterrupted customer service.

Unfortunately, in recent years the growing dependence on networks has been accompanied by an increase in the potential for unplanned network downtime. Several factors have contributed to this risk: IT systems have become more complex and widely distributed within organizations; corporate downsizing has reduced the number of IT personnel available to service the network infrastructure; and systems management can be more complex because hardware and operating system (OS) environments have become increasingly heterogeneous.

- **Network complexity:** Traditional client/server architecture with access to a few large, centrally colocated servers has given way to widely dispersed computing resources. In many organizations, networking and computing hardware are located in unmanned data centers, which may be difficult to access in a timely fashion.
- **Network maintenance and management:** Demands on IT staff are greater than ever, and many organizations must accomplish more using fewer resources in today's economic environment where reductions in total cost of ownership (TCO) are paramount. As IT staff is reduced, mean time to repair (MTTR) critical networking and computing resources may increase.
- **Management complexity:** In increasingly heterogeneous environments, IT staff must access a myriad of devices in the data center, including different

types of servers as well as network devices such as routers, firewalls, and power supplies.

For large organizations, overall downtime costs can average 3.6 percent[1] of annual revenue. Application problems are the single largest source of downtime, causing 30 percent[2] of annual downtime hours and 32 percent[3] of downtime cost, on average. Other sources of downtime include problems with network products, security products, servers, service providers, cables and connectors, and e-commerce.[4] Downtime hours are evenly split between outages and service degradations, but outages cost more—an average of 58 percent[5] of downtime costs come from outages and 42 percent[6] from degradations.

To help reduce network downtime, decrease the need for IT staff, and centralize network management—a particularly important goal for the ever-growing number of IT departments that manage distributed data centers—organizations can implement network management strategies with high availability in mind. An approach designed to maximize uptime includes the following considerations:

- The need for remote access to servers and third-party devices that are unresponsive or cannot be reached over the network
- The requirement to manage heterogeneous hardware, applications, and operating systems
- The importance of redundancy to help ensure high availability for the enterprise
- Security requirements for accessing network devices remotely

Understanding serial console port management

Data centers rely on a wide range of hardware devices to support the network infrastructure, including switches, routers, hubs, mail servers, and file-and-print servers as well as mainframes and workstations. A typical data center environment also involves several operating systems. Administrators require an efficient and secure method to access and control an extensive variety of devices on the network.

Network administrators normally access and manage servers, routers, and other data center devices using HTTP, Secure Shell (SSH), or Telnet protocols across *in-band* network connections that rely on an enterprise's main communication network. However, in-band management methods can be used only when a device—and the network to which it is attached—is properly functioning. *Out-of-band* management techniques allow administrators to access devices over a separate route so they can control systems even when the main network is down. By providing direct "backdoor" access, the out-of-band network management approach enables administrators to diagnose and reconfigure devices when the hardware is not responsive to normal in-band requests. An out-of-band connection can be used to control a computer through every stage of operation—from booting up (a pre-OS state) to fully operational status. The time—and therefore, cost—required to restart access to infrastructure components can be minimized when out-of-band technology is deployed properly.

Out-of-band connections are most commonly implemented through serial console ports. Serial console ports are a logical choice for controlling remote and local computers on the network for several reasons:

- Most data center devices have at least one serial port for console access.
- Communication standards that govern serial ports are well defined.
- Many text-based applications allow for local or remote interaction with serial ports.
- Serial ports can be simple, reliable, and flexible enough for use in multiple-OS environments.

The console port is the management interface that is commonly used by most network appliances as well as many of the operating systems popular in today's data centers—including Linux®, UNIX®, Sun™ Solaris™, and Microsoft® Windows Server™ 2003 operating systems. In fact, recovery services such as Windows® Emergency Management Services (EMS) as well as Telnet and SSH protocols are designed to use a serial console port to gain access to network devices. Because serial console port access is pervasive, serial console port management tools offer an excellent method for accessing local and remote serial devices through in-band networks, as well as providing out-of-band connectivity options such as dial-up and serial terminal access.

The Equinox™ CCM Console Manager is designed to provide reliable and secure serial-over-IP access to the console ports of servers and other serially managed devices. By providing both in-band and out-of-band console management, the CCM Console Manager gives administrators the remote access that can be essential to efficient network device management and troubleshooting (see Figure 1).

Providing reliable, centralized access to serial devices

The CCM Console Manager provides a common interface that allows administrators to interact with various network components—such as servers, routers, power management devices, telecommunications equipment, network switches, firewalls, and other serial-accessible devices. The CCM Console Manager can communicate with these devices over a local area network (LAN) or wide area network (WAN) using in-band management methods, or over dial-up or direct terminal connections using out-of-band methods.

The CCM Console Manager is compatible with industry-standard Telnet and SSH clients. These clients enable secure, remote

[1, 2, 3, 4, 5, 6] Source: Infonetics Research, "Large Companies Lose 3.6% of Annual Revenue to Network Downtime," February 11, 2004, http://www.infonetics.com/resources/purple.shtml?nr.upna04_down.021104.shtml.

Dave is on vacation.

He’s been notified of a problem with his company’s servers.
He’s able to manage those servers from anywhere.
He’s not cutting his vacation short.

He’s right here.

MANAGE FROM ANYWHERE.

The Dell remote console switch provides the traditional functionality of KVM switching with the added benefit of **KVM Over IP**. Using Dell’s remote console software, simply point-and-click to take control of your servers.

Whether you’re at the rack, in your office or across the globe your data center will always be within your reach.

Take **control** of your servers with
the **Dell Remote Console Switch.**

Digital Availability. **Easy as DELL**

www.dell.com

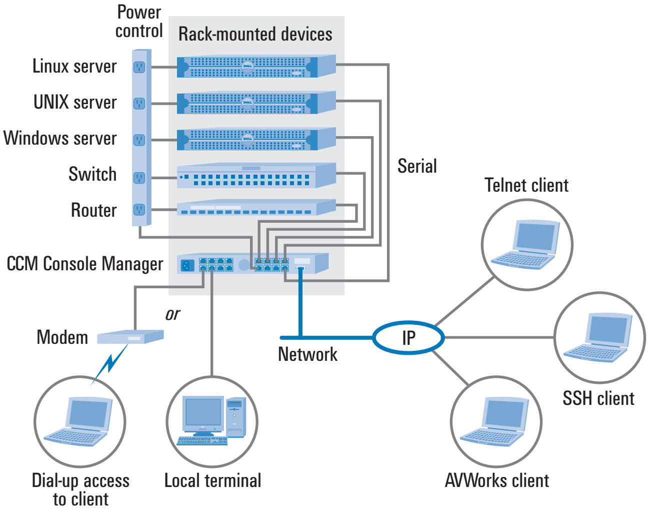


Figure 1. Equinox CCM Console Manager topology

out-of-band access to network devices. By launching a Telnet session directly to the CCM Console Manager, the administrator is presented with the command-line interface (CLI) of the appliance. This interface allows the administrator to access the attached devices or the CCM Console Manager itself. The CCM840 and CCM1640 Console Managers also can be accessed using Avocent® AVWorks® management software, which provides a graphical user interface (GUI) that helps enable administrators to consolidate device discovery, installation, administration, access, and control into a single cross-platform application (see Figure 2). Clicking on a desktop icon within AVWorks launches a console session to the selected device. Centralized access and control of serially managed devices on the network can help decrease administrative overhead while improving the stability of network systems.

AVWorks provides a list of all network devices and provides access to these devices based on assigned permissions. By clicking on a device, an administrator can establish a connection between AVWorks and the target device. The CCM Console Manager then transmits data regarding the status and health of the attached device to AVWorks for compilation and viewing. Using AVWorks, administrators can easily locate a device and establish a secure in-band or out-of-band Telnet or SSH session in which they can alter device configuration parameters, diagnose and recover from system faults, view statistics, and perform various other control-related functions.

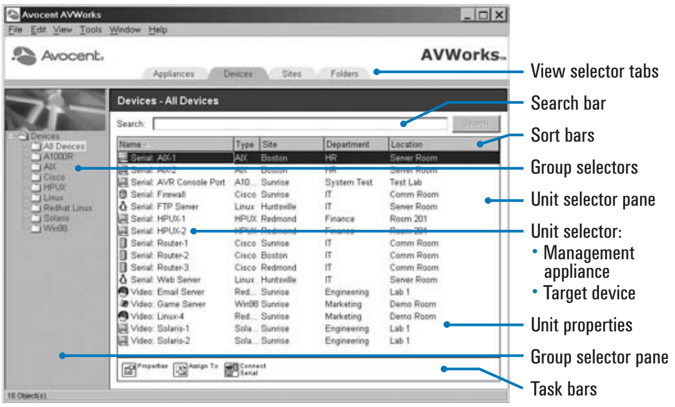
Additional management features, such as offline history buffering and hardware break suppression, distinguish the CCM Console Manager from simple terminal servers, which are often used for similar purposes. Equinox CCM Console Manager Simple Network Management Protocol (SNMP) management information base 2 (MIB2) and enterprise traps integrate directly with existing SNMP management systems to indicate events such as failed authentication and offline history buffer. This capability helps ensure that IT staff does not miss important information such as data pertaining to traffic on the attached devices, failures, and user lockouts.

To permit access to existing user accounts, the CCM Console Manager is compatible with servers that use the Remote Authentication Dial-In User Service (RADIUS) protocol. In addition, the CCM Console Manager contains an on-board database that allows for the creation of 64 on-board user accounts with individual passwords. Specific access rights may be set for each account. Advanced features such as access control levels helps ease the task of assigning user access rights. The CCM Console Manager offers three levels of access control: Appliance Administrator, User Administrator, and User. The CCM Console Manager provides a 10/100BaseT (8- and 16-port) and 10/100/1000BaseT[7] (48-port) Ethernet connection for attaching to the network, and up to 48 serial ports for attaching to the serial console ports of servers, routers, and other devices. Individual users may have up to 48 simultaneous Telnet sessions, depending on the number of ports available per CCM Console Manager.

Using security features for secure connections

Security is an important concern for remote IT administrators because network connections—particularly in distributed environments—can span both public and private subnetworks. The CCM Console Manager offers several attributes that help provide secure local and remote IP connections:

- **Permissions:** Not all administrators need or should have the same access rights. The CCM Console Manager enables IT staff to assign user-specific access rights and privileges, restricting access to devices based on set permissions.
- **Client authentication:** Authentication software such as RADIUS is frequently used to identify individuals and provide



Devices - All Devices

Search:

Name	Type	Site	Department	Location
Serial: AIX-1	AIX	Boston	HR	Server Room
Serial: AIX-2	AIX	Boston	HR	Server Room
Serial: AVR Console Port	A10...	Sunrise	System Test	Test Lab
Serial: Firewall	Cisco	Sunrise	IT	Comm Room
Serial: FTP Server	Linux	Huntsville	IT	Server Room
Serial: HPUX-1	HPUX	Redmond	Finance	Room 201
Serial: HPUX-2	HPUX	Redmond	Finance	Room 201
Serial: Router-1	Cisco	Sunrise	IT	Comm Room
Serial: Router-2	Cisco	Boston	IT	Comm Room
Serial: Router-3	Cisco	Redmond	IT	Comm Room
Serial: Web Server	Linux	Huntsville	IT	Server Room
Video: Email Server	Red...	Sunrise	Engineering	Lab 1
Video: Game Server	Win98	Sunrise	Marketing	Demo Room
Video: Linux-4	Red...	Sunrise	Marketing	Demo Room
Video: Solaris-1	Sola...	Sunrise	Engineering	Lab 1
Video: Solaris-2	Sola...	Sunrise	Engineering	Lab 1

Figure 2. Avocent AVWorks interface

[7] 1000BaseT 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.

access to key network components based on user privileges. The CCM Console Manager is compatible with the RADIUS protocol, which is widely used in network environments for the Linux OS and embedded network devices such as routers, modem servers, switches, and so forth.

- **Encryption:** Console port management tools are designed to allow IT administrators to encrypt sensitive data using protocols such as SSH. The CCM Console Manager contains an on-board SSH v2 server to support encryption—a key security feature not available when using plain-text terminal servers.

Providing redundancy for 24/7 network access

Data centers are rapidly evolving to accommodate higher expectations for growth, consolidation, and security. Stringent demands for uptime and service availability, coupled with new technology and protocols, make data center design efforts more challenging and demanding.

High availability translates into a fully redundant architecture in which all possible hardware failures are predictable and deterministic. This implies that each component has a predetermined failover and fallback time, which is unrealistic at best.

Because redundancy is a critical aspect of maintaining uptime in the network infrastructure, administrators should consider combining the CCM Console Manager with other in-band and out-of-band management tools. The CCM Console Manager can act as a universal secure access point—an aggregator—to myriad devices in the data center. Although most enterprises have high-end network strategies for in-band access, these implementations assume that the network and OS are fully functional. Any key point of failure may require out-of-band access, or in the case of a failing server, access at the BIOS level of operation.

The following are some of the advantages of combining the CCM Console Manager with other in-band and out-of-band management tools and strategies:

- **Remote access:** One way to help prevent catastrophic failure is to locate a portion of the enterprise data center at a remote location. The CCM Console Manager can provide out-of-band access through a third-party Telnet, SSH v2, or HyperTerminal dial-in connection.
- **KVM management:** Keyboard, video, mouse (KVM) switches provide access to Windows-based servers through a keyboard, video, and mouse. The CCM Console Manager can be deployed in conjunction with a KVM switch to provide universal access to the multiple devices and operating systems in the data center—and can be used to manage the KVM switch too. The CCM Console Manager uses a software client similar to the Dell™ 2161DS Remote Console Switch, providing a familiar interface to administrators who have deployed this KVM technology.
- **Access to failed devices and networks:** Another important aspect of a redundant data center is access to a device when its OS—or the network on which it resides—is not operational. In such situations, the CCM Control Manager is one of the few tools that can provide access.
- **Remote access controllers:** Dell Remote Access Card III (DRAC III) or other Baseboard Management Controllers (BMCs) provide access to servers that incorporate such controllers. The CCM Console Manager can act as an aggregator for these servers as well as for other devices or servers that do not use BMCs. The CCM Console Manager can also provide security.
- **Device monitoring:** Catching problems before they become critical or negatively affect business is important. The CCM Console Manager supports MIB2 traps, which interface with SNMP managers to provide e-mail or message alerts when a device in the data center is failing.

Maintaining network uptime and data access

Network downtime can be costly to any organization, and employing the right network management tools can help maintain network uptime and data access. A reliable, easy-to-use appliance that offers both in-band and out-of-band connection methods using a single interface can be an effective management tool, particularly in multiple-OS and mixed-hardware data center environments.

The Equinox CCM Console Manager is hardware and software independent, offering local and remote access to heterogeneous network devices through one console. By providing secure, centralized access to network devices, whether they are operational or not, remote console management products such as Equinox CCM Console Manager can provide an effective way to help reduce downtime—streamlining IT management and lowering organizational costs while helping to improve access to mission-critical network systems and information.

Lisa Stout (lstout@equinox.com) is a strategic account manager for Equinox Systems, an Avocent company and original equipment manufacturer (OEM). Lisa has more than 12 years of experience in the high-technology industry, and has a B.A. in Communications and an M.B.A. from St. Edward's University.

Ron Rasmussen (rrasmussen@equinox.com) is the director for OEM and strategic sales at Equinox Systems, an Avocent company. Ron has more than 25 years of experience in the data communications industry with 17 years in hardware and software sales and marketing. He previously taught computer programming and designed networks at ITT Technical Institute.

FOR MORE INFORMATION

http://www.avocent.com

http://www.equinox.com

Using WMI Scripting

for System Administration

Windows® Management Instrumentation (WMI), which offers administrators a rich management scripting tool set, is an integral part of the Microsoft® Windows family of operating systems—which includes Microsoft Windows 2000, Windows XP, and Windows Server™ 2003. This article introduces WMI, explains basic scripting techniques, and provides examples for automating systems management processes using Dell™ OpenManage™ tools.

BY SUDHIR SHETTY

Windows® Management Instrumentation (WMI) is the underlying management technology for Microsoft® Windows operating systems. Based on industry standards, WMI helps enable consistent, uniform management control and monitoring of systems throughout an enterprise. WMI allows system administrators to query, change, and monitor configuration settings on desktop and server systems, applications, networks, and other components of the IT infrastructure. System administrators can create a wide range of systems management and monitoring scripts that work with WMI using the WMI scripting library.

Overview of WMI architecture

Figure 1 shows the three primary layers of the WMI architecture: WMI data providers, WMI management infrastructure, and WMI consumers.

WMI data providers

The data providers retrieve information from managed resources. A managed resource is a logical or physical component that can be accessed and managed using WMI. Examples of standard Windows resources that can be managed using WMI include computer systems, disks, peripheral devices, event logs, files, folders, file systems, networking components, operating system subsystems, performance counters, printers, processes, registry settings, security, services, Microsoft Active Directory® directory service, Windows Installer, and Windows Driver Model device drivers. WMI data providers access the managed resources using the appropriate application programming interfaces (APIs) and expose the data to the WMI infrastructure using a standards-based, object-oriented data model.

The operating system is bundled with standard providers for accessing standard operating-system resources such as processes, registry settings, and so forth. In addition, WMI enables software developers to add and integrate additional providers that expose data and management functionality unique to their products. For example, Dell provides a Common Information Model (CIM) provider, which exposes additional asset information through WMI. The *Dell™*

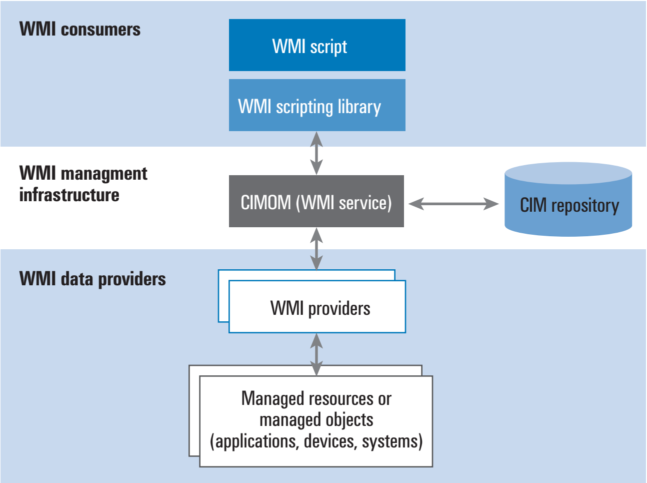


Figure 1. Three primary layers of the WMI architecture

OpenManage™ Server Administrator CIM Reference Guide contains additional information on the data model supported by this provider. The documentation is available online at http://docs.us.dell.com/docs/software/svradmin.

WMI management infrastructure

The WMI infrastructure consists of the Common Information Model Object Manager (CIMOM), which manages the interaction between consumers and data providers. All WMI requests and data flow through the CIMOM. The WMI service acts as the CIMOM and is similar to other operating system services. For example, it can be stopped and started using the following commands:

```
net stop winmgmt
net start winmgmt
```

Management applications, administrative tools, and scripts make requests to the CIMOM to retrieve data, subscribe to events, or perform management-related tasks. The CIMOM retrieves the provider and class information to service the request from the CIM repository. The CIMOM uses the information obtained from the CIM repository to pass the consumer request to the appropriate WMI provider.

The CIM repository holds the schema, or *object repository*, that models the managed environment. CIM uses a rich object-oriented data model and the notion of classes to represent information about managed resources. For example, a Win32_LogicalDisk class represents logical disk information on a computer. Instances of this class, such as information about the C: drive, D: drive, and so on, can be retrieved by querying the CIMOM. Attributes, or *properties*, of this class, such as FreeSpace, Name, and Size, can be queried remotely. Because the operational state of most resources changes frequently, information is typically read on demand to ensure that up-to-date information is retrieved.

CIM classes are organized into namespaces. Each namespace contains a logical group of related classes representing a specific area of management.

WMI consumers

WMI consumers are management applications or scripts that access and control management information available through the WMI infrastructure.

Introduction to scripting

Windows scripting provides a rich environment in which administrators can develop scripts. Scripts can access Component Object Model (COM) objects that support automation or standard access methodologies. This capability enables scripts to access COM-based technologies, such as WMI or Active Directory Service Interfaces (ADSI), to manage Windows subsystems.

The examples in this article use VBScript as the scripting language because of its widespread usage. Other scripting technologies (for example, Microsoft JScript® development software) that support COM automation can also be used to access WMI objects. Available on Windows XP and Windows Server™ 2003 is a command-line tool for accessing WMI information, called WMIC (Windows Management Instrumentation Command-line).

Figure 2 provides a simple example of a WMI script, which lists the logical disk information for the local computer. To invoke

```
strComputer = "."
  Set objWMIService = GetObject("winmgmts:" _
    & "{impersonationLevel=impersonate}!\\" & strComputer & _
    "\root\cimv2")
  Set colDisks = objWMIService.ExecQuery _
    ("SELECT * FROM Win32_LogicalDisk")

Const CONVERT_TO_MB = 1048576
For each objDisk in colDisks
  WScript.Echo "Name:" & objDisk.Name
  WScript.Echo "Size(MB):" & Int(objDisk.Size / CONVERT_TO_MB)
  WScript.Echo "Free Space (MB):" & _
  Int(objDisk.FreeSpace /CONVERT_TO_MB)
Next
```

Figure 2. Sample WMI script for listing logical disk information

the script, save the listing into a file called disk.vbs and run the following command:

```
C:> cscript disk.vbs
```

Cscript.exe is a console-based scripting host that is packaged with the Windows Script Host (WSH) and is used for running scripts.

Note that the script in Figure 2 can be adapted to access a variety of information about managed resources. Information about managed resources can be viewed by downloading WMI administrative tools from http://www.microsoft.com/downloads/details.aspx?FamilyId = 6430F853-1120-48DB-8CC5-F2ABDC3ED314&displaylang = en. CIM Studio, which is bundled in the WMI administrative tool set, lets administrators view class information in the CIM repository. For other sample scripts that perform a variety of useful system administration tasks, visit *Dell Power Solutions* online at http://www.dell.com/magazines_extras.

Figure 3 lists additional useful tasks that system administrators can script.

Scripting to execute processes on remote Dell systems

Administrators can write scripts to execute processes on remote systems, a technique that can be used in many ways to manage a network of Dell systems in the enterprise. Examples in this section illustrate the remote execution of Dell OpenManage Server Administrator command-line interface (CLI) commands on a Dell PowerEdge™ server. To remotely execute other processes on target nodes, administrators can customize these examples. A complete command reference can be found in the *Dell OpenManage Server Administrator Command-Line Interface User's Guide*, which is available on the documentation CD that ships with every PowerEdge server or online at http://docs.us.dell.com/docs/software/svradmin.

Windows entity	Description	WMI-scriptable usage model
Event logs	Provide a repository of activities on a computer	• Retrieve and query event logs • Configure event log properties • Back up and clear event logs
Printers	Manage printers in the enterprise	• Monitor printers and print jobs • Manage printer connections on client systems
Registry	Contains configuration settings for the operating system, applications, and services under Windows	• Create new registry entries • Delete and update registry entries • Back up the registry
Files/folders	Contain valuable enterprise data	• Manage files and folders (create, delete, and update) • Enumerate files • Manage shares • Monitor the file system
Disks and file system	Manage disks and file systems to ensure continued access to data	• Manage disk partitions • Manage logical disk drives • Manage disk space
Active Directory users	Manage user account information in Active Directory	• Create user accounts • Manage user accounts • Delete user accounts
Computer asset information	Contains an inventory of hardware information, including the operating system and other software installed on the remote computer	• Install, upgrade, and remove software • Retrieve system information, such as operating system, memory, and disk space • Manage computer startup and recovery settings
Computer roles	Manage computers and computer roles in Active Directory	• Create, manage, and delete computer accounts in Active Directory • Manage the roles assigned to computers
Processes	Manage running instances of an application or an executable file	• Monitor processes • Create and terminate processes • Enumerate additional process properties

Figure 3. Examples of scriptable administrative tasks

Based on industry standards, WMI helps enable consistent, uniform management control and monitoring of systems throughout an enterprise.

Figure 4 shows an example script that executes a process on a remote system. To invoke the script on a remote server, called MachineA in this example, save the preceding listing into a file called remote.vbs and run the following command:

```
C:> cscript remote.vbs MachineA "omreport system
   esmlog -outa c:\%COMPUTERNAME%_ESMLOG.TXT"
```

The script in Figure 4 initiates the remote execution of the process and does not wait for the process to complete. Executing the script exports the Embedded Server Management (ESM) log into a text file on the remote target. This capability offers a rich set of possibilities for the remote scripting of PowerEdge Servers using Server Administrator CLI commands, including:

- `omconfig`: Configures values using the CLI. Administrators can use specific values for warning thresholds on components or prescribe what action a system should take when a certain warning or failure event occurs. Administrators can also use the `omconfig` command to assign specific values to a system's asset information parameters, such as the purchase price of the system, the system's asset tag, or the system's location.
- `omdiag`: Runs diagnostic tests against system hardware to isolate problems.
- `omhelp`: Displays short help text for CLI commands.
- `omreport`: Produces reports of a system's management information.
- `omupdate`: Installs the latest update packages for the system's BIOS, firmware, and drivers.

Administrators can run sequences of these commands using a batch file on a remote managed node. For a sample script that executes a batch file on a remote node, visit *Dell Power Solutions* online at http://www.dell.com/magazines_extras.

```
strComputer = WScript.Arguments.Item(0)
strCommand = WScript.Arguments.Item(1)

Set objWMIService = GetObject("winmgmts:" _
  & "{impersonationLevel=impersonate}!\\" & strComputer & _
  "\root\cimv2:Win32_Process")

errReturn = objWMIService.Create(strCommand,null,null,intProcessID)
if errReturn = 0 Then
  Wscript.Echo strCommand & " was started with a process ID of " _
  & intProcessID & "."
Else
  Wscript.Echo strCommand & " could not be started due to error " & _
  errReturn & "."
End If
```

Figure 4. Sample WMI script for executing a remote process

Performing a remote software update under Dell OpenManage

Administrators can perform a remote software update for BIOS, firmware, and drivers by using the `omupdate` command or by using a Dell Update Package.

`omupdate` **CLI command.** To execute the `omupdate` command, an administrator would invoke the following:

```
omupdate biosupdate path="\\network_share\file_name "
```

This sample command refers to the network share where the update file resides. In the absence of a network share, administrators can copy the file to a temporary directory on the local system before performing the update.

Dell Update Package. Another option is to copy the Dell Update Package, which is a self-contained .exe file, onto the remote system. The .exe file can be directly executed on the remote node by passing its command-line arguments for a silent installation of the package and rebooting the system, if necessary. The WMI script listed in Figure 5 illustrates a remote software update that passes in the computer name and the name of the Dell Update Package. In the script, note that the Dell Update Package is invoked with the `/s /r` option to perform a silent update and reboot of the system, if necessary.

For a scheduled update, administrators can use the Windows Scheduled Task Wizard on the management station to perform the update during off-peak hours.

Security considerations

WMI scripts run in the security context of the administrator running the script. To run scripts on a remote computer, a user needs administrative access on the remote system, thus helping to prevent malicious individuals from running destructive scripts on a remote server. Remote WMI scripting leverages Distributed COM (DCOM) for interacting with the WMI service running on the remote system.

The scripts in this article set the impersonation level to “impersonate.” This setting implies that the WMI service uses the administrator’s security context to perform the requested operation. If the administrator’s security credentials are not adequate, errors such as “Access Denied” will occur when running the script.

WMI provides a rich security infrastructure that can be configured on a per-namespace basis. By default, the administrator’s group has full control of WMI on both local and remote computers, and users in other groups do not have remote access to the computers. To view a computer’s security settings, invoke wmimgmt.msc from the command prompt, view the properties of the root node, and look

```
strComputer = WScript.Arguments.Item(0)
strDupFile = WScript.Arguments.Item(1)

'* First copy the Dell Update Package to the remote system
Set objFSO=CreateObject("Scripting.FileSystemObject")
ObjFSO.CopyFile strDupFile, "\\" & strComputer & "\C$\tmp\" & _
  strDupFile

'* Launch the process on the remote system
Set objWMIService = GetObject("winmgmts:" _
  & "{impersonationLevel=impersonate}!\\" & strComputer & _
  "\root\cimv2:Win32_Process")

strCommand = "C:\tmp\" & strDupFile & " /s /r"
errReturn = objWMIService.Create(strCommand,null,null,intProcessID)
if errReturn = 0 Then
  Wscript.Echo strCommand & " was started with a process ID of " _
  & intProcessID & "."
Else
  Wscript.Echo strCommand & " could not be started due to error " & _
  errReturn & "."
End If
```

Figure 5. Sample WMI script for a remote software update

PASSING ARGUMENTS THROUGH A TEXT FILE

Several scripts in this article involve passing arguments through the command line. An alternative is to pass arguments as a text file. For example, to perform a software update on a set of PowerEdge 2650 servers, administrators can export an external database to list several servers in a text file, such as Server1, Server2, and Server3.

This text file can then be passed as a command-line argument to the example script listed in Figure A.

```
Set objArgs = Wscript.Arguments
Const ForReading = 1
Set objDictionary = CreateObject("Scripting.Dictionary")
Set objFSO = CreateObject("Scripting.FileSystemObject")
Set objTextFile = objFSO.OpenTextFile(objArgs(0), ForReading)

'* Read the file which contains a list of machines, one per line
I = 0
Do While objTextFile.AtEndOfStream <> True
  StrNextLine = objTextFile.ReadLine
  objDictionary.Add I, strNextLine
  I = I + 1
Loop

'* For each machine, retrieve the number of services running on that machine
For Each objItem in objDictionary
  strComputer = objDictionary.Item(objItem)
  Wscript.Echo "Accessing remote machine" & strComputer
  Set objWMIService = GetObject("winmgmts:" _
    & "{impersonationLevel=impersonate}!\\" & strComputer & _
    "\root\cimv2")
  Set colServices = objWMIService.ExecQuery _
    ("SELECT * FROM Win32_Service")

  Wscript.Echo strComputer, colServices.count
Next
```

Figure A. Sample WMI script for reading arguments from a text file

at the security tab of the dialog box. Administrators can alter security settings at the namespace level and grant different permissions (remote access, method execution, or write) to different sets of users (domain or local).

A powerful tool for performing administrative tasks

WMI scripting is designed to provide a powerful mechanism that helps system administrators to perform remote administration tasks. The techniques presented in this article help provide an introduction to the power of WMI scripting. The Microsoft Web page "Scripting Access to WMI" (http://msdn.microsoft.com/library/default.asp?url=/library/en-us/wmisdk/wmi/scripting_access_to_wmi.asp) and the book *Microsoft Windows 2000 Scripting Guide* by the Microsoft Windows Resource Kit Scripting Team provide additional valuable information about WMI and scripting.

Sudhir Shetty (sudhir_shetty@dell.com) is a member of the Enablement Technologies team at Dell, working on systems management console solutions. He has more than 10 years of experience in designing and developing software components for user interfaces; database access; and networking in application areas such as systems management, systems prototyping and simulation, and tiered, distributed computing. Sudhir has an M.S. in Computer Science from The University of Texas at Austin.

Unattended Server Installation

Using the Remote Floppy Boot Feature of the Dell Remote Access Controller

By automating the process of installing operating systems on servers, enterprises can improve their operational efficiencies. This article describes a method of unattended installation that uses the Remote Floppy Boot feature of the Dell™ remote access controller to deploy the Microsoft® Windows Server™ 2003 operating system on Dell PowerEdge™ servers.

BY SHELLEY PALMER-FETTIG AND STEPHAN MAAHS

To help decrease total cost of ownership (TCO) and increase return on investment (ROI), many enterprises are looking to streamline IT processes. One way to help lower costs is to automate recurring, time-consuming tasks. In particular, automating the server installation process can help reduce the amount of time required for system rollouts—which can be critical when servers must be deployed soon after they are ordered. Automation also allows for a standard build, which helps decrease the amount of time administrators must spend configuring each installation. Moreover, automated (also known as *unattended*) installations are designed to eliminate the need for administrators to physically access the servers, whether the systems are local or remote.

Dell™ remote access controllers (RACs) enable unattended installations through the Remote Floppy Boot feature. RACs, which include the Dell Remote Access Card (DRAC) III and the Embedded Remote Access (ERA) controller, allow administrators to perform management operations remotely over a LAN or a phone line. To implement unattended installations, administrators can combine the functionality of the Remote Floppy Boot feature with tools provided by the Dell OpenManage™ Deployment Toolkit. The toolkit includes DOS-based utilities and scripts for configuring and deploying Dell servers using Microsoft® Automated Deployment Services, the Altiris® Deployment Solution™ tool, Windows® scripting, and other deployment tools. This article describes the unattended installation of the Microsoft Windows Server™ 2003 operating system on Dell servers installed with RACs. In addition, the article discusses how to build some of the key components required to install an unattended system.

Understanding the unattended installation process

An unattended installation using the RAC generally proceeds as follows:

1. An administrator or a script issues a RAC command to remotely boot the target bare-metal server—that is, a server with no operating system installed.
2. The target server uses Trivial FTP (TFTP) to contact the deployment server for the floppy image it needs for booting; the location of the image is specified in the RAC configuration, by the administrator using the RAC graphical user interface (GUI) or by a command in the installation script.
3. An administrator or a script changes the target server's boot order so that it boots from the remote

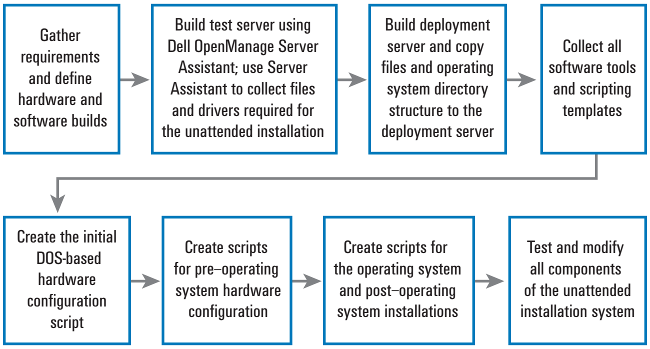

Gather requirements and define hardware and software builds
Build test server using Dell OpenManage Server Assistant; use Server Assistant to collect files and drivers required for the unattended installation
Build deployment server and copy files and operating system directory structure to the deployment server
Collect all software tools and scripting templates
Create the initial DOS-based hardware configuration script
Create scripts for pre–operating system hardware configuration
Create scripts for the operating system and post–operating system installations
Test and modify all components of the unattended installation system

Figure 1. Overview of tasks required to set up an unattended installation using the RAC

floppy image until the hardware configuration process has been completed. The Dell OpenManage Deployment Toolkit includes a utility to toggle the boot order, which enables this change to be scripted.

4. The deployment server responds by delivering a DOS-based remote floppy image to the RAC of the target server.
5. The bootable floppy image configures the target server hardware using the Dell OpenManage Deployment Toolkit. A utility sets the STATECFG variable in the BIOS, which tracks the number of reboots to represent the server's process stage.
6. An administrator or a script may need to reboot the target server two or three times to complete the hardware configuration phase.
7. After the last reboot needed for the hardware configuration phase, scripts included on the remote floppy image map a network share to a distribution directory on the deployment server, which contains the Windows operating system files for installation, as well as additional files such as drivers, tools, and applications to be copied to the target computer. Brief instructions on how to create the distribution directory can be found in the *Microsoft Windows Server 2003 Deployment Kit* or online at http://www.microsoft.com/windowsserver2003/techinfo/reskit/deploykit.mspx.
8. An administrator or a script changes the boot order of the target server so that it boots from the hard drive. The Dell OpenManage Deployment Toolkit includes a utility to toggle the boot order, which enables this change to be scripted.
9. After the operating system is installed, the target server will reboot once more. Additional installation instructions may run from scripts on the target server to perform silent installations of various applications.

Building a deployment server to prepare for an unattended installation

To implement the unattended installation process, administrators must prepare several components that support the installation, including a test server, a deployment server, and deployment script files. Figure 1 summarizes the general process for building the components required by an unattended installation.

One key component of an unattended installation is the deployment server. This component, which is set up as a TFTP server to facilitate remote network access, acts as a central repository for all deployment files and can be used successively as a testing space, a repository for a network deployment, and a repository to build a bootable deployment disk.

Administrators can use Dell OpenManage Server Assistant to collect the files that control the installation of Microsoft Windows on the target server. The administrator copies these and other files to the deployment server. Figure 2 lists the major components of the deployment server and the sources of the files that make up these components.

Building the bootable floppy image

The floppy image, which is required for the Remote Floppy Boot feature, contains the applications that need to be loaded at boot time. The Remote Floppy Boot feature (see Figure 3) can be used only on a server that boots to a DOS partition, so administrators must create a bootable MS-DOS 6.22 disk that contains the following scripts and files:

- **autoexec.bat:** Script that configures the environment while the configuration process runs on the target server; created by an administrator
- **dellpre.bat:** Core script that configures the hardware; created by an administrator
- **racsetup.bat:** Script that reconfigures the RAC settings when the server no longer needs to boot from the floppy image; created by an administrator
- **Dell OpenManage Deployment Toolkit:** Utilities that are used in dellpre.bat to drive the hardware configuration of the target server
- **racadm.exe:** Command-line utility to communicate with the RAC
- **Other DOS utilities:** Tools that are used in the scripts

Deployment server component	Population method
Operating system–specific directory, such as z:\w2k	Copy files collected using Server Assistant to this directory
An \i386 directory within the operating system–specific directory, such as z:\w2k\i386	Copy contents of the \i386 directory from the Windows operating system installation media to this directory
System-specific directory, such as z:\pe1650, for each PowerEdge platform	Copy files from the \pe** directory of the Server Assistant CD to the root directory of the network share on the deployment server
Web-based RAC GUI and Racadm command-line utility	Install Remote Console using the Dell OpenManage Systems Management CD
Floppy image	Copy the image of the boot floppy to the TFTP location given to the RAC

Figure 2. Components of the deployment server

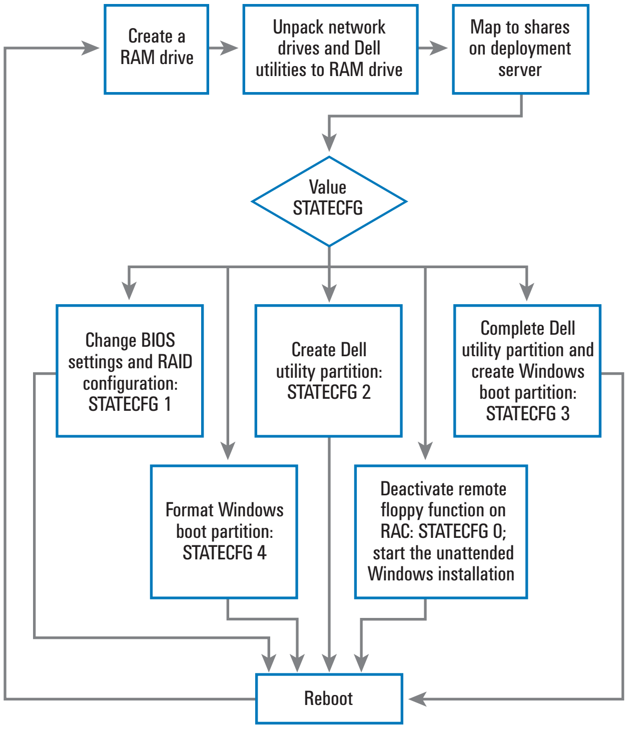

Create a RAM drive
Unpack network drives and Dell utilities to RAM drive
Map to shares on deployment server
Value STATECFG
Change BIOS settings and RAID configuration: STATECFG 1
Create Dell utility partition: STATECFG 2
Complete Dell utility partition and create Windows boot partition: STATECFG 3
Format Windows boot partition: STATECFG 4
Deactivate remote floppy function on RAC: STATECFG 0; start the unattended Windows installation
Reboot

Figure 3. Remote Floppy Boot flow process

Administrators should develop and test the unattended installation setup using only the conventional floppy until the setup works without any problems. Then, administrators can use the Dell Remote Floppy Utility to create an image of the bootable floppy disk; the Remote Floppy Utility is included on the Dell OpenManage Systems Management CD and is available after a management station installation in the c:\program files\dell\rac\mt directory. Once created, the image can be uploaded onto the deployment server.

Creating scripts for the unattended installation

Administrators must create several scripts that reside on the bootable floppy image, including dellpre.bat, a script that controls the initial DOS-based hardware configuration, and racsetup.bat, a script that contains the commands to reconfigure the RAC remotely. Guidelines to create the scripts for the boot floppy are contained in the user's guide of the Dell OpenManage Deployment Toolkit. The toolkit also contains script templates to help administrators script the hardware-build portion of an unattended installation.

Before the actual Windows operating system setup starts, some hardware settings for the server must be reestablished, such as changing the boot order and deactivating the Remote Floppy Boot in the RAC (which is controlled by racsetup.bat). After the pre–operating system hardware configuration completes, the unattended Windows installation begins. This process is controlled by an answer file, typically named unattended.txt, which was created during the Server Assistant installation. The answer file is a script that automatically answers questions during the operating system setup, thus eliminating the need for user interaction. Administrators can edit the file to support their remote installation requirements.

If necessary, administrators can create scripts that must be executed after the initial operating system installation. These post–operating system installation scripts can install, for example, service packs, hot fixes, and the Dell OpenManage Server Administrator software.[1]

Optimizing the unattended installation process

The RAC Remote Floppy Boot feature helps to ease the unattended installation of local as well as geographically dispersed Dell PowerEdge systems. The time devoted to creating an auto-installation process may pay for itself in the long run, resulting in greatly improved installation times and a standard server build across the data center as an additional benefit. By performing unattended installations using the RAC, administrators can take advantage of existing technologies to streamline routine procedures and help free IT staff to deal with more critical issues—all helping to reduce the TCO and increase the ROI of an enterprise's IT infrastructure.

Shelley Palmer-Fettig (shelley_palmer@dell.com) is a senior consultant in the Enterprise Solutions Group at Dell. Shelley has been involved in systems management for servers throughout her Dell career, from helping to start a Network Operations Center in IT to direct consulting for Dell Professional Services.

Stephan Maahs (stephan_maahs@dell.com) is a system consultant for the Global Segment and Large Corporate Accounts divisions of Dell, serving as a subject-matter expert in systems management. Previously, he worked at IBM as a system consultant for IBM® Netfinity® servers. Stephan has a degree in industrial engineering from the University of Applied Sciences in Giessen-Friedberg, Germany.

FOR MORE INFORMATION

RAC local and remote command-line interface:
http://premiersupport.dell.com/docs/software/smdrac3/RAC/en/racadmug/racadmbk0.pdf

Dell OpenManage Deployment Toolkit:
http://support.dell.com/filelib/format.aspx?releaseid=r69795

[1] For more information, see "Customizing Unattended Installation of Server Administrator for Windows and Linux" by Kit Lou and Mohammad Dhedhi in *Dell Power Solutions*, May 2003.

Installing the .NET Client for Dell OpenManage IT Assistant 6.5

The user interface for Dell™ OpenManage™ IT Assistant is implemented using the Microsoft® .NET framework beginning with the IT Assistant 6.5 release. Three new methods provide access to the IT Assistant .NET client: local installation on each workstation, remote access using Microsoft Windows® 2000 Server Terminal Services, and browser-based access using Terminal Services.

BY TERRY SCHROEDER AND MARY JEAN RAATZ

Effective with the release of Dell™ OpenManage™ IT Assistant 6.5, the Microsoft® .NET framework is now required to run the IT Assistant client application. In prior releases, the IT Assistant client ran in a browser running Microsoft Java™ Virtual Machine (JVM™).

With the release of IT Assistant 6.5, administrators can no longer simply point a browser to http://*itaservername*/itassistant to launch the IT Assistant client. Instead, administrators must choose from three new access methods, all available using the new .NET client:

- Install the new IT Assistant .NET client locally on each workstation that needs to access the IT Assistant application.
- Install Microsoft Windows® 2000 Server Terminal Services on the IT Assistant server and use Remote Desktop Connection software from each client.
- Install Terminal Services on the IT Assistant server and install Remote Desktop Web Connection software on the server. This provides access to IT Assistant through a Microsoft Internet Explorer browser.

Installing the IT Assistant .NET client locally

Installing the IT Assistant .NET client locally is the simplest method, but also the least secure from an access standpoint, because the administrator can rely on only IT Assistant identification and passwords for authentication. To install the IT Assistant .NET client using the graphical user interface (GUI) installation process, perform the following steps:

1. Insert the Dell OpenManage Systems Management CD.
2. In the Installation dialog box, select “Custom Setup.”
3. Choose the destination directory and click “Next.”
4. Select the IT Assistant Settings option on the Management Station Software screen, and click “Next.”
5. Deselect the IT Assistant Services check box on the Select IT Assistant Settings screen.
6. Click the OK button to return to the Management Station Software screen.
7. Deselect all applications other than IT Assistant, and click “Next.” The Installation Summary screen displays the components to be installed.
8. Click “Next” to continue with the installation. The IT Assistant components will be installed.

To deploy the IT Assistant .NET client application using a silent script, perform the following steps:

1. Visit *Dell Power Solutions* online at http://www.dell.com/magazines_extras, and copy the silent installation options script into a new Windows Notepad file.
2. Search and replace `localhost` with the name of the IT Assistant server.
3. Save the file to disk using a file name with an .iss extension. (In this example, ITAUIonly.iss is the file name.)
4. Copy the contents of the ITA65 subdirectory from the Dell OpenManage Systems Management CD to a network share. This command can also be run from the CD if it is inserted in the client workstation.
5. Execute the following command from the ITA65 subdirectory:

```
setup standalone -s -f1"c:\ITAUIonly.iss"
```

This command assumes that the .iss file is on the client workstation; however, the file can reside anywhere on the network.

Running the IT Assistant client using Terminal Services

To run the IT Assistant client using Terminal Services, first ensure that Terminal Services is installed on the server where the full installation of IT Assistant will be deployed.[1] During the installation process, Terminal Services prompts the administrator to choose either Remote Administration mode or Application Server mode. Remote Administration mode is easier to install but is limited to two sessions and can be used only by members of the Administrator Group. Application Server mode is more flexible but requires a license server to be configured. In addition, the license server must be in installation mode when the IT Assistant user interface is installed.[2]

Once Terminal Services is installed on the server, IT Assistant can be installed on the server. Be sure to follow the proper steps in the *Dell OpenManage IT Assistant User's Guide*[3] based on the mode in which Terminal Services was installed (either Remote Administration or Application Server mode).

A client running Windows XP already has Remote Desktop Connection software installed. For clients that need to run IT Assistant but are not Windows XP clients, install Remote Desktop Connection software locally on each workstation.[4] The IT Assistant client can now be run from Terminal Services.

Running the IT Assistant client from a browser

To run the IT Assistant client from a browser, perform the steps to install Terminal Services as described in the preceding section. Then install Remote Desktop Web Connection software on the server.[5]

From the client workstation, enter the following address in a Web browser: http://*itaservername*/tsweb. Be sure to replace *itaservername* with the name of the server running the IT Assistant Services.

Follow the prompts to establish a Terminal Services connection by logging in from a Web browser, and then open IT Assistant from within the Terminal Services session. This provides access to the IT Assistant interface.

Choosing among options for accessing IT Assistant

The release of IT Assistant 6.5 provides administrators with several choices for accessing the IT Assistant user interface. In addition, two of these access options allow administrators to control the IT Assistant environment through Terminal Services—an approach that is designed to enable organizations to take advantage of Terminal Services security features.

Terry Schroeder (terry_schroeder@dell.com) is an enterprise technologist in the Advanced Systems Group at Dell. He has an M.S. in Library Science and Information Management and a B.S. in Social Sciences from Emporia State University.

Mary Jean Raatz (mary_jean_raatz@dell.com) is a systems consultant in the Advanced Systems Group at Dell. Mary Jean has a B.A. in Information Systems from the University of Northern Iowa.

FOR MORE INFORMATION

Dell OpenManage IT Assistant Discovery Service:
http://www.dell.com/downloads/global/solutions/DiscoveryWhitepaper_Final.pdf

[1] For complete instructions on installing Terminal Services, visit http://www.microsoft.com/windows2000/technologies/terminal. Click "Windows 2000 Terminal Services Licensing FAQ" to learn about Microsoft licensing requirements for Terminal Services. Dell does not provide client access licenses (CALs) for Terminal Services. Administrators must operate within the stated Microsoft requirements by purchasing appropriate CALs.

[2] Administrators should fully understand and follow Microsoft licensing requirements for Terminal Services. For more information, visit http://www.microsoft.com/windows2000/en/server/help/default.asp?url=/windows2000/en/server/help/sag_termsrv_topnode.htm, click "Getting Started with Terminal Services," and then click "Enabling Terminal Services."

[3] The latest version of this document is available from the Dell support Web site at http://support.dell.com.

[4] For more information and to download the installation file, visit http://www.microsoft.com/downloads/details.aspx?FamilyID=a8255ffc-4b4a-40e7-a706-cde7e9b57e79&displaylang=en.

[5] For more information on installing Remote Desktop Web Connection software, visit http://www.microsoft.com/downloads/details.aspx?FamilyID=E2FF8FB5-97FF-47BC-BACC-92283B52B310&displaylang=en.

Using Dell OpenManage Server Assistant 8.x to Optimize Installation of Dell PowerEdge Servers

Dell™ OpenManage™ Server Assistant 8.x provides features designed to improve operating system (OS) installation on Dell PowerEdge™ servers. These features include network adapter teaming, OS installation replication, network download installation, and support for customized installation scripts. This article explores how these features can help system administrators deploy various operating systems on Dell PowerEdge servers.

BY MICHAEL E. BROWN, NIROOP GONCHIKAR, NATHAN MARTELL, AND GONG WANG

Dell™ OpenManage™ Server Assistant 8.x is a redesigned tool that can help provide seamless operating system (OS) installation for Microsoft® Windows®, Red Hat® Linux®, and Novell® NetWare® software on Dell PowerEdge™ servers. Although this tool has traditionally been used in small business environments, additional features—including network adapter teaming, OS installation replication, network download installation, and support for customized installation scripts—may benefit medium and large businesses as well. This article presents typical enterprise scenarios that help illustrate the benefits of these features to system administrators, and explains the process steps involved in implementing these features.

Making OS installation simple and efficient

To set up a new server in a network environment, administrators typically must go through a lengthy series of steps, which may include:

- Obtaining a list of devices installed on the server and downloading device drivers from the vendor's Web site.
- Downloading and upgrading RAID controller firmware with the latest compatible version for the desired OS.
- Configuring and creating a RAID container using the RAID controller's utility. Then, downloading the compatible RAID controller's driver before starting the OS installation on the newly created container. *Note:* Without the compatible driver, the OS installation is likely to fail.
- During OS installation: Going through a few interview pages (which can be tedious) to configure the desired OS.
- After OS installation: Installing the latest service pack and loading drivers for the network adapters, video adapter, and other devices onto the server until they all can function in the OS.
- Manually adding the server to the network domain.

This entire process can be time-consuming and complex, depending on the OS type and version as well as the dependencies that exist between versions of various service packs, firmware, and drivers.

Dell OpenManage Server Assistant was designed to make this process more streamlined and less time-consuming. The Dell OpenManage Server Assistant CD ships with every Dell PowerEdge server. This tool has two key functions: The first and primary function—known as install mode—is to provide tools to configure the PowerEdge server and install the OS. The install mode can be activated by booting the server with the Server Assistant CD. Server Assistant then provides a mechanism to optimize the installation of the selected OS. It also facilitates a uniform, simple configuration of RAID controllers and automatically installs the correct driver set for all supported devices, without downloading or guesswork.

The second function of Dell OpenManage Server Assistant—known as service mode—is to deliver the latest versions of the BIOS, firmware, diagnostics, and optimized drivers. In this mode, administrators can insert the Server Assistant CD into a system running Microsoft Windows or Red Hat Linux. If the server's auto-play feature is enabled (Windows only), a screen will appear that leads administrators to create the diskette(s) of the latest-version firmware, drivers, device utilities, and diagnostic tools for the selected platform and OS from the Server Assistant CD. Using these diskettes, administrators can install or update an already-running PowerEdge server.

For more information about key Dell OpenManage Server Assistant functions, visit *Dell Power Solutions* online at http://www.dell.com/magazines_extras.

New features in Dell OpenManage Server Assistant 8.x

In addition to the traditional functionality provided in previous Server Assistant releases, Dell OpenManage Server Assistant 8.x introduces several new features to meet the needs of real-world enterprise scenarios: express versus advanced mode for OS configuration; network adapter teaming; service pack or RPM™ (Red Hat Package Manager) deployment through network installation; customized installation scripts; and OS replication. These new features are illustrated in the following sections.

Configuring servers in express or advanced mode

In Dell OpenManage Server Assistant 8.x, configuration pages for RAID, network, and OS are split into two modes: express mode and advanced mode. In express mode, the basic configuration tasks can be completed more quickly. Figure 1 shows the differences between these two modes for Server Assistant configuration interview pages.

	Express mode tasks	Advanced mode tasks
RAID configuration	• Select RAID level • Configure RAID size (Requires two steps)	• Select RAID level • Select array disks • Configure RAID size • Configure stripe size, stripe depth, and read/write policy (Requires four steps)
Network configuration	• Configure IP settings for each network adapter: – DHCP or static IP – For static IP, configure IP address and subnet mask (No network adapter teaming)	• Configure one or more network adapter teams • Select network adapter teaming service provider (Intel or Broadcom), if applicable • Select network adapters for teaming • Select network adapter teaming mode (Adaptive Load Balancing or Adapter Fault Tolerance, for example) • Supply network adapter team name • Configure IP settings (Advanced mode is not available for NetWare)
OS configuration for Microsoft Windows platforms	Provide basic OS information: language, organization, user name, product ID, system name, installation directory, license type, and client license number; workgroup or domain name; and Domain Name System (DNS) server, gateway, and Windows Internet Naming Service (WINS) server	In addition to providing basic OS information: • Select OS components (IIS, DNS, DHCP, and WINS) • Configure SNMP settings • Configure service pack installation over the network • Create customized installation script
OS configuration for Red Hat Linux platforms	Provide basic OS information: system name, domain name, firewall setting, language, root password, DNS server, and gateway	In addition to providing basic OS information: • Configure additional RPMs for installation over the network • Create customized installation script
OS configuration for Novell NetWare platforms	Provide basic OS information: language, server name, Internet domain name, DNS server, and gateway; as well as NDS installation information such as tree name, server context, administrator context, and administrator name and password	In addition to providing basic OS information: • Select OS components: NetWare FTP server, DNS/DHCP service, Web server, wide area network (WAN) traffic manager services, Novell Distributed Print Services™ (NDPS®), multimedia server, Web search, and news server • Configure SNMP settings

Figure 1. Tasks that can be performed using Dell OpenManage Server Assistant 8.x: Express mode versus advanced mode

Teaming network adapters for improved performance and reliability

Network adapter teaming is a mechanism that enables administrators to bond two or more physical network adapters to a single logical network device. The teaming method is designed to increase server uptime and performance. Dell OpenManage Server Assistant 8.x helps administrators to configure network adapter teaming quickly and easily during the OS setup. The tool is designed to enable uniform configuration of network adapter teaming regardless of the type of network adapter installed. Using Dell OpenManage Server Assistant 8.x, administrators are not required to download extra utilities or use complicated vendor-specific interfaces to configure teaming after the OS installation. Furthermore, the tool is designed to enable administrators to set up network adapter teaming without being required to understand its technical underpinnings.

In Dell OpenManage Server Assistant 8.x, administrators can set up network adapter teaming, through the advanced mode network configuration page, by selecting one of the service providers: Broadcom or Intel. Broadcom® Advanced Server Program is available only for Red Hat Linux. Intel® PROSet is available for both Microsoft Windows and Red Hat Linux. Overall, teaming requires two or more network adapters. Administrators may team network adapters from different vendors as long as one adapter in the team is from the vendor that is providing the teaming functionality.

Intel adapter teaming

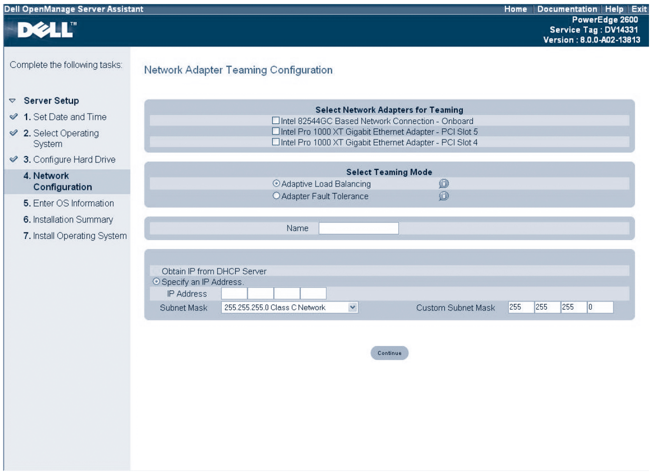
Intel PROSet offers two teaming modes: Adaptive Load Balancing and Adapter Fault Tolerance. In Adaptive Load Balancing mode, a team of two or more network adapters can be configured to help increase the system's transmission and reception throughput. This setup is designed to increase the server's network throughput by allowing transmission over two to eight ports to multiple destination addresses, while incorporating adapter fault tolerance. Only the primary adapter receives incoming traffic and transmits broadcast/multicast packets and non-routed protocols packets.

In Adapter Fault Tolerance mode, a team of two or more network adapters is configured to provide a backup network connection. This setup is designed to permit mixed models and mixed connection speeds as long as at least one Intel network adapter is in the team. A failed primary adapter will pass its Media Access Control (MAC) address and Layer 3 address to the secondary, or *failover*, adapter—this capability is designed to help ensure server availability to the network.

Figure 2 shows the configuration page for Intel PROSet network adapter teaming in Dell OpenManage Server Assistant 8.x. Administrators must assign a static IP address for the logical teaming device.

Broadcom adapter teaming

When administrators select Broadcom as the service provider in Dell OpenManage Server Assistant 8.x, Smart Load Balancing is the only available teaming mode. Smart Load Balancing is designed to offer fault-tolerant teaming and load balancing. All adapters in the team have separate MAC addresses. Operating on Layer 3 addresses, this mode can help balance both incoming and outgoing traffic loads. If a link on any port fails, Smart Load Balancing is designed to divert traffic automatically to other ports in the team.


Dell OpenManage Server Assistant
Home | Documentation | Help | Exit
DELL™

PowerEdge 2600
Service Tag : DV14331
Version : 8.0.0-A02-13813

Complete the following tasks:

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

Network Adapter Teaming Configuration

Select Network Adapters for Teaming
□ Intel 82544GC Based Network Connection - Onboard
□ Intel Pro 1000 XT Gigabit Ethernet Adapter - PCI Slot 5
□ Intel Pro 1000 XT Gigabit Ethernet Adapter - PCI Slot 4

Select Teaming Mode
● Adaptive Load Balancing
○ Adapter Fault Tolerance

Name

○ Obtain IP from DHCP Server
● Specify an IP Address.
IP Address
Subnet Mask 255.255.255.0 Class C Network
Custom Subnet Mask 255 255 255 0

Figure 2. Dell OpenManage Server Assistant 8.x Network Adapter Teaming Configuration page for Intel adapters

Installing service packs or RPMs from the network

One of the most crucial steps following an OS installation is to apply any service packs or kernel updates that have been released. This step is designed to ensure that any problems identified in the OS, such as security risks and critical bug fixes, are resolved—creating a robust OS environment. The longer an OS remains up and running without service pack or kernel updates, the greater the potential risk of a vulnerability being exploited or a critical failure causing server downtime.

A new feature in Dell OpenManage Server Assistant allows administrators to specify a network location from which service packs (Windows systems) or RPMs (Linux systems) can be downloaded and installed. Administrators can pull the service packs or RPMs from either a Network File System (NFS) or Server Message Block (SMB) share. Dell OpenManage Server Assistant downloads the service packs or RPMs to a temporary partition before the OS installation starts, and applies the service packs or RPMs immediately after the OS installation is complete. This feature helps administrators minimize the exposure time of known security and functional risks. It requires that the server be connected to a network that supports Dynamic Host Configuration Protocol (DHCP).

For a Windows OS, Server Assistant supports only the installation of service packs. For Linux, administrators can use a wildcard ("*") to install all the RPMs in a specific directory. This capability is designed to allow for kernel updates as well as the installation of any applications and utilities that can be installed with RPMs.

Customizing installation scripts

Following an OS installation, administrators often standardize the OS environment by adding users, configuring directories, or even mapping network share drives. The Custom Install Script feature available through the advanced mode of the OS interview page in Dell OpenManage Server Assistant is designed to provide a convenient method for configuring a server with settings tailored for an organization's environment. This feature, coupled with the OS Installation Replication feature highlighted in the next section, can be a powerful tool to deploy the OS to Dell PowerEdge servers with uniform and customized configurations. Administrators can use the Custom Install Script feature by going to the advanced mode of the Enter OS Information page and entering the script in the provided text box. This feature is available only on Windows and Linux.

When installing Windows, the Custom Install Script is executed toward the end of the Windows installation process, during the graphical user interface (GUI) portion of the installation through the cmdlines.txt file. For this reason, Dell best practices do not recommend that applications or utilities requiring full OS support be executed through the Custom Install Script feature. For more information about the Custom Install Script feature, see "Automating Server Installation and Upgrade" in the *Windows 2000 Server Resource Kit* (http://www.microsoft.com/ windows2000/techinfo/reskit/en-us/default.asp?url = /windows2000/techinfo/reskit/en-us/deploy/dgcb_ins_jqxv.asp).

When installing Linux, Dell OpenManage Server Assistant executes the Custom Install Script after the OS installation is complete, during the first OS boot. Dell OpenManage Server Assistant automatically adds the Custom Install Script to the /etc/rc.local script and then removes the Custom Install Script after execution. *Note:* The Custom Install Script must include only commands that terminate. For example, commands in Linux such as `ping 192.168.1.1` should not be used and will stall the OS installation from completing the first boot. For sample Linux and Windows Custom Install Scripts, visit *Dell Power Solutions* online at http://www.dell.com/magazines_extras.

Replicating the OS image for multiple servers

Dell OpenManage Server Assistant 8.x is designed to allow administrators to set up multiple Dell PowerEdge servers with identical configurations. After going through the interview pages for one installation, administrators can create a replication profile and use that profile to help automate OS installations on other systems that are configured identically. The OS replication feature is designed to enable administrators to save time creating RAID, configuring OS settings, setting up network adapter teaming, and so on. Dell OpenManage Server Assistant 8.x can support OS replication on Microsoft Windows, Red Hat Linux, and Novell NetWare.

To replicate an OS installation, administrators must first select the check box for "Save Profile for Replication" on the Installation Summary page. Server Assistant will then save the installation profile to the location on the hard drive indicated on the Summary page. Then, administrators must copy the profile files to a diskette after the OS installation is complete. Administrators boot the server to be replicated using the Server Assistant CD and insert the diskette as soon as the system starts to boot from the CD. A text message indicates the progress of the replication, and administrators can follow the instructions on the screen to complete this process.

Deploying servers with express installation

When using Dell OpenManage Server Assistant to set up a server, a novice administrator can take advantage of the express mode pages with no need to access the advanced mode pages. The express process requires administrators to enter only the minimal settings to install and run the desired OS. Expected input would be RAID type and size, basic IP address information, and necessary OS configuration information for the installation to occur.

	Initial time to install OS	Subsequent time to install OS*
Novice user and Windows 2000 Server (express mode)	10.5 minutes	3.5 minutes
Small-to-medium or large business and Windows Server 2003 (advanced mode)	14.5 minutes	8 minutes

**Assumes administrator becomes more familiar with tool on subsequent installations*

Figure 3. Time required for administrator to navigate Dell OpenManage Server Assistant 8.x when installing a Windows OS on a Dell PowerEdge server

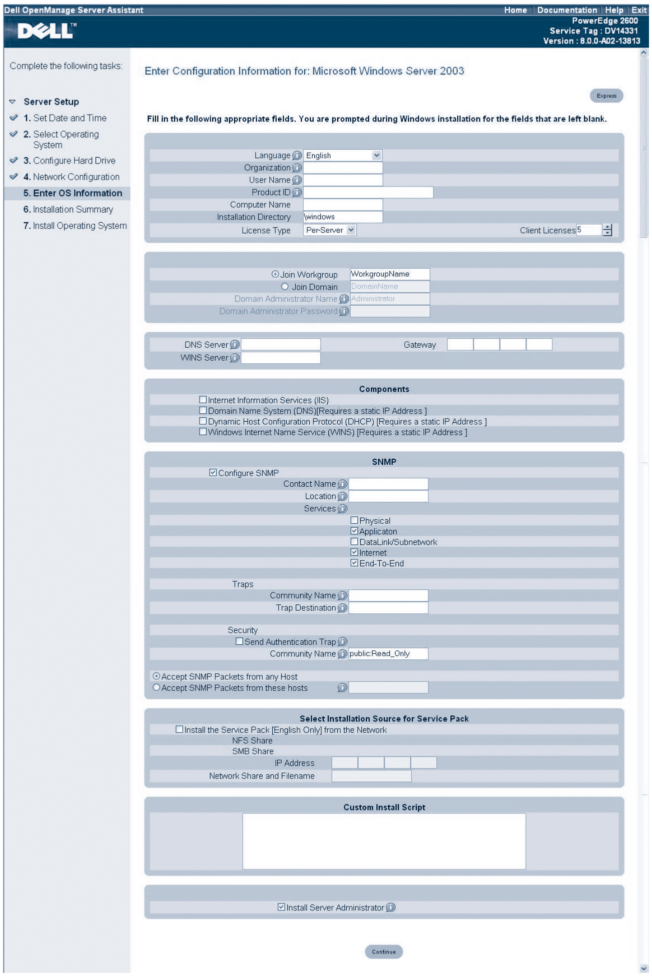
Figure 3 shows the results of a Dell usability evaluation report conducted in December 2003 that measured the time required for the Server Assistant interview pages to complete during Windows installations.[1] After these interview pages completed, the remaining unattended process for the Microsoft Windows Server™ 2003 OS, for example, required approximately 30 minutes.

Setting up server operating systems with advanced installation

For small-to-medium enterprises, a typical OS installation requires more configuration or more options than a novice user environment. In many cases, only certain drives are used to build a RAID array, especially if the system has six or more drives available. A network adapter team may also be required to help provide greater fault tolerance or improved network performance. Simple Network Management Protocol (SNMP) must be configured to send alert traps for system status. Other services such as Microsoft Internet Information Services (IIS), DHCP, and Terminal Services also may need to be installed on the server.

Under such circumstances, administrators must use the advanced mode pages for RAID configuration, network adapter configuration, and OS configuration in Dell OpenManage Server Assistant. To build an advanced RAID configuration, administrators should enter the RAID type; select the drives to be included in the RAID array; and then enter the size, name, stripe size, and read/write policies of the array. To set up a network adapter team, administrators must select the provider (Intel or Broadcom, if applicable); the adapters to be grouped; the type of team (Adapter Fault Tolerance or Adaptive Load Balancing, for example); and the team name and IP settings. On the advanced mode OS configuration page, administrators must enter SNMP configuration settings—such as community, rights, and destination address—and select the desired services to install. Figure 4 shows the Dell OpenManage Server Assistant 8.x GUI for the advanced mode

[1] Based on the Dell OpenManage Server Assistant usability evaluation report performed by Dell Labs in December 2003 on a PowerEdge 2600 server running Microsoft Windows 2000 Server and using RAID-0 and on a PowerEdge 2600 server running Microsoft Windows Server 2003 and using RAID-5, three hard drives, adapter teaming, and SNMP. The required time for installing an OS depends on processor speed, memory size, RAID level, and so on. Actual performance will vary based on configuration, usage, and manufacturing variability.



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

PowerEdge 2600
Service Tag : DV14331
Version : 8.0.0-A02-13813

Complete the following tasks:

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

Enter Configuration Information for: Microsoft Windows Server 2003

Express

Fill in the following appropriate fields. You are prompted during Windows installation for the fields that are left blank.

Language: English
Organization:
User Name:
Product ID:
Computer Name:
Installation Directory: \windows
License Type: Per-Server
Client Licenses: 5

● Join Workgroup: WorkgroupName
○ Join Domain: DomainName
Domain Administrator Name: Administrator
Domain Administrator Password:

DNS Server:
WINS Server:
Gateway:

Components

□ Internet Information Services (IIS)
□ Domain Name System (DNS)[Requires a static IP Address]
□ Dynamic Host Configuration Protocol (DHCP) [Requires a static IP Address]
□ Windows Internet Name Service (WINS) [Requires a static IP Address]

SNMP

☑ Configure SNMP
Contact Name:
Location:
Services:
□ Physical
☑ Applicaton
□ DataLink/Subnetwork
☑ Internet
☑ End-To-End

Traps
Community Name:
Trap Destination:

Security
□ Send Authentication Trap
Community Name: public:Read_Only

● Accept SNMP Packets from any Host
○ Accept SNMP Packets from these hosts

Select Installation Source for Service Pack

□ Install the Service Pack [English Only] from the Network
NFS Share
SMB Share
IP Address:
Network Share and Filename:

Custom Install Script

☑ Install Server Administrator

Continue

Figure 4. Dell OpenManage Server Assistant 8.x advanced mode Enter Configuration Information page for installing Windows Server 2003

Enter Configuration Information page in a Windows Server 2003 environment.

An advanced Server Assistant installation offers several advantages over both a manual installation and an express installation: flexible and easy RAID configuration even though the settings are more advanced; no need to install advanced network adapter teaming software and then learn the interface for setting up a team; easier SNMP setup; and additional time savings for network adapter teaming and SNMP configuration.

Advanced installation in large enterprises

The most advanced features of Server Assistant are designed for large enterprises. For example, administrators in large organizations are most likely to take advantage of the network download feature, customized installation scripts, and OS replication. Administrators can use the network download feature to install a service pack or an RPM during the OS installation. Service packs or RPMs can be installed automatically through the Dell OpenManage Server Assistant 8.x OS advanced mode. Later installations using OS replication can be completed with very little time invested, because a replicated installation skips the attended interview question pages and goes straight to the OS installation. OS replication also helps ensure that subsequent systems are installed with identical system settings.

Particularly in large enterprises, administrators may need to implement security patches, system policies, and standard application software across all the systems in their network. On the advanced mode OS configuration page in Server Assistant, a customized installation script is designed to provide a method for performing certain actions at the end of the installation process before the OS boots for the first time. This capability can help reduce the time required to set up a new system that complies with the enterprise's network security policy and software usage requirements.

The OS replication feature is designed to help administrators who are setting up multiple identical systems minimize the time they spend installing the OS. All pertinent information for the initial OS installation is saved to files, which can then be copied to a diskette and used in conjunction with the Server Assistant CD to facilitate another OS installation on the same system (in case of catastrophic failure) or a system configured with identical hardware. This capability can help significantly reduce system setup time, especially if several servers must be deployed in the same or a similar fashion.

Streamlining OS installation for efficient systems management

Enhanced features in Dell OpenManage Server Assistant 8.x are designed to help organizations of all sizes save time and thus lower costs. When deploying a Dell PowerEdge server on a network, administrators can use Server Assistant to help streamline the OS installation process—helping make the first step in systems management an efficient and well-executed one.

Michael E. Brown (michael_e_brown@dell.com) 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.

Niroop Gonchikar (niroop_gonchikar@dell.com) is a software engineer in the Dell Enterprise Software Development Group. Before joining Dell, he interned at Lucent Technologies. Niroop has a B.S. in Computer Science from the University of Virginia.

Nathan Martell (nathan_martell@dell.com) is a software developer in the Dell Enterprise Software Development Group. Before joining Dell, he worked as a software developer for Netopia, Inc. Nathan has a B.S. in Computer Engineering from the University of Kansas.

Gong Wang (gong_wang@dell.com) 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 an instructor at Wuhan University. Gong has an M.S. in Human-Computer Interaction and an M.S. in Experimental Psychology from Georgia Tech.

Planning Considerations for Intel Extended Memory 64 Technology on Servers and Workstations

Intel® Extended Memory 64 Technology (EM64T) is designed to expand the memory addressing capabilities of the 32-bit Intel architecture (IA-32) while the IA-32 systems continue to run the vast number of existing 32-bit x86-based applications. This article examines 32-bit, x86-based workstation and server applications whose performance may be improved by implementing EM64T. In addition, this article defines specific application scenarios in which the 64-bit Intel Itanium® architecture may help provide better performance than x86-based Intel systems running EM64T.

BY JOHN COOMBS AND JOHN FRUEHE

The recently announced Intel® Extended Memory 64 Technology (EM64T) is designed to increase the memory addressing capabilities of the 32-bit Intel architecture (IA-32) while the IA-32 systems continue to run existing 32-bit, x86-based applications. Intel Xeon™ and Pentium® processor–based systems that incorporate EM64T are expected to become available in the second half of 2004. EM64T-enabled systems are designed to fill the memory addressability gap between highly concurrent 64-bit Itanium® processor–based database and high-performance computing (HPC) servers and Pentium processor–based desktop systems and servers running 32-bit office productivity, Web, and file serving applications. EM64T can help enable Dell™ Precision™ workstations running Linux® operating systems with 64-bit addressing or Microsoft® Windows® XP to create larger and more detailed models and analyze or simulate more complex environments. Dell PowerEdge™ servers incorporating EM64T have the potential to help increase performance on memory-intensive applications as well as to help provide rich, layered commercial and information services to client and Web environments, which have become much more demanding and sophisticated over the last ten years.

The contrast between 32-bit and 64-bit architectures

The memory address ceiling of mainstream Intel processor–based architecture was last raised in 1985, when x86 memory addressing was increased from 16 bits to 32 bits. Before 1985, the 16-bit Intel 80286 processor allowed 2^{16} bytes (representing 65,536 memory locations) to be directly addressed—enough to catalog the population of a small city. In 1985, the 32-bit 80386 processor, enabled by the 32-bit Microsoft Windows NT® and Microsoft Windows 95 operating systems, was designed to address nearly 2^{32} bytes (representing roughly 4.3 billion memory locations)—enough to catalog every person in the world based on the reported global population in 1975.

Bits	Binary	Number of memory addresses (equivalence in bytes)	Relative scale
4	24	16	
9	28	256	
10	210	1,024 (1 KB)	
16	216	65,536 (64 KB)	Population of a small city
32	232	4,294,967,297 (4 GB)	Reported population of the world in 1975
64	264	18,446,744,073,709,600,000 (18 exabytes)	Estimated number of grains in a cubic mile of sand

Figure 1. Differences in scale between architectures

Understanding how addressability can scale may be helpful when considering what applications can best take advantage of 64-bit architecture. Although 64-bit architecture might initially appear to only double memory capacity, moving from 32 bits to 64 bits actually squares the available memory, providing a theoretical limit of 4.3 billion times 4.3 billion addressable memory locations. The resulting 18 exabytes[1] of memory addressability theoretically provides the capability to individually name each grain in a cubic mile of sand (see Figure 1).

Applications that may benefit from 64-bit memory addressing

In Intel Xeon, Itanium, and certain Pentium processors, 64-bit addressing is designed to provide a large, flat memory space that is easily addressed by several memory-intensive applications (see Figure 2). Such applications include databases; digital content creation (DCC) and mechanical computer-aided design (MCAD); Monte Carlo simulation; geophysical analysis and mechanical computer-aided engineering (MCAE); and directory, e-commerce, and messaging servers.

Databases. A very large memory address space can allow database administrators to cache all—or a significant part—of an Oracle® or Microsoft SQL Server™ database in buffer RAM. This approach can enable many more in-memory processes and help provide greater scalability using 64-bit memory addressability as compared to 32-bit addressability. Increased addressability and caching also may help provide improved online transaction processing (OLTP) performance on 64-bit Dell PowerEdge servers, because it helps enable both faster context switching and reduced I/O.

Digital content creation and mechanical computer-aided design. A very large memory address space can help enable designers in the areas of DCC and MCAD to create large 3-D models—in the 6 GB to 8 GB range—and virtual environments on platforms such as the Dell Precision workstation. Compared to 32-bit platforms, the increased memory addressability of 64-bit technology could also enable designers to assemble multiple components—for example, putting a virtual piston into a virtual engine block—quickly and without sacrificing detail simply to fit within the memory limits imposed by a 32-bit Windows process.

Monte Carlo simulation. Server and workstation applications—such as financial risk management and molecular modeling—that employ Monte Carlo techniques to simulate and test a broad range of very detailed outcomes are likely to benefit from the detail and range made possible by 64-bit addressing.

Geophysical analysis and mechanical computer-aided engineering. The 64-bit memory addressing on Intel Xeon and Itanium processor–based servers and HPC clusters can help geophysicists analyze complex environments in much greater detail than is possible when using 32-bit memory addressing. For example, such increased addressing capability may help to improve the likelihood of finding and recovering oil and natural gas.

Compared to 32-bit addressing, the 64-bit addressing capabilities of Itanium processors on Dell PowerEdge servers enable dense meshes for a detailed analysis of airflows around—and effects upon and between—structures in automotive, aerospace, and related industries that rely on MCAE applications. The 64-bit addressing also can help enhance visualization of server data on—and computational steering of server applications from—Intel Xeon and Pentium 4 processor–based Precision workstations, allowing greater interactivity and increased attention to detail than 32-bit variants.[2]

In Intel Xeon, Itanium, and selected Pentium processors, 64-bit addressing is designed to provide a large, flat memory space that is easily addressed by several memory-intensive applications.

Directory, e-commerce, and messaging servers. Directory servers on large corporate intranets can typically store and quickly provide access to more information about the user population when using a large, flat virtual address space. E-commerce servers may benefit from the faster encryption available with dedicated 64-bit registers as well as the larger memory and caches available with upcoming 64-bit Intel Xeon CPUs featuring EM64T, as compared with previous-generation 32-bit Intel Xeon processors. Messaging servers can usually be more easily consolidated and scaled when directories and indexes are enabled by 64-bit

[1] One exabyte = 2^{60} (1,152,921,504,606,846,976) bytes, or 1,024 petabytes.

[2] For more information, see "Windows XP 64-Bit Edition Version 2003 for Itanium-based Systems Technical Overview," Microsoft Corporation, http://download.microsoft.com/download/2/e/b/2eb773ec-944f-4000-a475-d8ee6834af62/WindowsXP64-BitforItanium.doc.

Memory used by example application and data set	64-bit workstation or HPC application example	Vertical market
3.2 GB	Finite element analysis: 5,000 elements with four degrees of freedom	MCAE
6 GB	Automated parking garage design: 500,000 components	MCAD
10 GB	Electronic design: 1,000,000-gate device	Electronic computer-aided design (ECAD)
26 GB	Video editing: typical two-hour movie	DCC
45 GB	Video rendering: one frame of special effects	DCC
56 GB	High-end engineering analysis: 1,000 time steps	MCAE
60 GB	Seismic analysis: multiple shots of one area	Geophysics
80 GB	Geographic survey: determining wireless paths through the city of Vienna with a 3-D geographic information system (GIS)	GIS

Figure 2. Examples of applications and data sets that can take advantage of 64-bit addressing

technology to scale well beyond the 32-bit addressability limit of roughly 4.3 billion messages.

The persistence of 32-bit applications

Many 32-bit applications are not pushing the upper limits of their 4 GB physical or virtual address space. Such applications can continue to run in 32-bit compatibility mode on a 64-bit platform, or in native mode on a 32-bit platform, for the foreseeable future without running out of address space and crashing. In addition, Web, file, and print servers may continue to run comfortably with a 32-bit operating system on 32-bit hardware, or in 32-bit mode with a 64-bit operating system and CPU, for many years to come. Similarly, very few, if any, of today's 32-bit consumer applications running on desktop or notebook PCs test the memory limits of a 2 GB to 3 GB virtual Windows process or push the 4 GB physical memory address limitation.

Some sophisticated users and independent software vendors (ISVs) use Intel and Microsoft extensions to the IA-32 platform such as Physical Address Extension (PAE) or Address Windowing Extension (AWE) to circumvent limitations of 32-bit addressing. These segmented memory extensions, which are similar to those that helped enable 1 MB memory addressing on the 16-bit Intel 80286 processor, work in some application environments, such as SQL Server, and not others—for instance, certain MCAD and MCAE applications. PAE and AWE are neither straightforward nor easy to use, and tend to be vendor- and application-specific. In addition, the processing power required to access and manage memory addresses beyond 4 GB can limit scalability because the processing cycles that PAE and AWE consume to address the additional memory can degrade application response as well as system performance. In short, PAE and AWE remain interim solutions on the path to industry-standard 64-bit workstations and servers.

Applications on EM64T versus Itanium servers

As EM64T becomes available in the second half of 2004 and beyond, distinctions may be drawn between the applications that are appropriate for EM64T and the applications that are appropriate for Itanium processor–based servers. Intel Xeon and Pentium 4 processor–based servers supporting EM64T are designed to run the spectrum of 32-bit applications. Next-generation Intel Xeon processors with EM64T, fast frontside buses, and fast clock cycles are designed to run 32-bit applications that require high native 32-bit performance. At the same time, they are intended to support the transition from 32-bit to 64-bit for Windows and Linux server applications by enabling 32-bit and 64-bit programs to run on the same hardware platform.

Cost- and performance-sensitive server applications that benefit from large data caches may also be appropriate for EM64T. Mixed application environments—for example, a 64-bit database running with a transaction monitor, virus checker, and other middleware not yet ported from the 32-bit environment—may also perform better on an EM64T-based server than on a similar 64-bit Itanium system.

For enterprises deploying applications on Itanium platforms, the application is often the first choice in the procurement cycle, followed by the operating system, and finally, the hardware. Organizations tend to buy an Itanium processor–based server to run a single application such as SAP or Microsoft SQL Server. In contrast, infrastructure servers, which tend to be Intel Xeon processor–based, are typically standardized at the organizational level and such servers are purchased as needed and deployed in various locations within the data center. For high-end and midrange servers running applications that are migrated from legacy 64-bit RISC systems, 32-bit x86 compatibility can be less critical.

IT managers deploying performance-oriented, 64-bit concurrent floating-point and database applications may want to consider Itanium-based systems like the two-processor Dell PowerEdge 3250 or the four-processor Dell PowerEdge 7250. Because many large e-commerce, enterprise resource planning (ERP), supply chain management (SCM), data warehousing, and business intelligence applications are largely concurrent, these applications are well suited to the Itanium architecture. Many Java™ Virtual Machines (JVMs) used in e-commerce, such as BEA® WebLogic JRockit™, can also benefit from just-in-time compilation, predication, and speculative loading—all features of the Itanium architecture. High-end, floating-point-intensive HPC applications such as reservoir simulation and computational fluid dynamics may also benefit from improved performance by taking advantage of the large address space, large caches, and concurrent throughput of an Itanium processor–based server.

The migration process from 32-bit to 64-bit

Best practices suggest that IT managers considering a migration from 32-bit to 64-bit architecture—or from RISC systems to 64-bit industry-standard systems—first perform a thorough needs analysis. Administrators must also work with their operating system and application vendors to discuss the timing of 64-bit product and feature availability, so they can plan their 64-bit hardware acquisitions to coincide with important software releases as 64-bit Intel Xeon, Pentium 4, and Itanium processor–based environments rapidly mature.

> Dell recommends that IT managers contact their operating system and application vendors early in the planning process when considering a 64-bit migration.

A thorough needs analysis can help determine which applications can take advantage of 64-bit architecture immediately, which applications can wait, and which applications will likely not migrate in the foreseeable future, if at all. For example, large database environments—which can benefit from the concurrent performance and large address space of a 64-bit Itanium processor–based platform—should target 64-bit platforms for additional performance and scalability. Application servers, which have less-stringent performance requirements, are more likely candidates for an EM64T migration. Meanwhile, client systems in even the largest and most complex database environments may have enough address space to perform adequately on the 32-bit platform.

Administrators may discover that large address spaces in themselves are not always advantageous. Benefits across various database environments are not uniform. Some data warehousing and decision-support applications perform full-table scans—bypassing data stored in buffer RAM—or invalidate tables frequently. In such cases, buffering data in memory may result in a performance penalty. Yet data warehousing and decision support may run considerably faster in a 64-bit Itanium environment because of the concurrency enabled by Itanium servers when they do not buffer a great quantity of data.

Dell recommends that IT managers contact their operating system and application vendors early in the planning process when considering a 64-bit migration. Many software vendors' 64-bit plans are just emerging, and others' plans are changing with the recent announcement of EM64T-based solutions. Compiled versions of either EM64T-based or Itanium processor–based 64-bit applications are required, so organizations may also continue to run 32-bit versions of applications while executing a planned migration to an EM64T-based or Itanium processor–based server.

Organizations must also be prepared for potential migration challenges, particularly with respect to the recently announced and yet-to-be-shipped EM64T-based systems. When 64-bit EM64T-based workstations and servers are released, both 32-bit and 64-bit versions of Linux and 32-bit versions of Windows will be available—but not 64-bit production versions of Windows. Although 32-bit operating systems will run on 64-bit hardware, they are designed to run only 32-bit applications. To run on 64-bit hardware with a 64-bit operating system, 32-bit and 64-bit applications will both require 64-bit device drivers. Drivers for the most pervasive devices may become available toward the end of 2004 and other drivers will arrive over time, in a manner analogous to that experienced with Windows NT in the first months after it shipped.

Important considerations for 64-bit migration

Intel Extended Memory 64 Technology is designed to greatly expand the memory addressing capabilities of x86-based workstations and servers while they run 32-bit x86-based applications. Enterprises running memory-intensive design, simulation, or analysis applications on 32-bit Intel processor–based workstations may want to consider migrating to 64-bit EM64T-based workstations beginning in the second half of 2004. IT managers with memory-intensive database, e-commerce, and messaging applications running on Intel Xeon or Pentium 4 processor–based servers may prefer to begin procuring EM64T-compatible hardware as soon as it is available and then plan a migration of operating systems and applications as these components become available from the software vendors. The robust support of Intel, Microsoft, and Dell for a mixed 32-bit and 64-bit environment can help smooth the migration to 64-bit architecture in both workstation and server environments. Managers of RISC-based UNIX® server environments, particularly servers dedicated to a single, large application, may consider migrating to Itanium processor–based servers because applications are available today.

John Coombs (john_coombs@dell.com) is a product planning senior consultant in the Dell Precision workstation business unit. John has an M.B.A. from the University of Chicago and a B.A. in Economics from Brown University, and has been in the technology field for 23 years.

John Fruehe (john_fruehe@dell.com) is a marketing strategist for the Dell Enterprise Product Group. He has worked at Dell for more than seven years; prior to that, he was at Compaq and Zenith Data Systems. John has a B.S. in Economics from Illinois State University and has been in the technology field for 13 years.

FOR MORE INFORMATION

Intel EM64T:
http://www.intel.com/technology/64bitextensions

An Introduction to the Intelligent Platform Management Interface

The Intelligent Platform Management Interface (IPMI) is the standards-based systems management interface used by Dell™ PowerEdge™ servers. This article introduces the base specifications of IPMI technology and discusses revisions made to IPMI in its 1.5 and 2.0 versions.

BY JORDAN HARGRAVE

Providing consistent cross-platform systems management functionality has historically challenged IT administrators. The wide variety of hardware and operating systems initially required system vendors such as Dell to create proprietary systems management solutions, which usually involved Simple Network Management Protocol (SNMP) agents. These SNMP agents also required each system vendor to develop plug-ins for enterprise management consoles such as the HP® OpenView® or BMC Software® Patrol® applications. The disadvantage: vendors had no common model for providing access to general system information such as service tags, BIOS versions, or system type.

A history of management headaches

In the early 1990s, the Distributed Management Task Force (formerly the Desktop Management Task Force), or DMTF, attempted to remedy the cross-platform systems management problem. This organization—created by Dell, Intel, Compaq, and HP—aimed to develop software-based specifications for systems management. Its first specification, the Desktop Management Interface (DMI), provided a large set of predefined object classes for a system, including hard disk, memory, CPU, and system chassis. DMI also included common objects for system events such as disk failure and chassis intrusion.

Any product that provided a standard DMI agent could be managed by a console that understood DMI. DMI 1.0 originally allowed querying only the local system; DMI 2.0 added the ability to query a system remotely. The next specification from the DMTF was the Common Information Model (CIM), which added object-oriented features, thereby allowing a greater reuse of code for common objects. This interface is currently used by Windows Management Instrumentation (WMI). The disadvantage of these purely software-based implementations, however, is that the server operating system (OS) must be up and running; if the system crashes, administrators cannot determine the root cause of the failure.

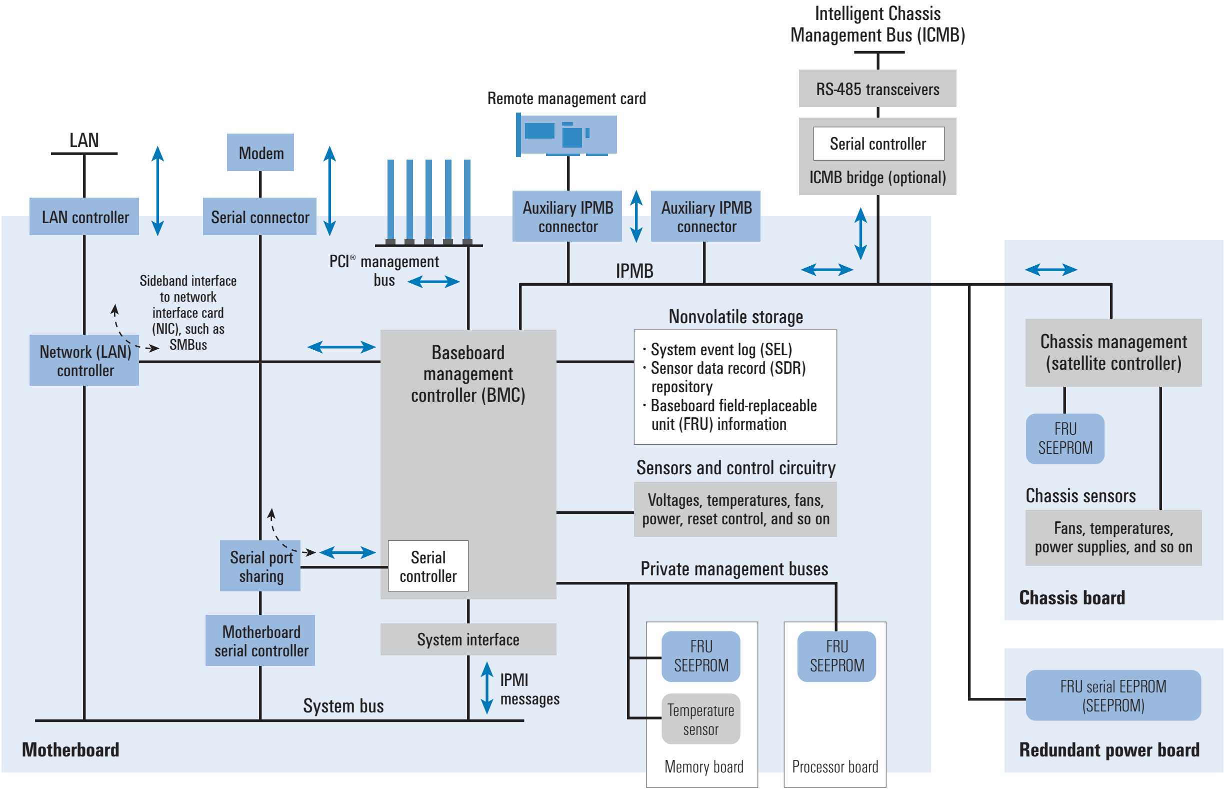


Figure 1. The IPMI architecture

Adding hardware to the mix

As systems became more powerful, they required increasingly robust remote systems management capability. This functionality was transferred from the software level to add-in or embedded controllers. Initial offerings from Dell, such as the Dell Remote Access Card (DRAC), were proprietary products that added new features to servers, such as a remote console and remote floppy boot. These cards can operate even if a server loses power or the OS crashes. They also can manage a server in a pre-OS state, such as during power-on self-test (POST) and BIOS configuration. Such hardware-based systems management implementations are complementary to the software agents. DMI/SNMP agents can provide most of the sensor and system information, while the card performs failure analysis at a remote console. The disadvantage of these cards is their proprietary nature—the agents and tools must be rewritten for each new release.

A new direction in systems management

In 1998, Intel, Dell, HP, and NEC developed the Intelligent Platform Management Interface (IPMI) as a specification for providing systems management capability in hardware. The IPMI specification provides a common message-based interface for accessing all the manageable features in a system. IPMI includes a rich set of predefined commands and interfaces organized by type of operation, such as reading temperature, voltage, fan speed, and chassis intrusion. Methods are also provided for accessing the system event log (SEL), hardware watchdog, and power control.

IPMI replaces or abstracts previous methods of accessing sensors through the systems management bus (SMBus) or intelligent interface controller (i2c). The IPMI specification is expandable, allowing original equipment manufacturers (OEMs) to define their own commands or sensor types. Many companies such as AMD, Fujitsu, and QLogic have developed products that use IPMI. Several releases and enhancements have been made to IPMI since its original 0.9 specification; the latest release is 2.0.

Understanding how IPMI works

The Baseboard Management Controller (BMC) is the heart of an IPMI-based system—it is responsible for monitoring and controlling all the manageable devices in the system. Figure 1 illustrates the central role of the BMC within the IPMI architecture. All access from the host OS is routed through the BMC, which can talk to other IPMI-aware devices in the system through the Intelligent Platform Management Bus (IPMB). Several vendors have developed inexpensive embedded BMC solutions, providing a cost-effective method for administrators to add IPMI support to a system. These controllers reduce server CPU usage by offloading system polling from the CPU.

The BMC contains nonvolatile RAM (NVRAM) storage for the SEL, sensor data records (SDRs), and asset information. The SDR area describes sensors that may be connected to the system. It stores information such as the sensor name, location, and thresholds. The BMC also is responsible for sending and handling events. These events can be thresholds that have been exceeded or triggers such as chassis intrusion. Actions to handle the events can include logging, power cycling, or issuing an SNMP trap (IPMI 1.5 only).

Several tools can be used for querying IPMI information on a system. The OpenIPMI project is an open source initiative to develop a suite of IPMI management utilities. It currently addresses only Linux® operating systems but may soon become available for other platforms. Intel has provided a set of reference IPMI drivers for Microsoft® Windows®, Novell® NetWare®, and UnixWare® operating systems.

Integrating IPMI into Dell PowerEdge servers

As one of the four original IPMI promoters, Dell is committed to using IPMI as its management interface for the Dell™ PowerEdge™ server line. Most of the current sixth-generation PowerEdge servers—the 1650, 2600, 2650, 6600, and 6650 models—support the IPMI 1.0 specification. The PowerEdge 1750, 3250, 7150, and 8450 also support IPMI. Dell OpenManage™ Server Assistant agents use IPMI to query server health, retrieve event logs, and perform power-control operations. The Dell Embedded Remote Assistant (ERA) and DRAC III cards communicate with a PowerEdge server through the IPMB. Dell plans to support IPMI 1.5 in future seventh- and eighth-generation PowerEdge servers.

The future of IPMI

Development of the IPMI specification is ongoing. The IPMI 1.5 specification, the latest approved specification, adds support for accessing the BMC through the serial port or a local area network (LAN). The LAN and serial connectors can be either dedicated to the BMC or multiplexed with the system connectors. This feature allows a server to be completely controlled from a remote system through the network or a modem. The advantage of this feature is that no agents are required on the remote system. Existing tools that already use IPMI can be easily modified to support the new transport methods, because the message interface remains the same. The LAN transport uses the Remote Management Control Protocol (RMCP), which employs User Datagram Protocol (UDP) datagrams. The server also can automatically send event notifications either through the serial or LAN port as SNMP traps, or to a pager.

The IPMI 1.5 specification has added features for controlling the system boot order. This allows the remote system to boot to a utility partition or to boot through the Preboot Execution Environment (PXE). Another new feature is terminal mode, which provides a remote text-mode console that can be used for viewing the BIOS configuration screen or server crash dump.

Furthermore, the IPMI 2.0 specification is currently under review. This version adds enhanced features for authenticating and encrypting the LAN/serial connection.

A new framework for cross-platform systems management

IPMI is an important step in the evolution of server management. It is available in a wide range of server and system platforms and can help reduce total cost of ownership by providing a consistent interface across systems. The large set of common commands and the ability to manage a failed system remotely can provide powerful tools for IT administrators.

Jordan Hargrave (jordan_hargrave@dell.com) is a senior software engineer at Dell. He has a bachelor's degree in Mathematics and Computer Science from Carnegie Mellon University.

FOR MORE INFORMATION

Intelligent Platform Management Interface:
http://developer.intel.com/design/servers/ipmi

OpenIPMI project:
http://www.sourceforge.net/projects/openipmi

Understanding Performance Benchmarks

Benchmarks provide objective information that can be used to compare computer platforms, components, operating systems, and specific system configurations. This article discusses characteristics of credible benchmarks, guidelines for evaluating benchmark results, and some of the main benchmarks used at Dell for assessing the performance of server, workstation, and client systems.

BY SHARON HANSON, DIEGO ESTEVES, AND CLINT ESPINOZA

At their best, performance benchmarks provide impartial information that can be used to evaluate and compare the performance of computer systems. Dell and the computer industry promote objective and credible benchmarking in various ways, including participation in standards bodies such as the Standard Performance Evaluation Corporation (SPEC), Business Applications Performance Corporation (BAPCo), Transaction Processing Performance Council (TPC), and Storage Performance Council. When properly run and documented, the benchmarks produced by these and other groups help provide objective information that can be used to compare computer platforms, components, operating systems, and specific system configurations.

Dell is committed to furthering industry practices that yield objective industry-standard benchmark results. Organizations can use these benchmarks to evaluate and compare Dell™ systems to competitors' systems. Dell also uses the benchmarks when developing new products and assessing new technologies.

The Dell benchmark philosophy is based on three tenets:

- Benchmark in a way that closely resembles how organizations use applications on Dell systems
- Ensure that anyone can reproduce results with a system shipped directly from Dell, using publicly available drivers
- Promote benchmark and run-rule changes that reflect this approach to benchmarking

This article discusses characteristics of credible benchmarks and presents high-level guidelines for evaluating benchmark results. It concludes with a list of the key benchmarks used at Dell to evaluate server, workstation, and client system performance.

Characteristics of credible performance benchmarks

A computer performance benchmark is a standard by which a computer system can be measured and judged. Many of the well-known benchmarks are developed and regulated by standards organizations such as SPEC and BAPCo. Just as common are unregulated benchmarks that measure system performance when running specific applications such as Adobe® Photoshop®, Microsoft® Exchange, Parametric® Pro/E®, or Id Software® Quake III® software. These benchmarks can help administrators evaluate system performance on a single, critical application such as Pro/E or Microsoft Exchange. Such benchmarks can be run—and their results reported—with varying degrees of flexibility.

In contrast, regulated benchmarks tend to have well-defined and documented methodologies, and their results are documented and reproducible. A good example is the SPEC® CPU2000 benchmark, which is produced by SPEC, a nonprofit corporation. According to SPEC, the organization's mission is to establish, maintain, and endorse a standardized set of relevant benchmarks. SPEC develops suites of benchmarks and also reviews and publishes submitted results from member organizations and other benchmark licensees.[1] The SPEC organization has industry-wide representation and its benchmark suites are well accepted and credible.

The SPEC CPU2000 benchmark provides performance measurements that can be used to compare compute-intensive workloads (both integer and floating point) on different computer systems. These compute-intensive benchmarks measure the performance of a system's processor, memory architecture, and compiler. CPU2000 consists of a set of objective tests that must be compiled and run according to SPEC run rules. SPEC provides the benchmarks as source code so they can be compiled to run on a variety of platforms, including industry-standard Intel® architecture–based systems and SPARC® processor–based Sun™ systems.

Those who are evaluating benchmarks should consider whether the benchmark workload is reasonably representative of the real-world applications that will be run on the system.

In addition, SPEC provides guidelines for legitimately optimizing the performance of tested systems on the benchmark. These guidelines are designed to ensure that the hardware and software configurations of tested systems are suitable to run real-world applications. The organization also requires a full disclosure report, which provides benchmark results and configuration details sufficient to independently reproduce the results. SPEC encourages submission of reports for publication on the SPEC Web site (http://www.spec.org). These reports undergo a peer-review process before publication. Because of these rigorous requirements, CPU2000 benchmark results that are published on the SPEC Web site are widely used to compare the CPU, memory, and compiler performance of client and server systems.

BAPCo, TPC, and the Storage Performance Council are also nonprofit corporations that provide industry-standard benchmarks widely used to compare the performance of client, server, and storage systems. TPC was founded to define transaction processing and database benchmarks. The BAPCo charter is to develop and distribute a set of objective performance benchmarks based on popular computer applications and industry-standard operating systems. The goal of the Storage Performance Council is to define, promote, and enforce vendor-neutral benchmarks that characterize the performance of storage subsystems.

Guidelines for evaluating benchmark results

When using benchmark results to evaluate and compare systems, administrators should understand the benchmark, be aware of system optimizations, and ensure comparable system comparisons, as follows.

Understand the benchmark

It is essential to understand which aspects of system performance a benchmark is testing as well as what the system's workload will be. Those who are evaluating benchmarks should consider whether the benchmark workload is reasonably representative of the real-world applications that will be run on the system. For instance, if a client system will be used to run mainstream business productivity applications, the BAPCo SYSmark® or Ziff Davis® Business Winstone® benchmarks are good candidates.[2] On the other hand, if the test subject is a workstation system that will be used primarily to run Pro/E, the Pro/E application benchmark is suitable. If possible, those who are evaluating benchmarks should focus on regulated benchmarks from standards bodies such as SPEC and BAPCo or on benchmarks that are standard industry applications.

Application benchmarks can be run with a variety of inputs, each of which attempts to represent different usage scenarios. For example, Adobe Photoshop performance varies greatly depending on the size of the image and the operations performed on it. Moreover, some Photoshop operations may be better suited or optimized for a particular system architecture. Even within a particular operation (such as the Gaussian Blur filter), the end user may be able to modify how the filter is applied. Different code algorithms may be used, resulting in significantly different performance results. These variables make it relatively easy to create a suite of Photoshop benchmark operations that greatly favor a particular system architecture. For this reason, Dell recommends that organizations look beyond summary benchmark results to help ensure that the operations performed are representative of their specific usage models.

Be aware of system optimizations

Some optimization of the tested system is expected and allowed on all benchmarks. SPEC outlines broad optimization guidelines in its run rules for each benchmark. The expectation of these guidelines

[1] For more information about SPEC, visit http://www.spec.org.

[2] For more information about the BAPCo SYSmark benchmark, visit www.bapco.com; for more information about the Ziff Davis Winstone benchmark, visit http://www.veritest.com/benchmarks/bwinstone/default.asp.

is to avoid optimizations that are so extreme as to render the system unsuitable for real-world applications. For example, when running SPEC benchmarks, Dell often uses publicly available compilers that support new CPU features. These features can improve system performance and better demonstrate the capability of Dell systems. This practice conforms to the spirit of the SPEC guidelines. The compilers are publicly available to software developers to use in building their own applications; therefore, the benchmark results are representative of possible real-world applications.

In contrast, it is not uncommon for a benchmark to be run on a system that has been specially tuned to do well on the benchmark. Such tuning can be so extreme that the benchmark results are neither credible nor useful. Even regulated benchmarks can be misused in this way, so it is important that the benchmark results include complete configuration information for the tested system.

Workload type	Benchmark
Business productivity	• SYSmark 2004 • Content Creation Winstone 2004 • Business Winstone 2004
Mainstream 3-D performance	• Futuremark® 3DMark®
CPU, memory subsystem, and compiler	• SPEC CPU2000 • Linpack
Gaming	• Quake III • Epic Games® Unreal Tournament 2003 • Ubisoft™ Splinter Cell®
Portable computer battery life	• Business Winstone 2002 BatteryMark® • BAPCo MobileMark® 2002
3-D graphics	• SPECviewperf® 7.1
Mechanical computer-aided design (MCAD)	• SPECapc for Pro/ENGINEER™ • NASTRAN
2-D graphics	• Photoshop • Autodesk® AutoCAD®

Figure 2. Typical client system benchmarks

Ensure comparable system comparisons

When comparing the benchmark results of systems from multiple vendors, test engineers should ensure that the tested systems and their benchmark settings are comparable. This requires organizations conducting benchmark tests to supply adequate documentation for system and benchmark configurations.

Benchmarking at Dell

Dell uses benchmarks throughout the technology assessment and systems development process to help ensure that Dell server and client systems provide the appropriate balance of performance, features, cost, quality, and reliability. Dell supports industry efforts to standardize performance benchmarks and is an active participant in all the standards bodies discussed in this article. Figures 1 and 2 list key benchmarks that Dell uses to evaluate the performance of server and client systems.

When used appropriately, benchmarks can provide valuable information that can help administrators compare and evaluate computer systems. In addition to benchmarks, many factors should weigh heavily in the evaluation process, including features, support, and price, as well as the ability to service, upgrade, and manage the system under consideration.

Workload type	Benchmark
Database	• Online transaction processing (OLTP): TPC-C • Decision support: TPC-H and TPC-R • Java™: SPECjbb®
Messaging	• Microsoft Exchange: MAPI (Message Application Programming Interface) Messaging Benchmark 2 (MMB2) and MMB3 • Lotus Notes®: Notesbench® • Simple Mail Transport Protocol (SMTP)/Post Office Protocol 3 (POP3): SPECmail® 2001
Web services	• Hypertext Transfer Protocol (HTTP): SPECweb®99 • HTTP over SSL (HTTPS): SPECweb99_SSL
File and print services	• Ziff Davis NetBench® • SPECsfs®
Storage	• SPC Benchmark 1™ (SPC-1) • Iometer
CPU and high-performance computing	• SPEC CPU2000 • Linpack • NASTRAN® • STREAM • Hierarchical INTegration (HINT)
Microbenchmarks*	• LMbench • Netperf

**A microbenchmark measures one specific feature of a system isolated from other features.*

Figure 1. Typical server benchmarks

Sharon Hanson (sharon_ hanson@dell.com) is a technical writer in the office of the Dell CTO. She has written and produced Dell white papers and technical articles on industry technology trends for the past eight years. Sharon has a B.B.A. from The University of Texas at Austin.

Diego Esteves (diego_esteves@dell.com) is a systems engineer and consultant currently working on Dell Precision™ workstation SPEC performance and independent software vendor (ISV) application certifications. Diego has a B.S.B.A. from Xavier University in Cincinnati, Ohio. He currently represents Dell on the SPEC CPU subcommittee, the body responsible for the industry-standard SPEC CPU2000 benchmarks.

Clint Espinoza (clint_espinoza@dell.com) is a storage performance engineer specializing in RAID adapter performance. Clint has a B.A. from Trinity University in San Antonio, Texas.

FOR MORE INFORMATION

SPEC: http://www.spec.org

BAPCo: http://www.bapco.com

TPC: http://www.tpc.org

Storage Performance Council: http://www.storageperformance.org

Exploring the Advantages of PCI Express

Parallel encoding schemes such as parallel ATA, SCSI, Peripheral Component Interconnect (PCI®), and Peripheral Component Interconnect Extended (PCI-X®) are being replaced by high-speed serial alternatives based on Peripheral Component Interconnect Express (PCI Express™) architecture. Designed to meet the requirements of today's high-bandwidth applications, PCI Express creates a point-to-point serial architecture that offers key performance, cost, and scalability advantages.

BY CHRIS CROTEAU AND PAUL LUSE

To enable quicker response to business needs, enterprises demand that the IT infrastructure process information faster, access shared data more efficiently, and improve end-to-end throughput. Such requirements are driving developments in processor, storage, and network architectures that can help create more scalable infrastructures by overcoming the inherent limitations of parallel protocols.

Serial architecture can help administrators accomplish business objectives by improving the performance and expanding the capabilities of network connections, drive interconnect protocols, and primary I/O buses for both server and desktop computing platforms. The Peripheral Component Interconnect Express (PCI Express™) specification helps pave the way by providing a system interconnect for computing and communications platforms. This article explores the concept of serialization, focusing on the PCI Express architecture and its potential to enable next-generation computing platforms for high-bandwidth enterprise applications.

Overcoming the limitations of parallel encoding

Serial protocols replace parallel encoding schemes at the physical layer with high-speed serial encoding. Already fundamental to certain data communications protocols—such as Fibre Channel, Asynchronous Transfer Mode (ATM), and Ethernet—serial architecture has recently emerged in areas where parallel encoding schemes have been prevalent for the past 15 years. The protocols for parallel ATA, SCSI, and Peripheral Component Interconnect (PCI®)—all of which are based on parallel encoding schemes—share many of the same characteristics and limitations, including high pin count, low frequencies, and shared buses. Serial technology can help mitigate each of these issues:

- **High pin count:** Serial protocols do not need to transfer a large amount of data in each clock cycle, so they require fewer data lines than parallel schemes. Fewer data lines fit into smaller cables, which potentially can lower costs.

- **Low frequencies:** Unlike serial data transfers, parallel transfers are frequency limited because of skew factors. Skew results from the requirement to synchronize all arriving data lines with one another at the receiving end of a parallel transfer. Lower frequencies translate into lower performance when the bus width is limited by physical and electrical considerations.
- **Shared buses:** Serial encoding schemes employ a point-to-point topology, whereas parallel encoding usually requires multiple devices to access a shared set of data lines. In the latter case, resource contention can lead to suboptimal performance.

Examining the move toward serial architecture for ATA and SCSI

With the emergence of Serial ATA and Serial Attached SCSI, the transition of disk protocols from parallel to serial architecture is well under way. Serial ATA, a high-speed point-to-point link for ATA, was created to address performance-limiting design issues in the ATA specification. Serial ATA increases bandwidth and simplifies topologies. Originally conceived as a desktop architecture, Serial ATA helps to improve performance and promises economic benefits that have prompted storage designers to consider it for enterprise applications.[1]

Serial ATA enhances the ATA specification with capabilities such as hot-plug drive swapping and native command queuing, which were not included in the parallel specification. By overcoming the signal timing issues inherent in parallel ATA, Serial ATA allows for longer cable lengths—and for narrower cable widths, which can help improve airflow within a system enclosure.

SCSI also is undergoing a parallel-to-serial conversion that promises new capabilities in addition to performance gains. Serial Attached SCSI, the latest advance in SCSI technology following Ultra320 SCSI, offers serial point-to-point links that replace the parallel bus—and thus reduce the overhead—of today's parallel SCSI topologies. Serial Attached SCSI is expected to be scalable to more than 16,000 physical devices in a single domain. In addition, it offers backward compatibility with legacy SCSI drivers and software by maintaining compatibility with the SCSI protocol. Because Serial Attached SCSI interoperates with Serial ATA devices, administrators can choose from a variety of drive types.

Migrating the I/O bus from PCI and PCI-X to PCI Express

The primary I/O bus is currently transitioning from a parallel to a serial architecture that is similar to ATA and SCSI. The PCI Express specification was developed to help minimize I/O bus bottlenecks within systems and to provide the necessary bandwidth for high-speed, chip-to-chip, and board-to-board communications within a system.

Parallel PCI has served as the dominant I/O bus for more than a decade. However, high-bandwidth applications—some available today and many anticipated in the future—and usage models place demands on CPUs, I/O devices, and the I/O bus that neither PCI nor Peripheral Component Interconnect Extended (PCI-X®) is equipped to meet, largely because of their parallel encoding architectures. For example, a high-volume, low-cost server is often used to support an I/O-intensive video-on-demand system. Such systems are required to serve multiple streams of time-dependent data concurrently. Specifications to provide this type of capability are built into PCI Express. Besides using serial encoding to help overcome the limitations of a parallel protocol, PCI Express offers the following features:

- **Cost-effectiveness:** Works with high-volume, low-cost components such as standard chips and connectors
- **Performance across market segments:** Meets the requirements of mobile, desktop, server, and communications applications
- **Scalability and life span:** Scales to accommodate future applications through its capability to aggregate links to gain additional bandwidth, and is designed with a life span to rival that of PCI
- **Compatibility:** Offers compatibility with previous PCI and PCI-X specifications and programming models
- **Data protection:** Helps provide improved data integrity and error handling

Overview of PCI Express architecture

More than an evolution of the PCI and PCI-X bus interface, PCI Express is a new, layered architecture that retains PCI and PCI-X usage and software programming models. PCI Express architecture is a point-to-point serial interconnect that uses low-voltage differential signaling (LVDS). One new component in the architecture is a switch that replaces the parallel I/O bus of PCI and PCI-X. Devices connect to this switch through point-to-point connections called *links,* which consist of one or more lanes. A lane is a set of differential signals used to transmit data. PCI Express bridges to PCI and to PCI-X are other critical components that will help facilitate the adoption of PCI Express.

The PCI Express architecture design offers the flexibility to aggregate lanes into higher-bandwidth links. A single lane, referred to as a ×1 link (pronounced "by one link"), can help support up to 250 MB/sec in each direction. For example, a ×4 link can help support up to 1 GB/sec in each direction (simplex) or 2 GB/sec in both directions (dual simplex). Figure 1 outlines the potential peak bandwidth capabilities projected in the PCI Express specification as lane widths are widened.

[1] The Serial ATA II Working Group is developing a specification to double overall bandwidth for Serial ATA to 3 Gbps.

Lane width	Peak simplex bandwidth	Peak dual-simplex bandwidth
×1	250 MB/sec	500 MB/sec
×2	500 MB/sec	1 GB/sec
×4	1 GB/sec	2 GB/sec
×8	2 GB/sec	4 GB/sec
×16	4 GB/sec	8 GB/sec

Figure 1. Potential bandwidth comparisons for PCI Express architecture

Figure 2 shows a simple PCI Express topology, identifying key elements within the architecture: the root complex and the end point, which are roughly equivalent to a PCI or PCI-X host bridge and a PCI or PCI-X device, respectively. Bridges can accommodate interaction of PCI and PCI-X devices with PCI Express devices in the same system.

PCI Express architecture also delivers features—such as advanced error logging, advanced reporting, power management, and quality of service—that system architects require to meet the demands of an enterprise environment. Because it requires fewer pins and signals than PCI and PCI-X to support each link, PCI Express architecture provides a smaller slot connector and smaller form factor that can help make system layout simpler and more cost-effective.

Role of PCI Express architecture layers

PCI Express is a layered architecture (see Figure 3). The software layer is responsible for the creation of a stable operating environment. This layer includes services such as enumeration and configuration of PCI Express devices, and the allocation of resources such as memory and interrupts. These types of services remain unchanged from those defined in PCI and PCI-X. Software compatibility with the PCI and PCI-X models is a key feature of PCI Express because it helps enable existing operating systems to boot PCI Express without software modifications. In addition, the runtime software model is distinguishable from PCI and PCI-X only in that future enhancements will be designed to leverage features unique to PCI Express.

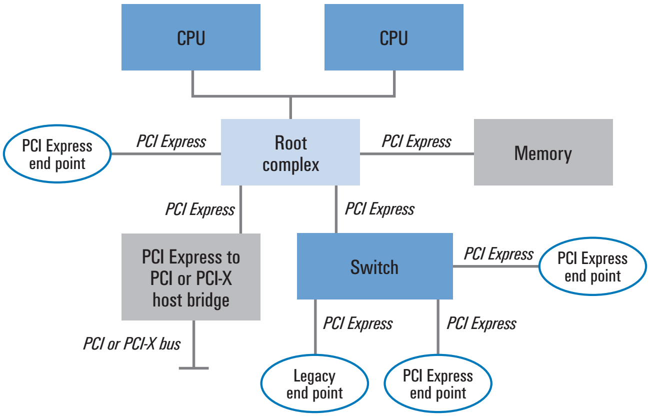
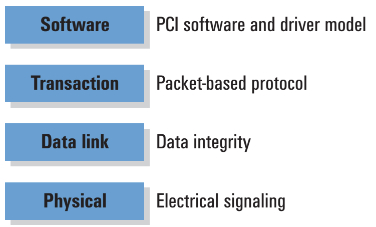


Figure 2. Simple PCI Express topology



Software	PCI software and driver model
Transaction	Packet-based protocol
Data link	Data integrity
Physical	Electrical signaling

Figure 3. PCI Express layered architecture

The transaction layer of the PCI Express architecture is responsible for the creation of outgoing request packets and for the completion of incoming packets. In addition, the transaction layer handles some aspects of flow control and power management. For a transmit operation, the transaction layer would field a read or a write—from the software layer, for example—then create the corresponding packet and pass it to the data link layer. For a receive operation, the transaction layer would accept data from the data link layer and complete the request in the operating system or an application in the software layer. Although all transactions in PCI Express are split transactions—transactions that are completed in multiple phases, which enables multiple transactions to be open or outstanding at one time—some do not require a response. Split transactions are one of the key features contributing to the efficiency of PCI Express.

The data link layer plays a critical role in PCI Express: it helps ensure that data is properly ordered across each link. A link cyclical redundancy check (LCRC) ensures data integrity, and sequence numbers handle the ordering of packets in the data link layer.

The physical layer comprises physical components required to configure and maintain communications across a link. These include mechanisms for link training, data scrambling, 8B/10B encoding, packet framing, and signaling the data onto the link.

Taking IT infrastructure to the next level

Serial technology already has become rooted in enterprise storage and networking architectures, and the current wave of migration from parallel to serial encoding schemes is gaining momentum throughout the IT infrastructure. Using PCI Express as a serial I/O interconnect can help reduce the overhead for enterprise systems and bus bandwidths by using low-cost components, while helping protect legacy software investments through backward compatibility.

Chris Croteau (christopher.c.croteau@intel.com) is a market development director at the Intel® Communications Group, Storage Components Division. He is responsible for market creation and the promotion of Intel building blocks within server and storage market segments.

Paul Luse (paul.e.luse@intel.com) is a senior technical marketing engineer specializing in silicon storage components at the Intel Communications Group, Storage Components Division. He has eight RAID-related U.S. patents pending.

Designing and Optimizing

Dell/EMC SAN Configurations Part 1

Dell/EMC storage area networks offer storage architects and administrators a wide range of design options and optimization settings. In the first of a two-part series, members of the Dell™ Server and Storage Performance team present collective best practices for designing logical storage units and optimizing storage processor cache settings.

BY ARRIAN MEHIS AND SCOTT STANFORD

Dell/EMC storage area networks (SANs) support various workload demands—from transactional databases and file sharing to media streaming and backup servers. This article examines how Dell/EMC SANs can benefit from optimized, well-planned storage processor (SP) and logical storage unit (LUN) designs.

Settings and configurations for SPs and LUNs can be interdependent or independent of each other and often affect back-end performance, depending on the type of workload, optimization, and design configurations. This close symbiotic relationship between SP and LUN configurations can quickly become complex, so storage architects who strive to design their SANs for maximum application or database performance will benefit from understanding this relationship.

SAN performance is affected by the workload read/write type, size, and activity; RAID group design and LUN allocation; SP cache settings; LUN binding and metaLUN design; and host bus adapter (HBA) performance tuning. This two-part series explores detailed concepts and processes for each of the preceding factors, and explains both the theoretical and practical performance advantages that can be realized by implementing properly tuned and optimized SAN configurations. Part 1 focuses on best practices for designing LUNs and optimizing SP cache settings.

Designing LUNs: Best practices

RAID storage offers large capacity, failover protection, high performance, or various combinations of all three. RAID groups can be subdivided into one or more LUNs; RAID groups represent the physical layer with which the SAN hardware communicates, while LUNs represent the logical layer with which the operating system (OS) communicates.

RAID groups

Figure 1 shows how individual physical disks are incorporated into a single RAID group. EMC® Navisphere® Management Suite—the storage management software for Dell/EMC SANs—supports up to 128 LUNs per RAID group[1] and the following RAID types:

- **RAID-5 (individual access array):** Provides data integrity using parity information that is stored on each disk in the LUN. This RAID type is well suited for multiple applications that transfer different amounts of data in most I/O operations.

[1] EMC Navisphere Management Suite version 12 or higher; for more information about EMC Navisphere, visit http://www.emc.com/products/storage_management/navisphere.jsp.

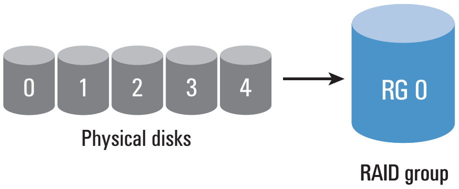


Figure 1. Incorporating five physical disks within one RAID group

- **RAID-3 (parallel access array):** Provides data integrity using parity information that is stored on one disk in the LUN. This RAID type is well suited for single-task applications—such as video storage—that transfer large amounts of data in most I/O operations.
- **RAID-1 (mirrored array):** Provides data integrity by mirroring (copying) its data onto another disk in the LUN. This RAID type provides the greatest data integrity at the highest cost in disk space; it is ideal for an OS disk.
- **RAID-0 (individual access array without parity):** Provides the same individual access features as RAID-5, but does not have parity information. As a result, if a disk in the LUN fails, the information on the LUN is lost. Also, RAID-0 is not technically RAID because the setup is not redundant.
- **RAID-10 (mirrored individual access array without parity):** Provides the same individual access features as RAID-5, but with higher data integrity. This RAID type is well suited for the same applications as RAID-5, but should be used when data integrity is more important than the cost of disk space.
- **Disk (individual disk):** Functions just like a standard single disk and, as such, does not have the data integrity provided by parity or mirrored data. This RAID type is well suited for temporary directories that are not critical.
- **Hot spare (global spare):** Serves as a temporary replacement for a failed disk in a RAID-5, RAID-3, RAID-1, or RAID-10 LUN. Data from the failed disk is reconstructed automatically on the hot spare—either from the parity or the mirrored data on the working disks in the LUN—so data on the LUN is always accessible.

LUN-to-RAID ratio

Although Dell/EMC SANs can support up to 128 LUNs per RAID group, dedicated RAID groups can help maximize performance. Dedicated RAID groups have a one-to-one relationship with a LUN; that is, only one LUN is allocated from the RAID group. The reason for this design is that, for each I/O operation on the LUN, the read or write still must be executed on the physical layer.

Understanding this concept is particularly important for applications such as Microsoft® Exchange Server 2003, because each client-side change (read or write) usually requires or expects a corresponding return acknowledgment. Delays in I/Os resolving to the physical layer or LUN level can cause upstream delays and message queues to build. For example, if three LUNs are bound to one RAID group, depending on the workload imposed on each LUN, all three LUNs could contend with one another because they are tied to the same physical resource. However, if each LUN were bound to its own RAID group, the LUNs would not compete for physical resources. Figures 2 and 3 illustrate how a RAID group can be allocated to one or multiple LUNs.

Binding multiple LUNs to a single RAID group is a common practice in file-sharing scenarios, where slices of capacity from the RAID group are divided among different users or groups that want to share information. File-sharing environments typically receive random accesses slightly similar to database workloads. However, file shares do not deploy or rely upon the delayed-write methodology at the application level, which is used in transactional database implementations.[2] In contrast, database workloads experience random I/O behavior and need fast, efficient flushes, or *commits*, from transaction log buffers to the database volumes. Therefore, storage architects should configure one LUN per RAID group for database workloads to reduce the possibility of LUN contention.

LUN and back-end loop symmetry

Just as implementing the optimal binding ratios for RAID groups and LUNs is critical to preventing LUN contention, so too is the concept of maintaining symmetry between LUNs and SP ownership. When binding LUNs using EMC Navisphere software, administrators have several options for assigning LUN ownership. LUNs can be automatically assigned to an SP or designated to a specific SP.

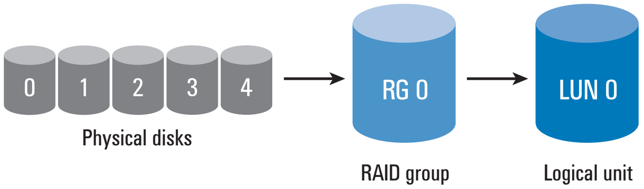


Figure 2. Incorporating physical disks into a RAID group that is mapped to a single LUN (dedicated RAID group)

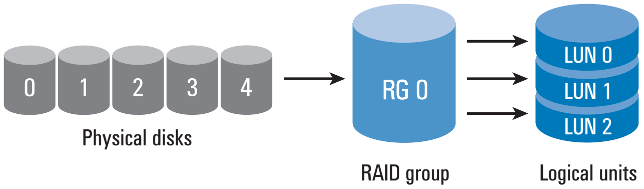


Figure 3. Incorporating physical disks into a RAID group that is mapped to multiple LUNs

[2] In file-sharing implementations, most network operating systems use *lazy writes*, in which writes are committed to the file system on the storage medium based upon predetermined OS settings.

Dell/EMC CX series arrays		Back-end loops (2 Gbps)	Drives per array
Entry-level	CX200	1	30
	CX300	1	60
Mid-level	CX400	2	60
	CX500	2	120
Enterprise-level	CX600	2	240
	CX700	4	240

Figure 4. Back-end loop and drive quantity matrix for Dell/EMC CX series storage arrays

Regardless of what method is used, administrators should ensure that symmetry between LUNs and SP ownership is maintained.

A typical technique used to formulate LUN symmetry is balancing by workload, which helps spread the workload evenly across the SPs and therefore reduce SP latency.[3] If LUN symmetry is not balanced and LUNs are heavily skewed on one SP, client actions originating from a mailbox housed on one SP and destined for a mailbox on a different SP can experience lengthy delays. The ultimate result of such delays can be higher SP cache misses.

All Dell/EMC CX Fibre Channel storage systems include dual SPs and either one or two back-end loops, depending on the model. LUNs should be designed for the best possible balance across the SPs and the back-end loops to help avoid or significantly minimize performance lag caused by overloading the SPs or back-end loops.

Back-end loops are external extensions from the SPs to the disk array enclosures that operate at 2 Gbps bandwidth—the more back-end ports, the more drive support, bandwidth, and I/O capability. Because each SP can be viewed as having two virtual connections to each individual disk on the system, the loops can help to balance workloads across the SPs. For example, the entry-level CX200 and CX300 each have one back-end loop to support additional drives. The mid-level CX400 and CX500 each have two back-end loops. The enterprise-level CX600 also has two back-end loops, but as shown in Figure 4, the next-generation CX700 has four back-end loops supporting the same amount of drives as the CX600.

Sample scenario: Creating symmetry in the SAN

Figure 5 depicts a hypothetical scenario in which a small business deploys a Dell/EMC CX400 storage array. The LUN layout shows an imbalanced symmetry for the SPs and back-end loops. Small, medium, and large workloads are represented respective to the LUN size.

In Figure 5, Storage Processor B (SPb), including both back-end loops (BE0 and BE1), is heavily overloaded with one large workload and two medium workloads. Storage Processor A (SPa) manages three light workloads. Figure 6 shows a more effective balance of LUN and loop symmetry, given the weight and type of workloads for this small business scenario. Although in Figures 5 and 6 each storage processor has two paths to every disk in the enclosure for failover purposes, these are hypothetical scenarios; the LUNs could just as easily span more than one back-end loop, depending on which loop the comprising disks reside.

An existing LUN can be easily shifted to a different SP using the trespass operation in EMC Navisphere. However, because the physical disk permanently resides on its own specific loop, reassigning LUNs to different loops requires destroying the LUN and recreating RAID groups based on physical disk location. To design more effective SP workload symmetry, this process is necessary.

Optimizing storage processor cache settings: Best practices

Cache settings are integral to SP and LUN configurations, specifically cache page size, cache flush watermarks, and cache allocation. Just as reading from or writing to memory is much faster than reading from or writing to disk, cache can help significantly reduce read and write latencies from the SP to and from the disk.

SP cache is subdivided into two types: read cache and write cache. The cache is present in, and also shared by, the SP memory; read and write caches can be enabled or disabled per SP. LUNs can be allowed or disallowed to use SP read and write caches. This capability is helpful when a workload targeted for a specific LUN will not benefit from cache—either read cache, write cache, or both—while the other LUNs owned by the SP will benefit from cache. So, when designing LUNs for optimum response times and low latency, administrators can disable cache on certain LUNs to make more cache available to LUNs that will benefit as well as to reduce unnecessary cache thrashing.

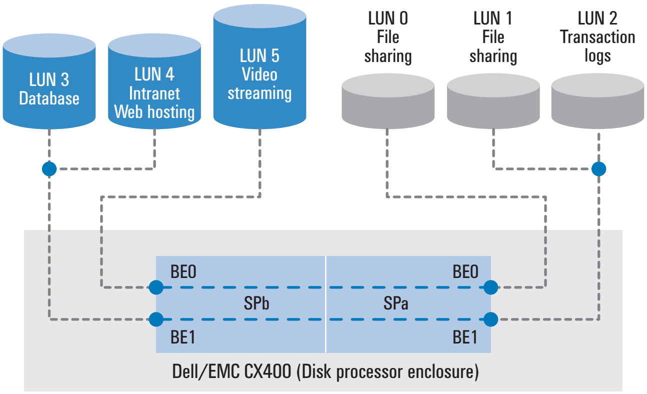


Figure 5. Imbalanced SP workload distribution for hypothetical small-business scenario

[3] EMC PowerPath® software can be used in combination with Navisphere to help provide load balancing among SPs. However, to fully leverage the capabilities of a Dell/EMC storage array, administrators should carefully maintain proper SP-to-LUN symmetry.

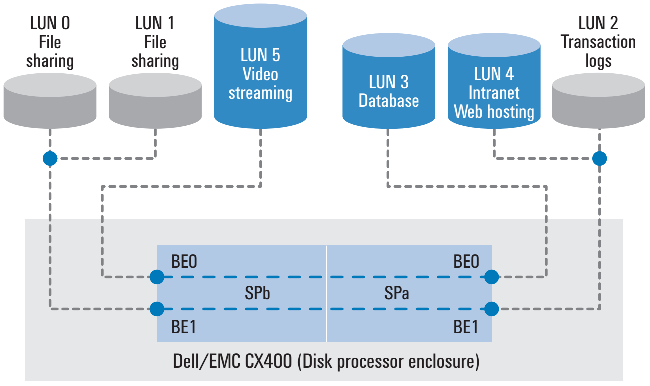


Figure 6. Balanced SP workload distribution for hypothetical small-business scenario

Cache page size

Page size sets the number of kilobytes stored in one cache page. A page is a portion of memory reserved for a specific block of I/O, which is the actual data being sent to or read from the LUN. Unlike physical disk I/O logic controllers, the SPs manage the read and write caches by pages instead of sectors—the larger the page size, the more continuous sectors the cache stores in a single page. EMC recommends the following page sizes:

- **General file-server applications:** 8 KB or 16 KB
- **Database applications:** 2 KB or 4 KB

Cache flush watermarks

A cache watermark determines when a processor flushes its write cache—also known as watermark processing. When an SP flushes its write cache, it writes dirty pages to disk. A *dirty page* is a write cache page with modified data that has not yet been written to disk. The high watermark is the percentage of dirty pages in the write cache; when the high-watermark threshold is reached, the SP begins flushing its write cache. The low watermark also is the percentage of dirty pages in the write cache; when this threshold is reached, the SP stops flushing its write cache. EMC recommends that the high watermark be set at 80 percent and the low watermark be set at 60 percent. Although watermark settings can be adjusted above and below the default recommendations, administrators should fully understand the implications of changing these settings and how those changes can affect cache miss rates or cache thrashing.

Read cache

Read cache relies heavily on prediction and the type of workload to which the LUNs are exposed. Read cache hits are extremely high during sequential reads of contiguous data streams. This type of scenario is typical in file-sharing environments where large amounts of sequential data may stream back and forth from the client to the storage system. In such an environment, administrators should allocate a large portion of the read cache to anticipate a high percentage of cache hits based on the workload.

Cache prefetching

The EMC read-prediction algorithm—also known as read-ahead caching or prefetching—is adaptive in nature. The SP will prefetch data (assuming prefetching is enabled for the corresponding LUN) and fill the read cache only after two sequential reads that share spatial locality. By using this algorithm, the SP assumes that if there were two reads from the same location, there is a high probability that it will need that data from the next sequential location again. The SP will first check the read cache for the data. If the data is not present, it will then check the disk. This algorithm can help reduce SP-to-LUN latency by filling the read cache with prefetched data before it is actually needed.

EMC Navisphere offers three types of cache prefetching:

- **None:** Disables all prefetch properties.
- **Constant:** Prefetches data of a constant length. This type of prefetching is recommended if the read data size is unvarying and regular in length. If selected, only *prefetch size* and *segment size* options are available. Prefetch size is the number of blocks of data to prefetch for one host read request. Segment size is the number of blocks of data to prefetch in one read operation from the LUN. An SP reads one segment at a time from the LUN, because smaller prefetch requests interfere less with other host requests.
- **Variable:** Prefetches data of variable length. If selected, only *prefetch multiplier*, *segment multiplier*, and *maximum prefetch* options are available. The prefetch multiplier determines the amount (in disk blocks) of data to prefetch. For example, if the prefetch multiplier is set to 4 and the amount of data requested is 2 KB, then the variable prefetch size is 8 KB (16 disk blocks). The segment multiplier determines the size (in disk blocks) of the segments that make up the prefetch operation. This option allows the variable prefetch size to be broken into smaller chunks of data, because smaller prefetch requests interfere less with other host requests. For example, if the segment multiplier is set to 2 and the amount of data requested is 2 KB, then the variable segment size is 4 KB (8 disk blocks). The maximum prefetch is the maximum number of disk blocks to prefetch; the default setting is 4,096.

In an environment where reads are sequential, such as a file server, administrators should select variable prefetching. Although Navisphere defaults to a prefetch multiplier of 4 and a segment multiplier of 4, the multipliers can be further tuned for specific needs. For example, if reads are very sequential and constant, such as with media streaming or backup servers, it may be best to select constant prefetching. Depending on the typical read size, regardless of whether variable or constant prefetching is used, the prefetch size may be increased to further bridge the gap between the SP and the cache. This will allow the cache to be loaded with more read data before it is needed by the SP, ultimately decreasing read latency.

In an environment where reads are random, administrators should select variable prefetching, with a prefetch multiplier of 1 and a segment multiplier of 1. That way, the prefetch data traffic will be reduced to avoid interfering with other host requests, but it will still allow for occasional read cache hits with minimal unused prefetched blocks.

In an environment where reads occur rarely (*very* random) or never, such as a database transaction log LUN in which only writes occur, administrators should set the prefetch type to none. Because no reads occur and the cache will never be used by the respective LUN, enabling any type of read-ahead logic to fill the cache is unnecessary. Ultimately, administrators may want to disable read cache at the SP level if read cache hits are seldom to none.

To use any of the read prefetch features, read cache must be enabled on the LUNs before they can benefit from prefetching and SP read cache.

Write cache

The Dell/EMC CX series Fibre Channel storage arrays introduce an additional step to the storage-system writing process, in which writes are performed in the SP write cache before the destination LUN. In a typical database application, new transactions or updates bound for a transaction log file housed on the transaction log LUNs must first be written to the SP write cache. However, the use of write cache applies to all write types, not just random database writes, so any write that takes place on the storage system will write to cache first. Specifically for typical database applications, when the cache is considered dirty, it is flushed to a transaction log file (buffer). Then, when the log file reaches a preset size, all writes are committed to the database.

High-availability (HA) cache vaulting is available in Dell/EMC CX series storage systems. HA cache vaulting determines the availability of storage-system write caching; if enabled, it will disable write caching when a single vault disk fails. Before disabling HA cache vaulting, administrators should consider the resiliency requirements of the application and storage array. Disabling cache vaulting simply allows write caching to continue on a failed drive. HA cache vaulting will have no effect on performance unless a drive fails, in which case the cache *image* will be quickly dumped to disk in an effort to save cached data.

Note: Write cache must be enabled on the LUNs before the respective LUN can benefit from SP write cache.

Advanced cache optimization

After using the previously discussed settings—read and write cache, cache page size, cache watermarks, and prefetching—as a starting point for tuning, administrators can further tune Dell/EMC CX series storage systems using EMC Navisphere Analyzer. Key Navisphere Analyzer counters such as read hit percentage; write hit percentage; write cache flush ratio; dirty pages percentage; throughput and bandwidth; and SP, LUN, and disk utilization provide solid indicators of the effect of each setting on I/O operations or effective bandwidth utilization.

Building a better SAN

LUN design methods and the interrelationships among LUN, RAID, and back-end loop symmetry can have a significant impact on achieving optimal I/O operations with low latencies. Storage processor cache optimizations further reveal how advanced storage settings can be utilized to support typical application workloads.

Performance, capacity, and redundancy are key considerations when determining the optimal storage solution for specific application workloads. EMC default performance settings can provide a solid foundation that can help administrators achieve optimal overall SAN performance. However, maximum performance can be obtained only through a thorough analysis of workload requirements and an understanding of the effects of performance settings and design concepts.

References

EMC Corporation. “EMC CLARiiON Fibre Channel Storage Fundamentals.” EMC Engineering White Paper, October 31, 2003. http://www.emc.com/products/systems/pdf/H1049_emc_clariion_fibre_channel_storage_fundamentals_ldv.pdf.

EMC Corporation. *EMC Navisphere R13 Help Tutorial.*

Arrian Mehis (arrian_mehis@dell.com) is a systems engineer on the Server and Storage Performance team in the Dell™ Enterprise Product Group. His responsibilities include Microsoft Exchange Server single-node performance analysis on Dell server, SCSI, SAN, and RAID solutions. Arrian has a B.S. in Computer Engineering with a minor in Information Systems from the Georgia Institute of Technology.

Scott Stanford (scott_stanford@dell.com) is a systems engineer on the Server and Storage Performance team in the Dell Enterprise Product Group. His current work focuses on Microsoft Exchange Server cluster benchmarking and server/storage performance analysis. 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/EMC:
http://www1.us.dell.com/content/products/category.aspx/storage?c=us&cs=555&l=en&s=biz

Enhancing Backup Performance

with the Second-Generation Dell PowerVault 136T

The second-generation Dell™ PowerVault™ 136T Fibre Channel tape backup library can help organizations improve performance to accommodate limited backup windows. Optimized for shared-storage environments, the high-performance PowerVault 136T can help protect critical data by helping make both backups and data verification feasible within allocated backup windows.

BY RICHARD GOLASKY

As the amount of storage in data centers continues to grow, IT organizations require increasingly efficient tape backup products that can maximize data archival speeds. Without high-speed tape hardware, backup operations may exceed the normally scheduled windows that are allocated for overnight backups. Administrators who arrive at work in the morning to discover that their backup jobs are still running may benefit from considering next-generation tape backup hardware.

The Dell™ PowerVault™ 136T Fibre Channel tape backup library is an enterprise-class library that is designed for use in storage area networks (SANs). Introduced in 2001, the first-generation PowerVault 136T—which incorporates Linear Tape-Open™ (LTO®) technology and Ultrium® 1 LTO (LTO-1) tape drives—can help provide reliable and efficient data protection for corporate data centers that require library-sharing in a SAN environment. The library is designed to maintain as many as six data streams to achieve multiple successful data protection operations within designated backup windows.

The second-generation PowerVault 136T Fibre Channel tape backup library is designed to provide improved performance over its predecessor, including a high-speed backup interface. In addition, the new library provides two main performance-enhancing features: LTO-2 tape drives and the embedded Fibre Channel–to–SCSI storage network controller.

A second-generation LTO-2 tape drive is designed to offer up to twice the performance of a first-generation LTO-1 tape drive. The LTO-2 tape drive is designed to have a native data transfer rate of up to 35 MB/sec and a data transfer rate of up to 70 MB/sec using 2:1 compression.[1] The LTO-1 drive, on the other hand, is designed to have a native data transfer rate of up to 15 MB/sec and a data transfer rate of up to 30 MB/sec using 2:1 compression.[2] Like the first-generation library, the second-generation PowerVault 136T is designed to simultaneously maintain six data streams.

Designed to handle the high-performance throughput of LTO-2 drives, the second-generation PowerVault 136T library is equipped with a second-generation Fibre Channel–to–SCSI storage network controller. While the new Fibre Channel–to–SCSI bridge includes the same functionality as the first-generation bridge, its architecture allows for efficient data streaming to high-performance

[1, 2] For more information about LTO-1 and LTO-2 tape drives, visit http://www.storage.ibm.com/tape/lto/3580/index.html.

LTO-2 tape drives.[3] Architecture enhancements to the storage network controller include the following:

- A 133 MHz Peripheral Component Interconnect Extended (PCI-X®) system interface bus
- A QLogic® ISP2312 dual Fibre Channel controller
- Small form-factor pluggable (SFP) transceivers
- PCI-X to dual channel Ultra320 Low Voltage Differential (LVD) interfaces

Testing the second-generation PowerVault 136T

To understand the level of performance that the new Fibre Channel–to–SCSI bridge and LTO-2 drives in the PowerVault 136T can help provide in a shared storage environment, Dell engineers ran a six-node backup test at Dell labs in April 2004. The team configured a SAN fabric test bed to replicate a six-server SAN using a Dell/EMC CX600 storage array. As shown in Figure 1, the test bed consisted of the following software and hardware components:

- Six PowerEdge 2650 servers running the Microsoft® Windows Server™ 2003 operating system
- QLogic QL2340 and Emulex® LP982 Fibre Channel host bus adapters (HBAs)
- Two Brocade® SilkWorm 3800 Fibre Channel switches
- PowerVault 136T Fibre Channel–attached library with six LTO-2 tape drives
- Dell/EMC CX600 Fibre Channel storage array
- EMC® PowerPath® software for load balancing and fault tolerance

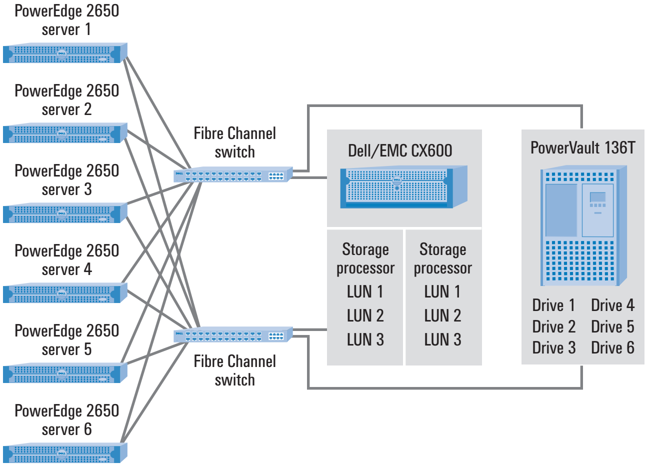


Figure 1. SAN fabric test bed that replicated a six-server SAN using a Dell/EMC CX600 storage array

Each Dell PowerEdge 2650 server was assigned a 170 GB logical storage unit (LUN) consisting of 100,000 noncompressible files that ranged in size from 5 KB to 200 KB. Testers used noncompressible data instead of compressible data to simulate an environment in which imaging and sound files are used. (Such files are usually noncompressible.) Using standard Dell CommVault® Galaxy® version 5 tape backup software, the team installed each server with a media agent software option that allowed Fibre Channel host servers to share a tape backup library.

In the performance tests, three phases were monitored: data backup, verification, and restoration. All performance results were obtained from the tape backup software job statistics.

In the backup test, each server initiated a single backup stream to the PowerVault 136T tape library; the backup library simultaneously maintained six backup streams from the six separate hosts. Once the backup jobs were completed, each server initiated data verification operations in which data segments were read back from the tape media to ensure that the data written to the tape cartridges was identical to the source data. Finally, in a third test, each server initiated a data restoration process in which all data that had been backed up from the server to the tape library in the first portion of the test was restored back to each server storage LUN.

Examining the throughput performance for each of the three test phases in this study can help administrators create appropriate backup schedules when using the second-generation PowerVault 136T in similar scenarios.

Examining the test results

Figure 2 shows performance results for the backup tests. The average backup speed for each server was approximately 29.0 MB/sec. The measured data transfer rate equaled 104.4 GB/hour per backup stream. All together, the six backup streams—each providing a data transfer rate of 104.4 GB/hour—totaled 626.4 GB/hour.

Since only one stream ran from each host server, the Dell team expected each backup stream to transfer data at a similar rate because the Fibre Channel–to–SCSI bridge in the PowerVault 136T is designed to efficiently maintain multiple data streams across its two Fibre Channel interfaces and four LVD SCSI controllers. To help accomplish this efficiency, administrators can implement equal load distribution across the PowerVault 136T by setting up proper channel zoning, which can be configured in the Fibre Channel-to-SCSI bridge. Using PowerVault 136T Storage Networking Controller (SNC) Manager software, administrators can zone any of the four LVD SCSI interfaces to either of the Fibre Channel interfaces on the bridge. This capability allows an administrator to assign a set number of tape drives to each Fibre Channel node.

[3] For more information about the next-generation Fibre Channel–to–SCSI storage network controller, visit http://premiersupport.dell.com/docs/stor-sys/136TLTO2/SNC2/en/index.htm.

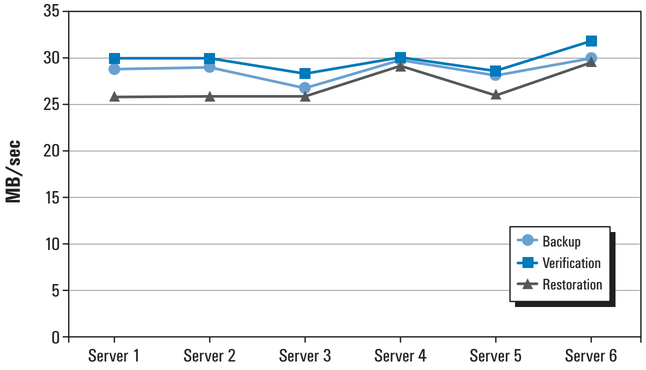


Figure 2. Dell PowerVault 136T test results for data backup, verification, and restoration

To understand this further, consider the following setup for the PowerVault 136T Fibre Channel–to–SCSI bridge:

- **SCSI bus 1:** Robotic controller, tape drive 1
- **SCSI bus 2:** Tape drive 2, tape drive 3
- **SCSI bus 3:** Tape drive 4, tape drive 5
- **SCSI bus 4:** Tape drive 6

Zoning SCSI bus 1 and bus 2 to the first Fibre Channel node will assign the robotic controller and first three tape drives to the first Fibre Channel interface. SCSI bus 3 and bus 4 can be zoned to the second Fibre Channel node, which assigns the last three drives to the second Fibre Channel interface. This type of setup allows for equal load distribution across both Fibre Channel interfaces, thus helping to improve performance.

After backup operations were completed, all six host servers simultaneously initiated data verification operations. Data verification is a process in which data segments on the tape media are read back from the tape drives and checked for inconsistencies against what was recorded in the tape backup software database. Matching segments indicate that the data was successfully written to tape.

The data verification process normally consumes less time than a backup operation because verification incurs minimal system overhead on the file system. During verification, data is not written to any device; instead, data segments are read from the media and compared to recorded information. Therefore, if a backup operation requires three hours to complete, data verification takes less than three hours. Because data verification adds time to the backup process, verification must always be taken into consideration when determining backup windows. Many data centers deactivate data verification because their backup windows are already too narrow, and they cannot afford to use network bandwidth for this purpose during production hours.

Because the data verification process in this test involved data transmission over the Fibre Channel link, the test team measured data transmission performance from the tape library back to the servers. Because considerably less system overhead occurs during data verification than during data backup, Dell engineers expected that the data verification performance would be equal to or higher than data backup performance.

Performance for the data verification process yielded the results shown in Figure 2. As predicted, the data verification process produced slightly higher performance results than those of the backup process. In Figure 2, the data verification performance measured on each host server was approximately 30 MB/sec.

In the third and final test, all six 170 GB data partitions were formatted to simulate a total data loss scenario. Each of the six host servers then initiated a restoration process to restore all 170 GB of data that had been backed up during the backup test. In the data restoration test, all six host servers read data from the PowerVault 136T tape drives and then transferred the data back to the data partitions on the storage array.

The restoration process differs from the verification process in that, during data restoration, all blocks of data must be read back from the tape cartridge, and each file must be reconstructed on the disk. This time-consuming process affects throughput, and thus data restoration operations are usually slower than either verification or backup operations. As shown in Figure 2, restoration performance on each host server was less than either backup or verification performance.

Improving backup speed and data protection

The second-generation PowerVault 136T tape backup library and the Fibre Channel–to–SCSI storage network controller are designed to provide high-speed backup capabilities and help increase the likelihood that data protection operations will be able to complete within the designated windows. The performance boost of the second-generation PowerVault 136T also helps make data verification feasible so that administrators can improve protection for critical data despite shrinking backup windows. Backward compatibility allows the first-generation PowerVault 136T to be easily upgraded to second-generation performance levels by upgrading the Fibre Channel–to–SCSI bridge plug-in module and replacing LTO-1 drives with LTO-2 drives. By upgrading in this manner, or by purchasing a second-generation PowerVault 136T, IT managers can help extend data center backup capabilities.

Richard Golasky (richard_golasky@dell.com) is a development engineer senior consultant in the Dell Enterprise Storage Data Protection Group. His responsibilities include all areas of tape backup, including Fibre Channel, SAN, network attached storage (NAS), and high-availability cluster and performance analysis. He has a B.S. in Electrical Engineering from Florida Atlantic University.

FOR MORE INFORMATION

LTO technology:
http://www.lto-technology.com

Moving Toward Storage Virtualization with Dell/EMC MetaLUNs

Storage virtualization enables IT managers to provision, manage, and expand storage capacity to individual host systems with minimal impact to production. Implementation of storage virtualization in the EMC® Navisphere® Management Suite is based on metaLUNs, which build on the mature, robust foundation of Flare™ logical storage unit (LUN) provisioning. Using the EMC approach to storage virtualization, IT managers can create separate types of storage groups depending on the desired level of protection and performance within the array.

BY MICHAEL L. GIERHART

Effective in release 12 of EMC® Navisphere® Management Suite—the array management software for Dell/EMC Fibre Channel arrays—EMC implemented an enhancement to the basic logical storage unit (LUN) and coined the term *metaLUN* to describe it. MetaLUNs provide enhanced LUN virtualization at the array level. Release 12 of EMC Navisphere Management Suite maintains the data protection, performance, and flexibility of previous-generation Navisphere software while introducing powerful metaLUN functionality.

In a Dell/EMC storage array, metaLUN functionality is layered between EMC Flare™ LUNs and EMC SnapView™ software (see Figure 1). The core code on Dell/EMC CX series arrays, Flare creates and manages RAID groups, including Flare LUNs within those RAID groups. All software applications above the metaLUN layer interact with a metaLUN in the same way that they would interact with individual Flare LUNs. When metaLUNs are used as the basic storage unit instead of Flare LUNs, they are designed to offer the following benefits: improved data protection and recovery, easier provisioning and expansion of storage, enhanced LUN performance, and the offloading of host-based virtualization to the storage array.

A metaLUN is essentially a mapping table to all of the Flare LUNs that it contains. The logical address of a block of data in a metaLUN is translated to a specific Flare LUN. The Flare LUN performs the actual reading and writing of the data to the storage processor's cache and to the specific disks in its RAID group. MetaLUNs require very few additional storage processor resources, but each metaLUN is currently limited to a maximum capacity of 2 TB.

A metaLUN is created when a Flare LUN is expanded using additional Flare LUNs; a Flare LUN can be expanded by consolidating Flare LUNs into a group. The resulting metaLUN inherits all the characteristics of the initial Flare LUN including the original LUN ID. This feature makes metaLUN creation transparent to the attached host server. The initial Flare LUN that is converted is known as the base LUN. All Flare LUNs that are added to a metaLUN conform to the characteristics of the base LUN.

Structuring metaLUNs for performance or capacity

Three methods of structuring metaLUNs exist: striping, concatenation, or a combination of both. A metaLUN structure comprises one or more components, and each component can contain one or more Flare LUNs. Depending on the CX series array, up to 16 components can be used in a metaLUN, and each component can hold up to 32 Flare LUNs. The striping of data is performed within a component, while concatenation is the process of adding more components to the metaLUN.

A striped metaLUN is designed to enhance performance by aggregating the throughput, measured in I/Os per second (IOPS), and bandwidth, measured in megabytes per second (MB/sec), of the underlying Flare LUNs contained in the stripe. Expansion of a metaLUN through striping is designed to increase both the performance and capacity of the metaLUN.

Implementing a striped expansion takes time, but once the striped expansion is initiated, it occurs in the background and is functionally transparent to the attached host server. Expansion rates can be set to low (4 GB per hour), medium (6 GB per hour), high (12 GB per hour), or ASAP. The default setting is high. The ASAP setting is designed to devote all of the array's storage processor resources to the expansion process, and therefore is not recommended in a production environment. When administrators implement striping within a component, all Flare LUNs used in that component must be identical in capacity, RAID type, and drive type (either Fibre Channel or ATA disks).

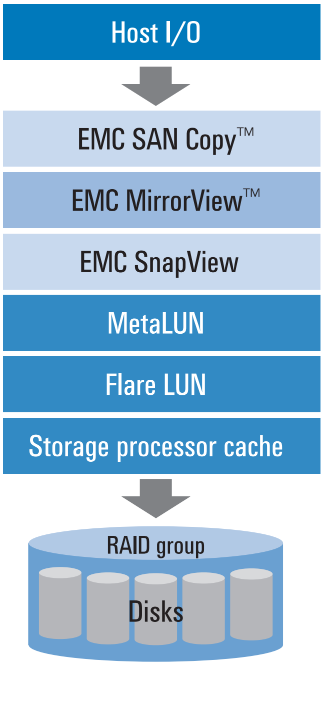


Figure 1. MetaLUN functionality in the storage array

Before the introduction of metaLUNs, host-based LUN virtualization was required to achieve higher performance from an array than was possible from a single Flare LUN. On Microsoft® Windows® operating system–based hosts, basic disks were converted to dynamic disks, allowing multiple LUNs to be collected into a dynamic, or *striped*, volume designed to provide either performance or capacity improvements. However, dynamic disks can be used only by nonclustered servers; Microsoft Cluster Service (MSCS) requires that all volumes in a cluster be basic disks, not dynamic disks—a restriction that limits the size, scalability, and performance of clustered hosts. By moving virtualization from the host to the Dell/EMC array, a high-performance metaLUN can be created and partitioned by the host as a basic disk.

Implementing a striped expansion takes time, but once the striped expansion is initiated, it occurs in the background and is functionally transparent to the attached host server.

Dell/EMC CX400 and CX600 storage arrays and the new CX500 storage array have two back-end 2 Gbps Fibre Channel arbitrated loops per storage processor. The new CX700 has four back-end 2 Gbps Fibre Channel arbitrated loops per storage processor. A striped metaLUN spanning multiple back-end loops is designed to provide superior performance to a host when compared to a simple Flare LUN on a single back-end loop.[1] The dynamic load-balancing features of EMC PowerPath® software are designed to enable high-performance access between multiple host bus adapters (HBAs) in the host and the Dell/EMC storage array.

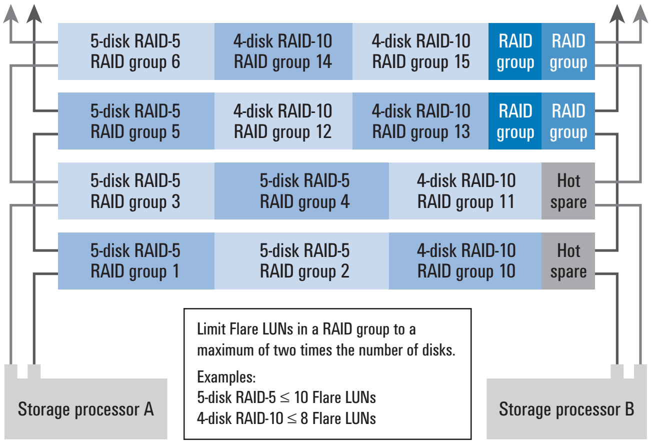
A concatenated metaLUN is designed to provide real-time capacity expansion. When administrators create a concatenated metaLUN, all Flare LUNs must be configured as either protected storage or nonprotected storage, and have the same drive type. Protected storage is derived from storage groups that use RAID-1, RAID-3, RAID-5, or RAID-10 levels. Nonprotected storage is derived from storage groups that use RAID-0 or single disks.

Concatenation is designed to provide immediate availability of additional capacity. A best-practices recommendation is to use the same RAID type and total number of spindles in each concatenated component of a metaLUN to help provide balanced, consistent performance across the entire metaLUN.

Simplifying RAID group creation and provisioning

Before Navisphere release 12, Flare LUNs were created in two ways. Multiple, small-capacity Flare LUNs could be created in one RAID group; or, a large-capacity Flare LUN could be created as a single-LUN RAID group. Properly creating a single-LUN RAID group for individual applications can be complex and can affect overall performance depending on the RAID type used.

[1] For more information on metaLUN performance considerations, visit https://powerlink.emc.com/HighFreq/folder_0/H519.3_clariion_fibre_chnl_dev_ldv.pdf.


5-disk RAID-5 RAID group 6
4-disk RAID-10 RAID group 14
4-disk RAID-10 RAID group 15
RAID group
RAID group
5-disk RAID-5 RAID group 5
4-disk RAID-10 RAID group 12
4-disk RAID-10 RAID group 13
RAID group
RAID group
5-disk RAID-5 RAID group 3
5-disk RAID-5 RAID group 4
4-disk RAID-10 RAID group 11
Hot spare
5-disk RAID-5 RAID group 1
5-disk RAID-5 RAID group 2
4-disk RAID-10 RAID group 10
Hot spare
Storage processor A
Storage processor B

Limit Flare LUNs in a RAID group to a maximum of two times the number of disks.

Examples:
5-disk RAID-5 ≤ 10 Flare LUNs
4-disk RAID-10 ≤ 8 Flare LUNs

Figure 2. A best-practices RAID group and hot-spare configuration

When multiple LUNs are created within a single RAID group, the individual LUNs cannot be dynamically expanded when the RAID group is expanded. Only a single-LUN RAID group can expand during a RAID group expansion.

For all types of SCSI and Fibre Channel storage, the level of available protection is set by the RAID type, and the required rebuild time is a function of the number, size, speed, and type of drives used in the RAID group. As the capacity of each drive doubles, the rebuild time can also double. In addition, in a RAID-5 group, the rebuild time becomes incrementally longer as the number of total drives increases. The overall performance of each LUN in a RAID-5 group is affected during the entire rebuild process. If more than one drive fails during the rebuild process, RAID-5 protection is lost and RAID-10 protection can also be lost.

Navisphere release 12 separates RAID group creation (which sets the level of protection) from LUN and metaLUN creation (the provisioning process), thereby simplifying storage management for administrators. Array administrators can easily create the type of protection (RAID-1, RAID-5, RAID-10, and so on) and the number of drives needed in each RAID group using the Navisphere graphical user interface (GUI). Storage group creation is a simple matter of selecting the individual drives to be used and committing the changes.

Configuring RAID groups according to best practices

Dell and EMC recommend smaller-size RAID groups of two to eight disks because these are designed to provide better data protection than larger-size RAID groups.[2] Smaller RAID groups have less chance of experiencing a multiple disk failure within the group. Additionally, the time to rebuild to a hot spare is faster with small RAID groups. Best-practice recommendations include the following:

- A two-drive RAID-1 set split between separate DAE2 disk array enclosures
- Either four-, six-, or eight-disk RAID-10 sets on the same DAE2
- Either three-, five-, or seven-disk RAID-5 sets on the same DAE2

A RAID group can contain 1 to 16 disks. Dell and EMC recommend four-disk RAID-10 sets for optimal performance and five-disk RAID-5 sets for optimal storage utilization. When using these basic configurations, organizations can select the location of RAID groups and hot spares for each DAE2 enclosure before the array is delivered (see Figure 2).

Provisioning LUNs from RAID groups

The building blocks of metaLUNs—Flare LUNs—can be used to create virtual storage from RAID groups. If an application requires additional performance or capacity, the metaLUN can be easily expanded. When performance is the primary consideration, striped metaLUNs are recommended; for large capacity, either striped or concatenated metaLUNs are recommended. When a RAID group rebuild is necessary because a single drive has failed, only the individual Flare LUNs within the particular RAID group are affected. A properly designed metaLUN should experience a partial degradation in performance only when accessing the affected Flare LUN.

To provision LUNs, array administrators can create a single Flare LUN for small storage requirements. Alternatively, for higher capacity requirements, administrators can create multiple Flare LUNs in different RAID groups and then combine these to create either a striped or concatenated metaLUN. By selectively creating Flare LUNs from available RAID groups, administrators can provision an array that is designed to offer balanced, predictable performance. For example, by using 15,000 rpm disks in small RAID-10 sets and provisioning the available storage in each RAID group to a single Flare LUN, administrators can form the building blocks for a large capacity, high-performance metaLUN. This type of metaLUN can be well suited for highly transactional databases, where efficient performance is the goal. Administrators also could create, for example, a striped metaLUN consisting of six high-performance

When performance is the primary consideration, striped metaLUNs are recommended; for large capacity, either striped or concatenated metaLUNs are recommended.

[2] For more information on best practices for Fibre Channel storage, visit https://powerlink.emc.com/HighFreq/folder_0/H519.3_clariion_fibre_chnl_dev_ldv.pdf.

RAID-10 Flare LUNs. Such a metaLUN could incorporate a total of 24 disks for I/O activity, which is greater than the 16-disk maximum limit of a single RAID group in Navisphere. If striped expansion is necessary, only four additional disks would be required to create a seventh RAID group and Flare LUN.

Alternatively, for efficient storage utilization, administrators could create six separate RAID-5 groups (see Figure 3). Each RAID group could be further divided into multiple Flare LUNs of equal size, creating a pool of available storage. The array administrator would need to create the RAID groups and storage pool only once, during the initial setup. Later, when an application dictates specific storage performance and capacity requirements, the array administrator could selectively choose from the pool to provision the metaLUN. The generally preferred method of metaLUN creation is through striping. When expansion is necessary, a Flare LUN from a different RAID group is used in the expansion process.

Improving array performance through metaLUN functionality

The addition of metaLUN functionality to Dell/EMC storage arrays is designed to provide LUN virtualization through striping or concatenation. To help enhance performance to the host at the LUN level, administrators can implement striped metaLUNs that span multiple back-end loops, alleviating the need for the host to provide storage virtualization. The performance from individual RAID groups can be set on the array, and this performance can be consistent and measurable.

MetaLUNs separate the data protection process—that is, RAID group creation—from the storage provisioning process that comprises LUN and metaLUN creation and expansion. In this way, the metaLUN approach is designed to enable array administrators to reduce the risks of degraded performance and potential data loss. The separation of the data protection and storage provisioning processes can also help improve the usability of future storage management products. For example, because they provide virtualization and expansion at the array level, metaLUNs enable products such as EMC VisualSRM™ storage resource management software—which can make standard calls to the array—to effectively automate the processes of storage provisioning and expansion.

To help enhance performance to the host at the LUN level, administrators can implement striped metaLUNs that span multiple back-end loops, alleviating the need for the host to provide storage virtualization.

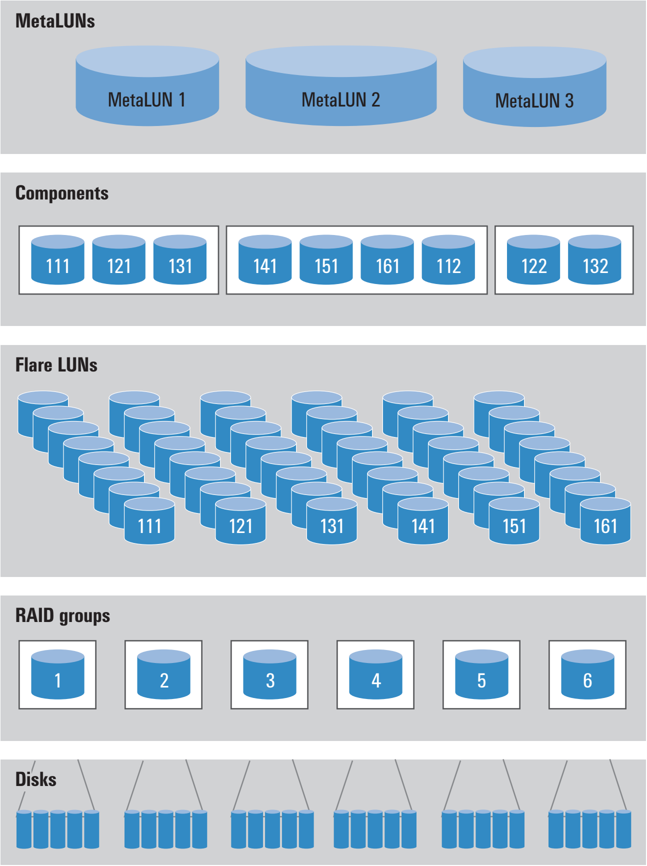


Figure 3. A metaLUN designed for efficient storage utilization

Michael L. Gierhart (michael_gierhart@dell.com) is an enterprise technologist in the Americas Advanced Systems Group at Dell. His area of expertise is storage area network (SAN)–based storage solutions and performance tuning. Michael has a B.S. from the University of California, Berkeley; has pursued post-graduate studies at Old Dominion University; and is a Microsoft Certified Systems Engineer (MCSE).

FOR MORE INFORMATION

Best practices for Fibre Channel storage:
https://powerlink.emc.com/HighFreq/folder_0/H519.3_clariion_fibre_chnl_dev_ldv.pdf

Dell/EMC storage arrays:
http://www.dell.com/storage

EMC Navisphere Management Suite:
http://www.dell.com/downloads/global/products/pvaul/en/navispheremanagementsuite.pdf

EMC VisualSRM:
http://www.emc.com/products/software/visual_srm/visual_srm.jsp

Deriving Maximum Performance

and Scalability with Dell PowerVault 770N and 775N Network Attached Storage Servers

Storage administrators must understand the performance capabilities of network attached storage (NAS) servers to enable these servers to perform at the appropriate level within their computing environments. This article examines the peak performance and scalability behavior of two mid-tier NAS offerings—the Dell™ PowerVault™ 770N and 775N NAS servers—in various enterprise scenarios.

BY WARD WOLFRAM AND PHILIP WONG

Network attached storage (NAS) servers offer an affordable option for additional data storage. NAS servers are comparable to traditional file servers but improve on traditional file servers by offering advanced file-sharing features—including support for multiple operating systems and files systems, and the ability to place the NAS server in a physically separate location from the application server. Because NAS servers are dedicated to file sharing, general services, drivers, and other items that are typically found on general-purpose servers have been removed. As a result, administrators cannot install applications such as Microsoft® SQL Server™ database, Oracle® Database 10*g*, and Microsoft Exchange on NAS servers. However, NAS servers are less complex than typical general-purpose servers, so they tend to be more reliable—and simpler to manage.

NAS servers also can provide shared networked storage to disparate types of servers and clients. Dell™ NAS servers enable shared storage by supporting the following file-sharing protocols: Novell® NetWare® Core Protocol (NCP), Network File System (NFS), Apple® AppleTalk® Filing Protocol (AFP), Common Internet File System (CIFS), HTTP, and FTP. For example, networks with hosts running the Microsoft Windows® operating system use CIFS to access shared storage, whereas hosts running the UNIX® operating system traditionally use NFS. Dell NAS servers offer the ability to concurrently support these different types of hosts within a single storage share.

For storage administrators to determine the appropriate NAS configuration required in their computing environments, they must understand the performance capabilities of NAS servers. This article examines peak performance and scalability of two mid-tier NAS servers: the Dell PowerVault™ 770N and PowerVault 775N storage servers.

Measuring NAS performance

The PowerVault 770N and 775N are basically identical, except that the PowerVault 770N supports additional Peripheral Component Interconnect Extended (PCI-X®) and internal hard drive slots in a 5U form factor. In February 2004, Dell engineers conducted performance and scalability tests for various enterprise configurations of the PowerVault 770N and 775N, which run a version of Windows built using the Microsoft Server Appliance Kit (SAK) version 3.

Two performance metrics help delineate the capabilities of NAS servers:

- **Maximum performance testing:** Determines the maximum performance boundary of the NAS server under stress. Maximum performance testing strives to achieve the best possible (peak) performance without manipulating the I/O.
- **Scalability testing:** Examines NAS performance behavior at different concurrent user loads. Scalability testing varies user loads, I/O ratios, user wait states, and file sizes to simulate the most common types of real-world user activity.

Results from maximum performance testing are used mainly to compare the performance of different NAS servers. The results can also identify the point at which NAS server resources become overutilized. However, by examining only maximum performance results, storage administrators may have difficulty determining performance behavior in various real-world environments. Instead, administrators can use scalability results to accurately determine which and how many NAS units fit best in a given computing scenario. Scalability testing can also identify networking configuration and deployment pitfalls to avoid.

Testing the PowerVault 770N for maximum performance

Dell measured the peak performance of a PowerVault 770N NAS server configured as follows:

- **Memory:** 4 GB of 266 MHz double data rate (DDR) SDRAM
- **Storage:** Four 15,000 rpm, 18 GB Ultra320 (U320) SCSI hard disk drives
- **Network connectivity:** Intel® PRO/1000+ LAN on Motherboard (LOM)
- **RAID controller:** Dell PowerEdge™ Expandable RAID Controller 4, Dual Channel integrated (PERC 4/Di) with 128 MB battery-backed cache (default settings); tested RAID levels were RAID-0 and RAID-5
- **Processor:** Dual Intel Xeon™ processors, including the Intel Xeon processor at 2.8 GHz with 512 KB of level 2 (L2) cache, the Intel Xeon processor at 3.06 GHz with 512 KB of L2 cache, the Intel Xeon processor at 3.06 GHz with 512 KB of L2 and 1 MB of level 3 (L3) cache, and the Intel Xeon processor at 3.2 GHz with 512 KB of L2 and 1 MB of L3 cache

All performance tests used 60 Dell OptiPlex™ GX1 desktop computers as network clients to generate network throughput and create the necessary client load. Each client contained a 1 GHz processor and an Intel PRO/100+ network interface card (NIC). Because the clients ran the Microsoft Windows XP Professional operating system, CIFS was used as the file-sharing protocol. The clients and the PowerVault 770N were connected through a Dell PowerConnect™ 3024 switch.

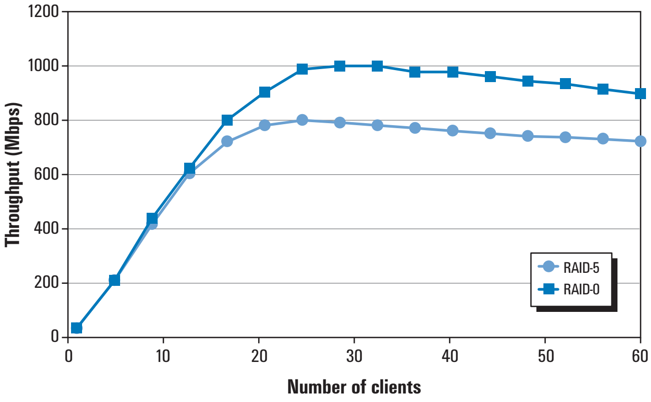


Figure 1. Effect of RAID type on throughput for the maximally configured PowerVault 770N

To produce performance results for the PowerVault 770N, Dell used version 7.0.3 of the Ziff Davis® NetBench® portable benchmark tool.[1] NetBench generates client workloads using file types similar to those supported by common Windows-based applications.

NetBench produces results for throughput and average I/O response time. Throughput, reported in megabits per second (Mbps), is the rate at which data is transferred between the server and its network clients. Average response time, reported in milliseconds (ms), is the average amount of time the server takes to respond to and complete network client requests at a given client load. Average response times are inversely proportional to throughput: the higher the response time, the lower the throughput. Greater throughput means faster performance. These results help reveal when the PowerVault 770N reaches a critical threshold, which is the point at which response times begin to increase and the NAS server runs in a suboptimal state.

Note that no correlation exists between the number of NetBench test clients and the scalability potential of the PowerVault 770N. To quickly obtain the peak throughput for a server during a test run, NetBench clients are asked to produce an unrealistic volume of common network file operations.

Effect of RAID configuration and processor cache

As shown in Figure 1, the peak throughput for RAID-0 on the maximally configured PowerVault 770N (containing 4 GB of DDR SDRAM and dual Intel Xeon processors at 3.2 GHz with 1 MB of L3 cache and Hyper-Threading enabled) is 1011 Mbps (126 MB/sec). The RAID-5 storage configuration achieves a peak throughput of 817 Mbps

[1] For more information about NetBench, see http://www.etestinglabs.com/benchmarks/netbench/default.asp.

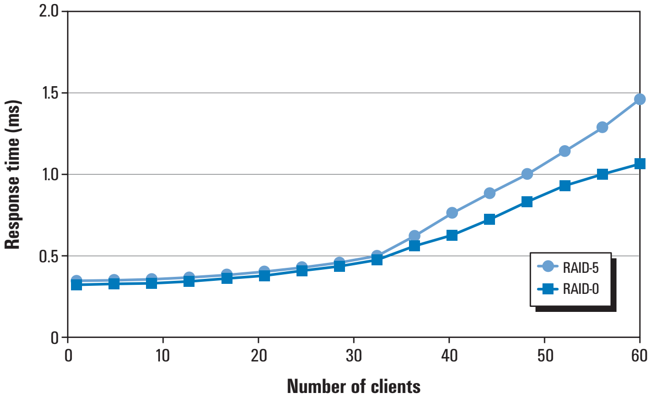


Figure 2. Effect of RAID type on average response time for the maximally configured PowerVault 770N

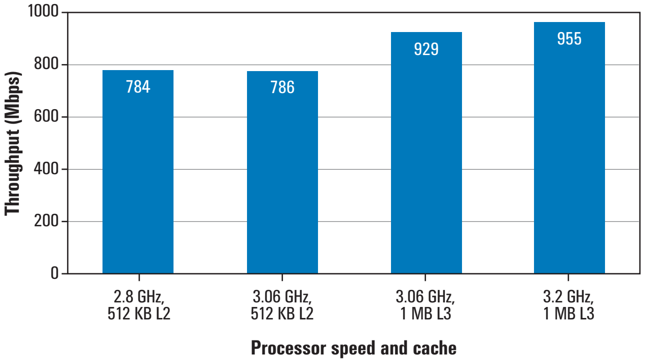


Figure 3. Effect of processor cache on throughput in RAID-5 configuration

(102 MB/sec), a 26 percent performance drop compared to RAID-0 throughput. This decrease is normal because the RAID controller requires additional computation to produce RAID-5 parity bits.

Figure 2 shows that the average I/O response times of the maximally configured PowerVault 770N increase after the server reaches its maximum sustainable throughputs for both RAID-0 and RAID-5. The difference between RAID-0 and RAID-5 performance also grows exponentially after the point of peak throughput. Storage administrators can use the tail end of the response time curves to determine when quality of service for a fully stressed PowerVault 77xN server begins to become unacceptable.

Processor cache also plays a vital role in NAS server performance, as shown in Figure 3. The additional L3 cache in the Intel Xeon processor running at 3.06 GHz boosts RAID-5 throughput by 18 percent, compared to the throughput of processors without an L3 cache. By maximizing the processor cache, storage administrators can improve NAS throughput.

Testing the PowerVault 775N for scalability

For scalability testing, Dell equipped PowerVault 775N hardware in maximum and minimum configurations, as shown in Figure 4. Each PowerVault 775N contained a PERC 4/Dual Channel (DC), which connected the server to a PowerVault 220S SCSI enclosure. Equipped with fourteen 15,000 rpm, 36 GB U320 hard drives in a RAID-10 configuration, the PowerVault 220S supplied a single file share for the test.

Six PowerEdge 2550 servers were used to generate user loads on a Gigabit[2] Ethernet test network. A seventh PowerEdge 2550 was configured as a domain controller to authenticate users for all user loads.

To generate user loads for the scalability testing, Dell used the Quest Software® Benchmark Factory® load-testing tool.[3] Benchmark Factory provides a framework for developing and running benchmarks that apply real-world stress on a test system. Dell configured Benchmark Factory to generate disk read-write ratios of 80 percent read, 20 percent write; 50 percent read, 50 percent write; and 20 percent read, 80 percent write. All tests applied 80 percent random and 20 percent sequential I/Os. Dell also defined the following I/O file sizes in Benchmark Factory:

- **Small:** Between 30 KB and 520 KB, inclusive
- **Medium:** Between 1 MB and 3 MB, inclusive
- **Large:** Approximately 6 MB

For all tests, Dell started with 102 virtual users and stepped up the number of users every 12 minutes by 102 to end with 3,672 users. Increasing the user I/O request rate and the total number of users within a test can significantly affect the amount of stress applied to a NAS server. Each user defined within the scalability test had an identical I/O request rate of one request every one to two seconds. In other words, each user issued at least one I/O every two seconds, but not before one second. Because typical production environments will not experience the extremely high user I/O request rates generated by Benchmark Factory, the performance results that are presented in this article should be viewed as worst-case real-world results. The PowerVault 775N may perform significantly better in typical enterprise environments.

Effect of user load on performance

Note that the user load boundary of 3,672 shown in the scalability results does not represent the maximum user load of the PowerVault 775N,

Maximum configuration	Minimum configuration
• Two Intel Xeon processors at 3.2 GHz with 512 KB of L2 cache and 1 MB of L3 cache • 4 GB system memory	• Two Intel Xeon processors at 2.4 GHz with 512 KB of L2 cache • 2 GB system memory

Figure 4. PowerVault 775N configurations for scalability testing

[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 Benchmark Factory, see http://www.benchmarkfactory.com.

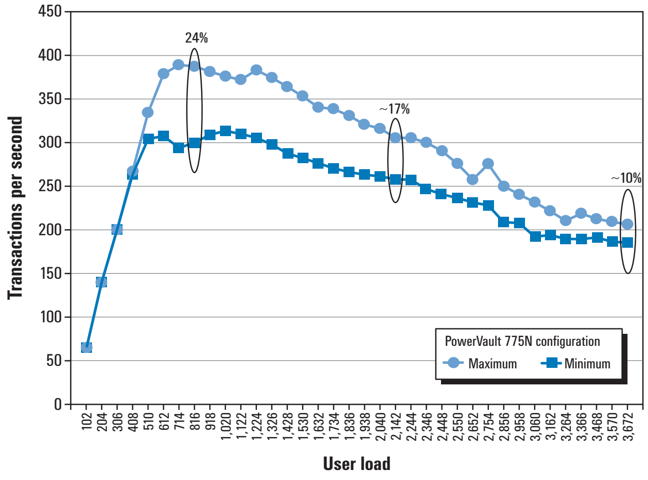


Figure 5. Effect of PowerVault 775N configuration on performance, using small file sizes and an 80:20 read-write ratio

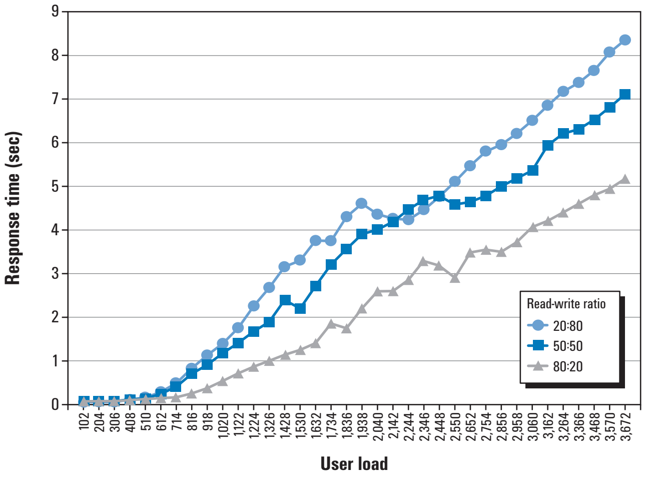


Figure 6. Effect of read-write ratios on a maximally configured PowerVault 775N, using small file sizes

nor does it indicate any type of failure. The number is simply the maximum licensing limit within Benchmark Factory.

Figure 5 shows that, as user load increased, the performance difference between the minimum and maximum PowerVault 775N configurations diminished. To benefit from the performance of a maximally configured PowerVault 775N, storage administrators may consider adding another NAS server as user loads increase.

The read-write ratio can significantly affect the performance of a maximally configured PowerVault 775N, as shown in Figure 6. For example, when I/O requests averaged 80 percent reads at a user load of 1,632, each user could expect the I/O request to be satisfied in approximately 1.5 seconds. If the average I/O request was 80 percent writes, each user could expect I/O requests to be satisfied in approximately 4 seconds.

Figure 7 can help storage administrators determine how many PowerVault 775N servers they may need. For example, if a 15-second response time is unacceptable in an environment with 1,530 users who typically request large files and perform 80 percent reads and 20 percent writes, an additional NAS server is recommended. If the administrator implements and properly configures a second PowerVault 775N such that the load is balanced between the two servers, response times to each of the 1,530 users can decrease from 15 seconds to 7 seconds. The reduction in response time can occur because the user load for each server decreases from 1,530 to 765.

Achieving optimal PowerVault 77xN performance

In this study, peak performance testing for the PowerVault 770N identified that PowerVault 770N resources become fully utilized at a throughput of 1011 Mbps. With an understanding of this result, administrators can effectively avoid overutilizing the PowerVault 77xN, keeping its operation at a highly available and functional state.

To help determine whether a PowerVault 77xN will perform at a satisfactory level for their specific enterprise configurations, administrators can consider the results from the scalability testing presented in this article. These results can help administrators determine the number of PowerVault 77xN servers required to achieve satisfactory NAS performance for a multitude of users.

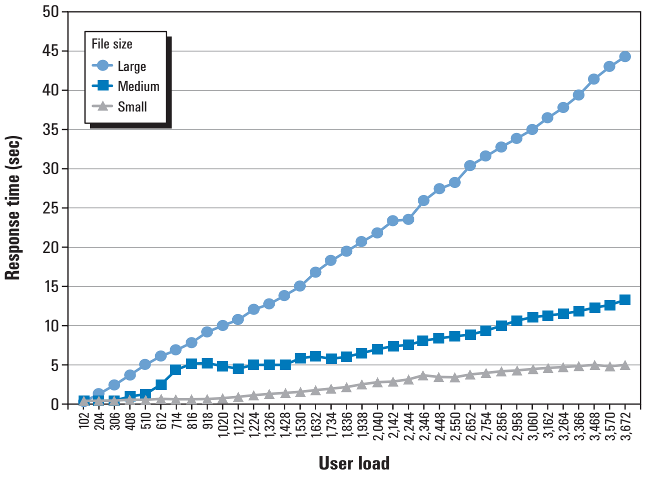


Figure 7. Effect of file size on a maximally configured PowerVault 775N, using an 80:20 read-write ratio

Ward Wolfram (ward_wolfram@dell.com) is a storage performance and solution engineer on the Dell Solution Enablement Lab and Technology Showcase team. His responsibilities include performance and best-practice analysis for storage area network (SAN), NAS, and tape backup systems. Ward has an M.S. in Computer Science from the University of Nebraska at Lincoln and a B.S. in Mathematics from Concordia University in Seward, Nebraska.

Philip Wong (philip_wong@dell.com) is a systems engineer in the Dell Enterprise Performance Analysis Lab. His current projects include NAS systems performance tuning and industry-standard benchmarking on the Red Hat® Linux® operating system. Philip has a bachelor's degree in electrical engineering from the University of Pennsylvania.

Achieving Comprehensive Enterprise Backups

with VERITAS Backup Exec Desktop and Laptop Option

Many server backup products do not back up all business data because employees often store a significant amount of data on their PC hard drives rather than on the enterprise network. To address this "hidden" data, the VERITAS Backup Exec™ *for Windows Servers* Desktop and Laptop Option automatically copies user data from desktops and laptops to existing network storage. By using this tool, IT administrators can help ensure a more comprehensive backup of enterprise data.

BY BRIAN GREENE

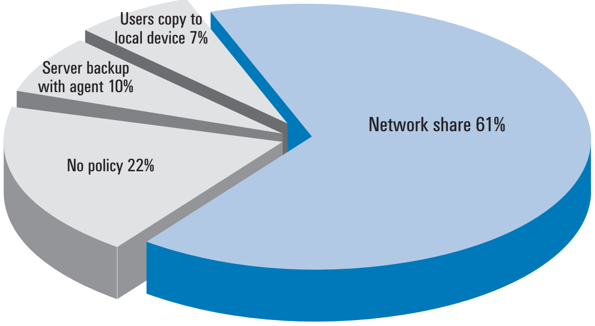
The typical enterprise uses a plethora of desktops and laptops. In such IT environments, the chances are high that all business data is not being backed up. Instead, the information is likely stored on individual desktop and laptop systems—areas often neglected by corporate backup policies, which tend to focus on servers and primary storage. This article examines the VERITAS Backup Exec™ *for Windows Servers* Desktop and Laptop Option, a tool that can help IT administrators back up the significant amount of business data located on desktops and laptops.[1]

Enterprise backups often miss critical business data

The most popular method for backing up enterprise data is to use network shares (see Figure 1). In many organizations, the policy is to rely on employee initiative to back up data that is not stored on network shares. Under these circumstances, users who need to protect data stored on their PCs must copy it to the server so the server backup product can save it. However, when users are responsible for backing up their own files, this backup often does not occur and organizations can find out the hard way that suggested backup policies are not being followed. For example, employees can be so busy with their primary job responsibilities that they do not manually copy important documents and e-mail messages to a network share. As a result, the only copy of many e-mail messages and documents may reside in the .pst files located on users' PCs. When something happens to an unprotected laptop or desktop, that information can be lost.

Local user file-protection policies were acceptable years ago when the data that resided on PCs was likely to be personal data. In today's environment, however, a large percentage of business data resides on users' computers. A network share backup method can be ineffective when business-critical data does not reside on the network.

[1] For more information about VERITAS Backup Exec *for Windows Servers*, please visit http://www.veritas.com/Products/www?c=product&refId=57.



Source: VERITAS survey (February 2004)

Figure 1. Popular backup methods: Network share is the most popular but can be the least effective

VERITAS Backup Exec software provides an option to back up data stored on PCs

The VERITAS Backup Exec *for Windows Servers* Desktop and Laptop Option—which is integrated with VERITAS Backup Exec storage software—automates existing data protection policies. It enables business data to be copied automatically from desktops and laptops to the network share, thereby freeing users from the responsibility of protecting their own data.

The Desktop and Laptop Option does not add any work to the daily backup schedule; instead, it uses existing network shares to offer automatic disk-based protection for business data, whether users are in the office or on the road. In addition, it enables users to restore their own files and maintain synchronization among multiple desktops and laptops.

By automatically copying user data to existing network storage or to the local PC, the Desktop and Laptop Option easily integrates into existing IT infrastructure and policies, helping to lower total cost of ownership (TCO) and enhance return on investment (ROI) for Backup Exec deployments.

Why a network share approach

To provide a product that could be used in most network environments without requiring a major investment in time and money, VERITAS Software Corporation designed the Desktop and Laptop Option to work with the infrastructure that enterprises already have in place. The Desktop and Laptop Option does not require a heavyweight, dedicated application server or use a large database to constantly communicate with all connected desktops and laptops. Instead, it uses Microsoft® Windows® networking protocols to authenticate users and to send files. The theory behind this product is that fewer moving parts mean less overhead, which can help reduce the backup time and effort for users and administrators alike.

In addition, VERITAS has tested the Desktop and Laptop Option on Dell™ systems. Figure 2 lists the Dell product lines and models that can support this option.[2]

Desktop and Laptop Option integrates into Backup Exec environments

The VERITAS Backup Exec *for Windows Servers* Desktop and Laptop Option comprises two main components:

- Desktop and Laptop Administration Console (for the administrator)
- Desktop and Laptop Agent (for the desktop or laptop user)

These components are used with Backup Exec software to help provide a complete data protection solution (see Figure 3).

The Desktop and Laptop Administration Console

The Desktop and Laptop Administration Console is part of the Backup Exec Administration Console. It can be run from a Windows-based workstation or the Backup Exec media server. The console provides administrators with a summary view of tasks and the backup status of desktops, laptops, and servers in the network.

All traditional administrative functions can be performed from the Administration Console, including setting up or changing profiles, checking status and alerts for users and servers, managing settings, designating or monitoring storage, and initiating restore jobs for individual PCs.

Once a profile is set up, the Desktop and Laptop Option can perform the specified backup tasks automatically. Profiles provide an easy way to logically group users by department or function as well as to tailor specific backup parameters to meet each group's needs. Within each profile, administrators can control:

- Schedules
- File versions
- Inclusion or exclusion of drives, folders, and files
- Maximum amount of storage designated for each user

The Desktop and Laptop Agent

The Desktop and Laptop Agent resides on each desktop and laptop. It protects the business data on the PC by backing up the designated drives, folders, and files from the hard drive to the appointed storage on a network share.

Dell product line	Models
Dell PowerVault™ network attached storage (NAS) servers	725, 745, 770, 775
Dell PowerEdge™ servers	400, 700, 1500, 1600, 1750, 2600, 4400, 4600, 6400, 6600

Figure 2. Dell hardware on which VERITAS has tested the VERITAS Backup Exec *for Windows Servers* Desktop and Laptop Option

[2] Dell has not certified these products for VERITAS Backup Exec *for Windows Servers* Desktop and Laptop Option; this information is the result of VERITAS testing on Dell hardware.

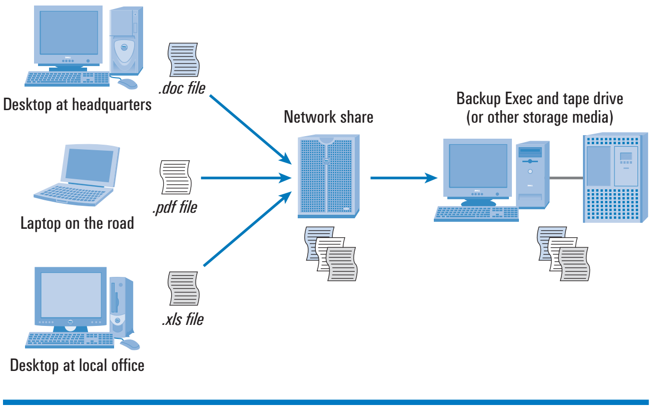


Figure 3. Protecting data on desktops and laptops using Backup Exec and the Desktop and Laptop Option

During the installation and setup process, administrators initially configure the Desktop and Laptop Agent on each PC by using profiles. Administrators can grant users the right to change their backup options, or administrators can deny all rights. Depending on the rights granted, users may be allowed to restore files, synchronize files, configure backup selections, set schedules, view history, and perform other functions.

Once protected, data can be restored at any time by the administrator or by the user. Using domain-level security for authentication, administrators can quickly and easily set up synchronization and recover files if a user changes a PC or needs to recover from a failure.

Backups use schedule-free method to provide high data protection

Reliably backing up laptops can be a challenge. Most backup products rely on a schedule to start the backup job. If the laptop is not connected during the scheduled start time, then the laptop misses the backup window and business data that resides on the laptop is not protected.

The Desktop and Laptop Option will run a backup whether the computer is offline or online. If the computer is offline, files are copied to a local data folder on the user's PC. When the computer reconnects to the network, the files are copied from the local folder to the designated storage location on the network. This functionality helps ensure a high level of file protection. In addition to backing up the data, this approach enables users to restore files while offline.

Furthermore, the Desktop and Laptop Option does not depend on a schedule. The continuous backup mode will copy PC files to a local data folder as they change or as they are saved, creating a local backup that can help protect business data if a scheduled network backup is missed. This schedule-free method also can help provide a higher degree of data protection than scheduled approaches because scheduled backups are often performed on a daily basis—which can result in a 24-hour window of data loss. Consequently, files saved in continuous backup mode using the Desktop and Laptop Option can be more up-to-date than files backed up on a daily basis.

Protecting Microsoft Outlook .pst files

The Desktop and Laptop Option backs up personal mail data—the .pst file—from the Microsoft Outlook® e-mail client by using the Microsoft Messaging Application Programming Interface (MAPI). This file can be backed up while Outlook is open and in use by the user. During the backup, only mail messages that have changed or are new since the last backup are copied to the network share. In this way, the Desktop and Laptop Option can help lower storage overhead and bandwidth usage by not copying the entire .pst file each time changes occur.

By leveraging Microsoft Outlook technology and using MAPI, the Desktop and Laptop Option does not have to scan the .pst file and compare that scan to a .pst file located on a backup server. VERITAS avoided that approach because file-scanning can affect system response—for example, by stalling the e-mail service or degrading overall performance. Instead, MAPI informs the Desktop and Laptop Option when something in the .pst file has changed and the Desktop and Laptop Option copies the changes to the storage location.

VERITAS Backup Exec Desktop and Laptop Option provides data protection beyond the server

Although many enterprises recognize that critical business data may reside on desktop and laptop systems, they often continue to back up only servers. Current desktop and laptop backup policies can be unreliable—or nonexistent.

The VERITAS Backup Exec *for Windows Servers* Desktop and Laptop Option can integrate into an enterprise's current storage and network infrastructure and extend the Backup Exec server backup to desktops and laptops. In this way, the Desktop and Laptop Option can help provide ongoing data protection and support for all desktop and laptop systems, whether users are in the office or on the road. In addition, the Desktop and Laptop Option can help streamline administration by providing an at-a-glance task summary and the backup status of desktops, laptops, and servers in the network. VERITAS Backup Exec *for Windows Server* with the Desktop and Laptop Option can help enterprises ensure that all business data is protected.

Brian Greene (brian.greene@veritas.com) is a senior staff product manager at VERITAS Software (http://www.veritas.com), a leading storage software company that offers products to improve application performance and provide data protection, storage management, high availability, and disaster recovery.

FOR MORE INFORMATION

VERITAS Backup Exec Desktop and Laptop Option:
http://www.veritas.com/offer?a_id=6742

FREE Subscription Request

❏ **Yes!** I want to receive *Dell Power Solutions* **Free.**

Check one:

❏ New subscription ❏ Renew subscription
❏ Address change ❏ Cancel subscription

Current subscriber ID (from mailing label): ________________

First name: ________________

Last name: ________________

Company name: ________________

Address 1: ________________

Address 2: ________________

City: ________ State, province, or territory: ________

Zip/postal code: ________ Country: ________

Telephone: ________________

E-mail address: ________________

Please complete the questions listed below to receive your free subscription to *Dell Power Solutions.* You must answer all questions to qualify for a free subscription.

1. **Which of the following best describes your job function?**
   - ❏ Management (CxO, VP, director)
   - ❏ IT manager or supervisor
   - ❏ Systems engineer/technical consultant
   - ❏ Systems analyst
   - ❏ System administrator
   - ❏ Network administrator
   - ❏ Project manager
   - ❏ Marketing/sales
   - ❏ Other

2. **How large is your company, in terms of annual sales volume?**
   - ❏ Less than $5 million
   - ❏ $5–$9 million
   - ❏ $10–$49 million
   - ❏ $50–$99 million
   - ❏ Greater than $100 million
   - ❏ Greater than $1 billion

3. **How large is your company, in terms of employees?**
   - ❏ Less than 200
   - ❏ 200–500
   - ❏ 500–1,000
   - ❏ 1,000–5,000
   - ❏ Greater than 5,000

4. **What is the principal industry served by your company?**
   - ❏ Education
   - ❏ Financial services and banking
   - ❏ Government and military
   - ❏ Healthcare
   - ❏ Hospitality
   - ❏ Internet or Internet service provider
   - ❏ Manufacturing
   - ❏ Retail
   - ❏ Telecommunications
   - ❏ Utilities
   - ❏ Other

5. **What Dell products does your company use?**
   - ❏ Desktops or notebooks
   - ❏ Servers or storage
   - ❏ All of the above
   - ❏ None of the above

6. **What operating systems does your company use?**
   - ❏ Windows
   - ❏ Novell
   - ❏ UNIX
   - ❏ Linux
   - ❏ Mixed
   - ❏ Other

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

06/04

Subscribe online at www.dell.com/powersolutions_subscribe

PLACE
POSTAGE
HERE

DELL™
POWER
SOLUTIONS

MAILSTOP 8456
DELL INC
ONE DELL WAY
ROUND ROCK TX 78682
USA

Using Red Hat Network Satellite Server to Manage Dell PowerEdge Servers

A subscription to Red Hat® Enterprise Linux® software includes Internet access to the Red Hat Network update service. The Red Hat Network provides a simple solution for updating, deploying, and managing Red Hat Enterprise Linux systems. Administrators also can set up a local Red Hat Network Satellite Server to enable additional features and functionality. This article discusses how a Dell lab installed and configured a Satellite Server for its Dell™ PowerEdge™ servers.

BY TODD MUIRHEAD AND PETER LILLIAN

Updating enterprise systems with the latest security, enhancements, and bug fixes is critical in today's IT environment. The Red Hat® Network provides the ability to deploy, manage, and update packages for the Red Hat Enterprise Linux® operating system (OS)—potentially saving time and support costs while keeping systems secure and up-to-date. Internet access to the Red Hat Network is included as part of a subscription to Red Hat Enterprise Linux software, enabling organizations to manage their enterprise systems regularly with scheduled updates or automatically as updates become available. By obtaining updates directly from Red Hat, administrators can procure patches, security fixes, and enhancements that have been tested and authenticated by Red Hat engineers.

For a large number of servers, updates can consume a significant amount of bandwidth. Additionally, some administrators need to distribute their own unique packages across their infrastructure. To address these requirements, Red Hat offers the Satellite Server model, which allows all subscribed Red Hat Network content and functionality to be available locally (see Figure 1). Content can be downloaded from the Internet as updates become available, or CD-ROM images of the updated content can be downloaded separately and then loaded locally. Once the updates have been loaded, the Satellite Server can be completely disconnected from the Internet, allowing Red Hat OS–based systems to be updated through the Satellite Server while isolated or disconnected from the Internet.

Exploring the advantages of a Red Hat Network Satellite Server

In many data centers, the number of Internet connections is minimized to maintain a high level of security. The Satellite Server model allows for an environment that is completely disconnected from the Internet. However, administrators must determine how to transport the updates to an isolated Satellite Server. The options are to connect the Satellite Server to the Internet, or to load all the updates onto CD-ROMs and load the update content onto the local systems.

The Satellite Server generally allows updates to occur faster than with Red Hat's hosted servers. Because the

Satellite Server and the servers that are being updated through the Satellite Server are on an internal network, the available bandwidth between them is usually much greater than a connection over the Internet in the hosted model. If many systems must be updated and managed, then the increased speed of the Satellite Server updates can be a deciding factor in determining which Red Hat Network model to use.

Using a Red Hat Network Satellite Server also allows for the creation of custom channels and the distribution of custom software packages. By creating a custom channel and loading RPM™ (Red Hat Package Manager) software packages into that channel, administrators can use the Satellite Server to distribute and later provide updates to those packages.

The Red Hat Network offers three modules: Update, Management, and Provisioning. In the Dell™ lab discussed in this article, only the Update and Management modules are used. The Provisioning module adds the abilities to deploy, configure, manage, update, and redeploy systems as well as the ability to take snapshots of systems for future rollback, if needed. This article examines how the Dell lab uses the Red Hat Network Satellite Server model to manage its Dell PowerEdge™ servers—including the installation and configuration of the Satellite Server, updating of systems, and deployment of custom updates and packages using the Update and Management modules.

The Red Hat Network provides all of its software content in the form of channels—each version of Red Hat Enterprise Linux has a separate channel. In the Dell lab, six channels are used on the Satellite Server: three channels for Red Hat Enterprise Linux 2.1—one each for the AS, ES, and WS OS types—and three channels for Red Hat Enterprise Linux 3 for the same three OS types.

Setting up the Red Hat Network Satellite Server

Administrators can install the Red Hat Network Satellite Server themselves or have it installed as part of a services engagement. Red Hat offers a five-day service that includes architecture assessment, installation, configuration, training, deployment, and troubleshooting.

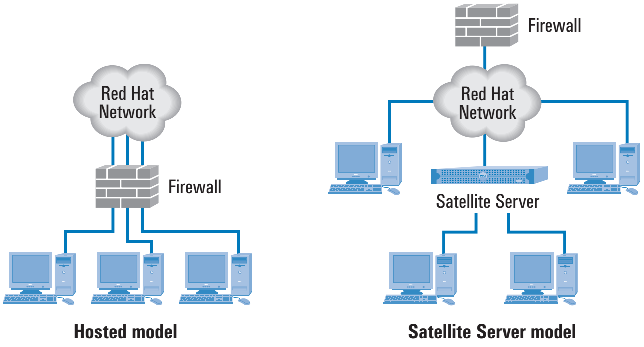


Figure 1. Two methods for updating servers from the Red Hat Network: The hosted model and the Satellite Server model

	Minimum requirements for Red Hat Network Satellite Server	Dell PowerEdge 2650 server
OS	Red Hat Linux AS 2.1	Red Hat Linux AS 2.1
CPU	Two Intel® Xeon™ processors DP at 2.4 GHz	Two Intel Xeon processors DP at 2.8 GHz
Memory	2 GB	4 GB
Internal disks	Three 36 GB drives	Five 36 GB drives
NICs	Not specified	Two 10/100/1000 Mbps (internal) NICs Two Intel PRO/1000XT Gigabit* Ethernet NICs
Disk controller	SCSI recommended	PowerEdge RAID Controller 3, Dual Channel integrated (PERC3/Di)

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

Figure 2. Red Hat Network Satellite Server: Minimum requirements for any deployment and configuration of the PowerEdge 2650 used as the Satellite Server in the Dell lab

Hardware requirements

In March 2004 the Dell lab installed and configured a Satellite Server for its Dell PowerEdge servers. Dell recommends using a Dell PowerEdge 2650 server for the Red Hat Network Satellite Server. Figure 2 shows the configuration details for the PowerEdge 2650 used in the Dell lab as well as the minimum requirements for any Satellite Server.

In the Dell lab, the five internal disks of the PowerEdge 2650 were configured as RAID-5 to provide the maximum amount of storage. One network interface card (NIC) is connected to the internal network and the other is connected to a network that has Internet access. Only one NIC at a time is used, and the firewall is always running to keep security at a high level.

Software requirements

The first step for the team at the Dell lab was to install Red Hat Enterprise Linux AS 2.1 on the Dell PowerEdge 2650. Using the latest version of Dell OpenManage™ Server Assistant, the team quickly installed the OS on the PowerEdge 2650, finishing in about 30 minutes. The firewall was configured on the server with only ports 22 (ssh), 80 (http), and 443 (https) open.

Satellite Server installation

A Satellite Server installation can be completed using the graphical user interface (GUI) or command-line interface (CLI)—both provide the same options. Using the GUI allows administrators to quickly and easily provide the correct information for all options. The Red Hat Network entitlement certificate, which is included as part of the Satellite Server purchase, must be provided during installation. The location of the entitlement certificate file can be specified, or the information can be entered in the provided spaces.

The Red Hat Network Satellite Server model can provide an excellent way to manage updates and software deployments for Red Hat Enterprise Linux systems.

Administrators also have the option of using an embedded database server or a separate database server. A database is required as part of the Satellite Server model to manage channel content as well as the package content of each internal server. In the Dell lab, the local embedded database option was selected for ease of installation.

The option to create a Secure Sockets Layer (SSL) certificate is also presented during installation. To help maintain the highest level of security for the system, administrators should complete this step.

A final option in the setup is to create a bootstrap script. This script can greatly simplify the client systems' configuration and connection to the Satellite Server. The required fields should be prepopulated with the necessary information based on information entered in previous steps. In the Dell lab, the bootstrap file was saved in the /var/www/html/pub/bootstrap directory.

During the installation process, the Red Hat Network Satellite Server will update itself to all the proper patch levels. Best practices strongly recommend that administrators reboot this server after completing the installation process. If a kernel update is included, then administrators must reboot so that the system can boot with the new kernel.

Satellite Server configuration

In the Dell lab, the two NICs in the Satellite Server were connected and configured so that one NIC has access to the Internet and the other NIC is connected to the isolated lab network. The NIC with access to the Internet is disabled by default at boot and only enabled during downloads of Red Hat Network content from the Red Hat hosted servers on the Internet. To ensure this, the Dell team created a simple script for calling the Red Hat Network update that includes the enablement and disablement of the NIC as part of the update download process.

Script execution includes downloading all entitled channels. The initial download of channel packages can take several hours, depending on the connection speed, because there are up to 3 GB of packages per channel. Subsequent channel updates download only the incremental changes and usually complete in less than an hour, depending on connection speed.

For the initial channel synchronization, loading the channel content through Red Hat Network CD-ROMs is recommended. The Red Hat Network Web site provides CD-ROM ISO image files that can be used to create the CD-ROMs. Using the CD-ROMs can greatly reduce the time required for the initial channel synchronization. The Satellite Server can still download incremental updates from the Internet if the initial process is performed using CD-ROMs.

Once the channel content has been loaded onto the Satellite Server, the server is ready for use. Administrators should enter the fully qualified host name of the Satellite Server into a Web browser to launch the administration tool. A Satellite Server administrator account must be created the first time the server is used; this account information will be required for entry on subsequent visits.

Updating and managing servers with the Satellite Server

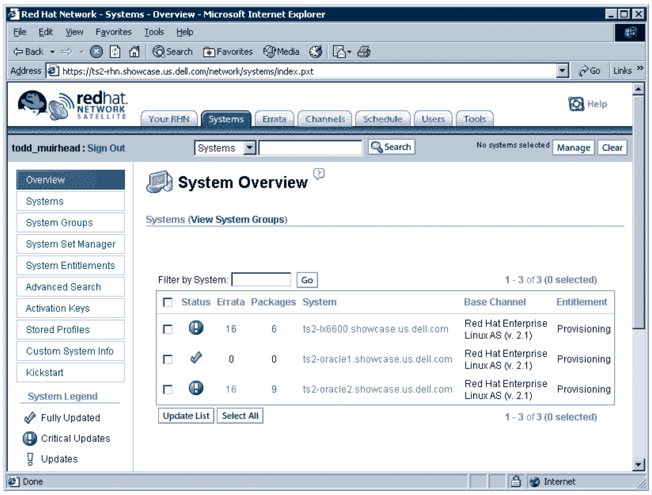
New clients can be added to the Red Hat Network Satellite Server by using the bootstrap script that was created during the Satellite Server installation. The bootstrap script for Red Hat Enterprise Linux 2.1 is bootstrap-AS-2.1.sh, and the script for Red Hat Enterprise Linux 3 is bootstrap.sh. The following command should be executed on the client with the appropriate script name:

```
wget -O - https://Satellite_Server_Name/pub/
   bootstrap/bootstrap script | /bin/bash
```

This command will download the script and then execute it in a bash shell. While this command is executed on each client, the bootstrap script updates the necessary Red Hat Network components, registers each system with the Satellite Server, and performs the initial update with the channel content to which the system has been subscribed. Any configuration files for already installed packages will be saved with .rpmsave or .rpmnew added to the end of the filename. In the Dell lab, the update of a system that was running the initial version of Red Hat Enterprise Linux AS 2.1 was completed in less than 5 minutes, even though the server required more than 100 updates.

Using the Satellite Server model, the Dell lab was able to completely automate the deployment of the latest Red Hat patches and updates to its Red Hat Enterprise Linux–based servers.

If the bootstrap script is not used to accomplish all these tasks, administrators can perform them individually. The *Red Hat Network Client Configuration Guide* and the *Red Hat Network Satellite Server 3.2 with Embedded Database Installation Guide* provide detailed instructions on the specific steps



□	Status	Errata	Packages	System	Base Channel	Entitlement
□	●	16	6	ts2-lx6600.showcase.us.dell.com	Red Hat Enterprise Linux AS (v. 2.1)	Provisioning
□	✓	0	0	ts2-oracle1.showcase.us.dell.com	Red Hat Enterprise Linux AS (v. 2.1)	Provisioning
□	●	16	9	ts2-oracle2.showcase.us.dell.com	Red Hat Enterprise Linux AS (v. 2.1)	Provisioning

Figure 3. Managing system updates with the Red Hat Network Satellite Server GUI

that must be taken to perform these tasks individually.[1]

Clients that have been added to the Red Hat Network Satellite Server will now appear in the management console where their current update status can be viewed (see Figure 3).

Using custom channels to deploy other updates

In addition to the default channels, custom channels can be added to the Satellite Server. Custom channels can contain packages that are not standard Red Hat packages, but they must be in the RPM format. If the software to be distributed and managed is not currently an RPM, then the `rpmbuild` command can be used to create an RPM. Several third-party software packages are already offered as RPMs and thus, in many cases, administrators will not be required to build one.

In the Dell lab, many of the servers are connected to a Dell/EMC storage area network (SAN). Every system attached to the SAN must have EMC® Navisphere® Agent software installed. This agent allows for communication between the Dell/EMC storage arrays and the server. A custom channel was created to provide the Navisphere Agent to the SAN-connected servers.

The default channels are known as parent channels; any custom channels that are created must be a child channel of one of the parent channels. In the Dell lab, a child channel called naviagent was created within the Red Hat Enterprise Linux AS 2.1 parent channel using the Manage Software Channels page of the Satellite Server administration tool. After clicking the Create New Channel button, the Dell team selected the parent channel name. Navisphere Agent was then specified as the new channel name and naviagent as its channel label.

To upload an RPM into a custom channel, administrators must use the `rhnpush` command. The Dell team downloaded the most recent Navisphere Agent RPM from EMC to the Satellite Server and used the following command to upload it into the naviagent channel:

```
rhnpush -c naviagent -nosig -server
    http://Satellite_Server_name/APP naviagent-
    6.4.0.5.0-1.i386.rpm
```

The `-nosig` option was used because this RPM was not signed with a security key. Multiple packages can be uploaded simultaneously simply by listing them at the end of the command. Administrators can create a digital signature for custom-created channels to ensure that packages did in fact come from the Satellite Server.

The Dell team added the naviagent channel for the SAN-connected servers. The next time the servers connected to the Satellite Server for updates, they received the new Navisphere Agent as well as any new Red Hat updates.

Keeping Red Hat Enterprise Linux systems up-to-date

The Red Hat Network Satellite Server model can provide an excellent way to manage updates and software deployments for Red Hat Enterprise Linux systems. Using the Satellite Server model, the Dell lab was able to completely automate the deployment of the latest Red Hat patches and updates to its Red Hat Enterprise Linux–based servers. Additionally, the ability to add custom channels to the Satellite Server allowed the lab to deploy and update other software needed on the servers. Besides providing an efficient method for updating the lab's Red Hat Enterprise Linux–based servers, the Satellite Server model can potentially save numerous hours of effort each time a new update is required compared to traditional manual methods.

Todd Muirhead (todd_muirhead@dell.com) 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.

Peter Lillian (plillian@redhat.com) is a sales engineer and business development manager for Red Hat, Inc. Peter has a B.S. in Engineering Physics from the University of Oklahoma and a Masters of Business Administration/Technology Management from the University of Phoenix. He is a Red Hat Certified Engineer® (RHCE®).

FOR MORE INFORMATION

Red Hat Network:
https://rhn.redhat.com

[1] These guides are available from the Red Hat Network Web site (https://rhn.redhat.com). To access them, administrators must log in and obtain entitlement by visiting the Help section.

Harnessing PXE Boot Services

in Linux Environments

The Preboot Execution Environment (PXE) is part of the Wired for Management (WfM) specification developed by Intel and Microsoft. This article describes the PXE process and explains how to set up a PXE environment on servers that run the Linux® operating system.

BY JOHN HULL, ROBERT HENTOSH, AND ROGELIO NORIEGA

Most computers sold today are compliant with the Wired for Management (WfM) specification. Initially created by Intel, WfM later became part of the Intel® and Microsoft® PC 98 specification, which defines minimum system requirements for PCs that run the Microsoft Windows® operating system (OS). The Preboot Execution Environment (PXE) is part of the PC 98 specification; it permits the system's BIOS and network interface card (NIC) to bootstrap a computer from the network. *Bootstrapping* is the process by which a system loads the OS into local memory so that it can be executed by the processor.

During a typical system startup, a boot program is loaded and executed before the much larger and more complex OS is loaded. The boot program is usually located on the first few blocks of the local hard disk. After the power-on self-test (POST), the system's BIOS routines load the boot program from the boot drive into memory. The boot program then reads the OS kernel from the hard disk into memory using BIOS calls.

PXE extends system startup functionality to allow the network to act as the medium from which the OS kernel can be loaded. This capability can simplify server deployment and maintenance for network administrators by allowing them to perform the following functions:

- Temporarily boot a server in a diagnostic environment, in which diagnostic tools can be downloaded from a remote server onto a local RAM disk
- Perform a network installation of the OS onto a server's local hard drive
- Boot a server into a network-mounted OS where no local hard disks are needed
- Boot a server into an OS downloaded from the network so that data can be recovered from a hard drive that fails to boot

This article explains the basic functionality of PXE as well as how PXE can be used in enterprise environments that run the Linux® OS.

Understanding how PXE works

When a system is powered on, the BIOS stored in the system's ROM goes through a POST. The POST checks memory and processors, and sets up the controllers and bus bridges that enable the system to operate. The BIOS will then load the boot program, which in turn loads the OS kernel. Using PXE, instead of loading the boot program from the hard drive the BIOS uses Dynamic Host Configuration Protocol (DHCP) to obtain an IP address for the network interface and to locate the server that stores the network bootstrap program (NBP). Like the boot program (or boot loader) in the hard drive environment, the NBP is responsible for loading the OS kernel into memory so that the OS can be bootstrapped.

The PXE specification provides IT administrators with a powerful industry standard that can be used to load an OS onto a system across the network.

In a PXE environment, the first requirement is to obtain an IP address for the network card, which is accomplished using a DHCP discover message. The message is broadcast on the local subnet, a DHCP server on the local subnet responds with a DHCP offer, and the DHCP server then assigns the IP address to the interface. The client responds to the server to acknowledge that it has received the packet and is using the IP address.

This exchange also includes information from the DHCP server about the IP address and file name of the NBP; the client uses that information to download the NBP over the network using Trivial FTP (TFTP). The client then executes the NBP, which helps load the OS kernel. The NBP is equivalent to GRUB (GRand Unified Bootloader) or LILO (LInux LOader)—loaders typically used in local booting. The NBP does not need to reside on the same physical system as the DHCP service.

Configuring a PXE server for a Linux environment

The Red Hat® Enterprise Linux 3 operating system (versions AS and ES) provide the software required to configure and manage a PXE server. These capabilities are provided through four packages: dhcp, tftp-server, syslinux, and redhat-config-netboot. The dhcp package provides network configuration information to clients, including redirecting clients to boot servers as requested. The tftp-server package provides capabilities that allow a client to download the NBP from the boot server. The syslinux package provides the NBP (pxelinux.0) that clients download through the TFTP server. And finally, redhat-config-netboot provides a graphical user interface (GUI) as well as a command-line interface (CLI), both of which can be used to help simplify the configuration of installation options or diskless boot images.

Four primary steps are required to get a PXE server up and running. First, the TFTP server must be enabled and configured to allow downloading of the NBP. Second, the DHCP server must be configured to send clients the address of the TFTP server and the location of the NBP on that server. Third, images for network installation or diskless boot must be created on the PXE server. Finally, individual client systems must be enabled to boot using PXE.

Step one: Configuring the TFTP server

When configuring the TFTP server, first verify that the pxelinux.0 NBP is present in the /tftpboot/linux-install directory. If it is not, copy it to that directory from /usr/lib/syslinux. Next, verify that the pxelinux.0 file is readable by everyone by entering `ls -l /tftpboot/linux-install/pxelinux.0` at a command prompt and changing the permissions on the file if necessary. Finally, enable the tftp service by entering `chkconfig tftp on` at a command prompt.

Step two: Configuring the DHCP server

For a client system to download the NBP, the DHCP server must allow clients to boot from the network, and the DHCP must direct the clients to the NBP on the TFTP server. To do this, several options must be present in the /etc/dhcpd.conf file:

- `allow bootp`: Directs the DHCP server to respond to bootp queries
- `allow booting`: Directs the DHCP server to respond to queries from particular clients
- `next-server` *`TFTP server IP`*: Provides the IP address of the TFTP server that contains the NBP
- *`filename`* `/tftpboot/linux-install/pxelinux.0`: Points to the NBP on the TFTP server

Note that the DHCP server and the TFTP server do not have to reside on the same physical system. If they are on separate systems, the `next-server` option directs clients to the correct system. For a full description of how to configure a DHCP server, refer to "Chapter 25. Dynamic Host Configuration Protocol (DHCP)" in the *Red Hat Enterprise Linux 3: System Administration Guide* (https://www.redhat.com/docs/manuals/enterprise/RHEL-3-Manual/sysadmin-guide/ch-dhcp.html). For more information on PXE-specific DHCP options, refer to the SYSLINUX/PXELINUX home page at http://syslinux.zytor.com/pxe.php.

Step three: Creating images for network installation or diskless boot

The next step is to configure OS installation images or diskless boot images and to create configuration files on the PXE server. To configure OS installation images, Red Hat Enterprise Linux 3 provides two useful methods: the redhat-config-netboot GUI and the pxeos CLI. Both provide simple interfaces to set up client installation images quickly. For a thorough explanation of how to use these tools, refer to "Chapter 14. PXE Network Installations" in the *Red Hat Enterprise*

```
default 0
prompt 1
timeout 100
display msgs/boot.msg
F1 msgs/boot.msg
F2 msgs/general.msg
F3 msgs/expert.msg
F4 msgs/param.msg
F5 msgs/rescue.msg
F7 msgs/snake.msg


label 0
  localboot 1

label 1
  kernel /rhel3/vmlinuz
  append initrd=/rhel3/initrd.img ramdisk_size=10000
ks=nfs:192.168.0.1:/ks/ks.cfg
```

Figure 1. Sample pxelinux configuration file

Linux 3: System Administration Guide (https://www.redhat.com/docs/manuals/enterprise/RHEL-3-Manual/sysadmin-guide/ch-pxe.html).

Step four: Enabling clients to boot using PXE

The final step in preparing an environment for PXE installations is to enable the systems to boot using PXE. Current Dell™ PowerEdge™ servers, Dell Precision™ workstations, and Dell OptiPlex™ desktops have PXE capability built in to their integrated and add-in network interfaces. To enable the PXE booting capability, reboot the system and press the F2 key to enter the system setup menu. Select the Integrated Devices menu item and press Enter. Next, scroll down to the Network Interface Controller menu items and cycle though the available options until "Enabled with PXE" is selected. Save the changes and exit the menu.

To boot using PXE, administrators have several options. For PowerEdge servers, on the initial Dell splash screen during boot, administrators can use the Boot from PXE option by pressing the F12 key. On Dell Precision workstations and OptiPlex desktops, they can press F12 for the Boot Menu option, which presents a choice of boot devices including the integrated NIC. Alternatively, to boot from PXE at every reboot, administrators may alter the boot sequence in the Setup menu to move the "Integrated NIC" boot device earlier in the boot order.

Configuring the client for a PXE boot environment

When a client downloads the pxelinux.0 NBP from the TFTP server, the NBP searches the /tftpboot/linux-install/pxelinux.cfg directory for an appropriate configuration file that tells NBP what to do next. The search proceeds in the following sequence:

1. The NBP converts the client's IP address into an uppercase, hexadecimal string (for example, 10.9.0.1 would be 0A090001).
2. The NBP then looks for a file with that name in the pxelinux.cfg directory. If it does not find a file with that name, it will remove the final digit from the string and look again (using the previous example, it would search for 0A09000).
3. The NBP will continue this process until it finds the file (in the given example, searching for 0A0900, 0A090, 0A09, 0A0, 0A, and finally 0). If it does not find a matching file name, it will look for a file named "default."

This process can be helpful because it allows administrators to specify boot image options tailored to a specific client system or a group of systems using the IP address. The default file and the file named after the uppercase, hexadecimal string both contain the same types of information:

- OS images to boot
- Default OS image to load
- Menus that can be displayed to the client
- Options for displaying a prompt, including how long to display it

The configuration file: Customizing boot options for the client system

Figure 1 shows a sample configuration file that is similar to the grub.conf or lilo.conf file. Each section of the file that starts with `label` specifically points to an image that the client may boot. Choosing option `0` will cause the OS image on the local hard drive to boot. Choosing option `1` will load the kernel and the initrd file specified on the `kernel` and `append` lines, respectively. The paths to the kernel and the initrd file are relative to the location of the pxelinux.0 file (thus the path is actually /tftpboot/linux-install on the TFTP server). Note that as in the grub.conf and lilo.conf files, kernel parameters can be added to the `append` line. For example, to boot using a specific Red Hat kickstart file, administrators can use the `ks=` option. Other kernel parameters also can be used, such as `noapic`, `ide=nodma`, or any other desired kernel parameter.

The capability to load an OS from the network can simplify server deployment and maintenance for network administrators.

The `prompt 1` command directs the NBP to prompt the user to choose a boot image, and to time out after the number of seconds specified by the timeout parameter. If the timeout is reached or the `prompt` parameter is not specified or set to 0, the boot image uses the default parameter.

The `display` parameter tells the NBP the location of a custom message to display to the user if the `prompt` parameter is specified. Once this message is displayed, the `F1` through `F7` parameters specify the messages that will be displayed subsequently if the user at the client system presses any of those keys. Note that, once again, the path to the message files is relative to the location of the pxelinux.0 file.

Once the user receives the `boot` prompt at the client system, kernel parameters also can be specified at this prompt when choosing the desired label to boot. For example, the `rescue` parameter allows system administrators to enter rescue mode remotely to help speed system recovery. Other options, such as kickstart file location or specific kernel options, may be specified as well.

> Current Dell PowerEdge servers, Dell Precision workstations, and Dell OptiPlex desktops have PXE capability built in to their integrated and add-in network interfaces.

Enabling remote system administration across the network

The Preboot Execution Environment specification provides IT administrators with a powerful industry standard that can be used to load an OS onto a system across the network. This specification comes standard on NICs that are provided within current Dell servers, workstations, and desktops. Red Hat Enterprise Linux 3 provides software and configuration tools that help enable administrators to configure a PXE server. Combining Dell hardware with Red Hat Enterprise Linux 3 can help IT administrators to build a powerful enterprise data center that is designed to perform remote OS installations, OS booting on diskless systems, and remote recovery of failed systems.

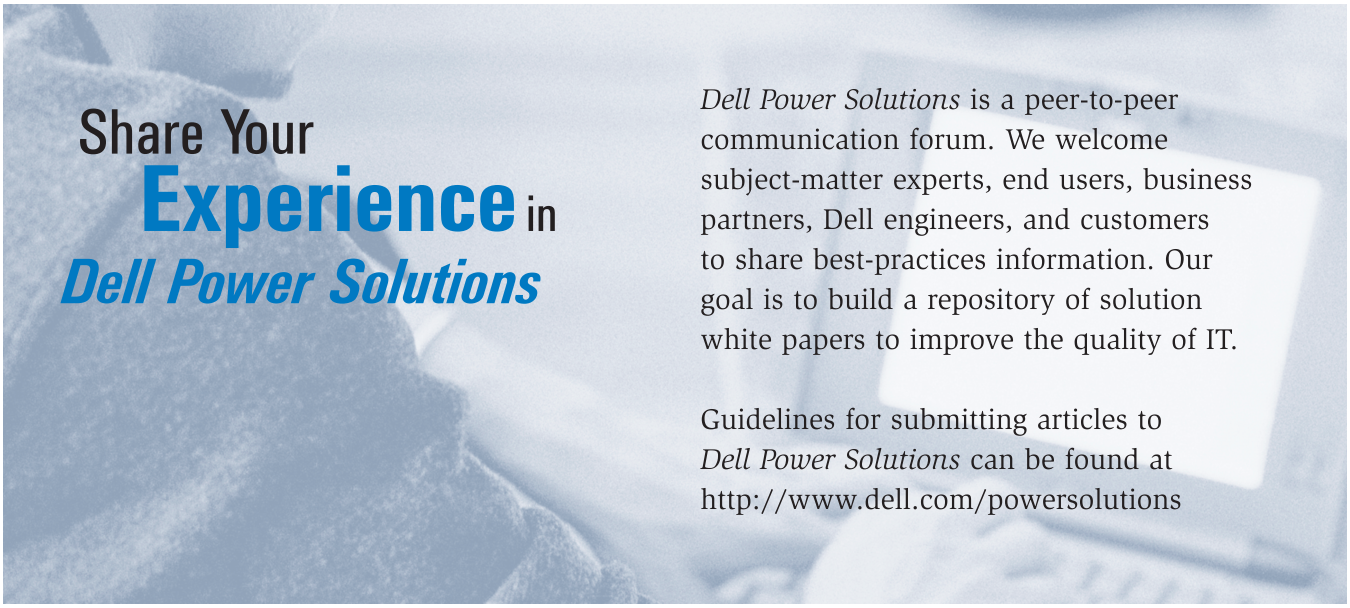
John Hull (john_hull@dell.com) is a software engineer at Dell and is the lead Linux engineer for Dell Precision workstations. He has a B.S in Mechanical Engineering from the University of Pennsylvania and an M.S in Mechanical Engineering from the Massachusetts Institute of Technology (M.I.T.).

Robert Hentosh (robert_hentosh@dell.com) is a senior consultant on the Dell Linux Engineering team. Before joining Dell, he worked for 12 years in the software industry, with experience in device driver development, communications interfaces, and Web-based applications. Robert has a B.S. in Electrical Engineering and a B.A. in Physics, both from Bucknell University.

Rogelio Noriega (rogelio_noriega@dell.com) is a senior analyst on the Dell Linux Engineering team and a technician for Dell Precision workstations.

FOR MORE INFORMATION

Dell and Linux:
http://www.dell.com/linux



Using
Linux Logical Volume Manager
on Dell PowerEdge Servers

To enable fast response to changing business needs in heterogeneous storage environments, administrators can use Red Hat® Linux® Logical Volume Manager (LVM) software to dynamically resize logical volumes. LVM can also be used to create snapshots, which help keep critical business data highly available.

BY MATTHEW WYGANT

With enterprise data growing at an ever-increasing rate, IT organizations require powerful tools to manage heterogeneous storage environments. Red Hat® Linux® Logical Volume Manager (LVM) software enables administrators to reallocate storage flexibly to meet changing business demands—enabling administrators to manage groups of disks that comprise multiple, individual disks (known as volume groups) rather than individual disks and disk partitions. In addition, LVM helps keep data highly available through a snapshot capability that enables the rapid creation of offline copies of stored data.

Red Hat began including LVM packages in its operating systems with Red Hat Linux Professional 8.0, and the updated version of Red Hat LVM has recently become part of the Red Hat Enterprise Linux 3 distribution. LVM version 1.0.5 is highly stable and provides many advanced features that are useful to Linux administrators, such as the capability to dynamically resize disks to accommodate growing storage needs. The snapshot capability, which is also new to LVM 1.0.5, offers built-in snapshot options for read-only data that can be used for backups.

To understand the usefulness of Red Hat Linux LVM in Dell™ storage and server environments, in March 2004 Dell support engineers configured a Dell PowerEdge™ 1750 server with an on-board PowerEdge Expandable RAID Controller 4, Dual Channel integrated (PERC 4/Di) adapter and a PERC 4/Dual Channel (DC) Peripheral Component Interconnect (PCI) adapter connected to a Dell PowerVault™ 220S enclosure. The storage enclosure was equipped with fourteen 73 GB drives constructed in four large RAID-5 containers. To replicate the test configuration described in this article, administrators must first create partitions on the disk. If a partition table does not already exist on the drive, enter the following command at the command prompt in the Linux operating system:

```
fdisk /dev/disk_name
```

To reread the disk partition table, enter:

```
partprobe
```

Setting up Logical Volume Manager on Linux

The steps to install LVM are fairly simple and require knowledge of just a few basic concepts. LVM comprises three main layers: physical volumes, volume groups, and logical volumes.

Physical volumes. The physical volume (PV) consists primarily of the underlying physical storage: for example, a Dell PowerVault 220S SCSI chassis with disks, a Fibre

Channel storage device, or even the internal disks on any Dell PowerEdge server. To participate in a PV, the physical disks must be initialized using `pvcreate`. In the scenario for this article, PVs reside on the devices /dev/sdb1, /dev/sdc1, /dev/sdd1, and /dev/sde1.

To start LVM, use the `vgscan` command, which creates the /etc/lvmtab file:

```
vgscan
pvcreate /dev/sdb1 /dev/sdc1 /dev/sdd1 /dev/sde1
```

Volume groups. A volume group (VG) resides on top of the physical layer; the operating system views the VG no differently than it would a regular SCSI or IDE disk. By creating two different VGs, administrators can contrast the overhead of LVM, see how added disks help distribute the workload, and speed up disk writes and reads.

To create a VG, use the `vgcreate` command:

```
vgcreate vg_one /dev/sdb1
vgcreate vg_two /dev/sdc1 /dev/sdd1 /dev/sde1
```

Logical volumes. A logical volume (LV) works the same way as partitions on standard disks. Logical volumes are created using the `lvcreate` command. When creating LVs, administrators can specify sizes two ways: One method uses megabytes and the other uses physical extents (PE), which by default are 4 MB blocks of disk space. With the default size, administrators can create a 2 GB LV by using 512 physical extents.

When creating an LV, use the `lvcreate` command with the `-l` switch to specify the number of physical extents:

```
lvcreate -l 512 -n lv_one vg_one
```

For convenience, administrators often prefer to indicate a megabyte size in the `lvcreate` command instead of the number of physical extents. The `-L` switch enables them to do so:

```
lvcreate -L 2048 -n lv_two vg_two
```

Note: In the preceding two sample commands, the switches (`-l` and `-L`) are case sensitive.

After creating the two LVM partitions, format them as ext3 file systems using the `mke2fs` command.[1]

```
mke2fs -j /dev/vg_one/lv_one
mke2fs -j /dev/vg_two/lv_two
```

The partitions can now be defined in the /etc/fstab file:

```
/dev/vg_one/lv_one /lvm1 ext3 defaults 1 2
/dev/vg_two/lv_two /lvm2 ext3 defaults 1 2
```

Once the devices are input, mount them using the `mount -a` command, and then use the `df -h` command to output their partition sizes. LVM addresses physical disks in a sophisticated manner, so adding or removing a single physical disk from the server does not prevent LVM from starting on boot. When administrators issue the `pvcreate` command, unique drive information is written to each disk that has been initialized. LVM relies on finding this unique information on boot when `vgscan` is run from the /etc/rc.sysinit file. These steps are all done before the /etc/fstab file is used on boot; thus, administrators can safely mount the file systems through the /etc/fstab file.

LVM performance. To demonstrate the performance of logical disks created by LVM versus physical disks, administrators can use the data test (dt) package[2] to execute test runs using the following commands:

```
dt of=/lvm1/testfile1 bs=2k limit=4G runtime=1h
   log=~/lvm1.log
dt of=/lvm2/testfile2 bs=2k limit=4G runtime=1h
   log=~/lvm2.log
```

LVM also includes its own tools for collecting and reporting statistics. The `lvmsadc -v` *`log_name`* command can be used to collect I/O statistics, and the resulting log file can be viewed with `lvmsar -fv` *`log_name`*.

Resizing logical volumes

Before resizing LVs, administrators should create a current backup of their data to safeguard against file system corruption. Several commands can be used to resize an LV, but each has a caveat. The `e2fsadm` command is generally regarded as the most comprehensive for checking the ext2 and ext3 file systems to ensure that no corruption occurs while resizing. The `e2fsadm` command requires that the LV be unmounted. (Other resize commands will run on a mounted file system, but data corruption is a risk.) The `e2fsadm` command automatically spawns the `lvextend` or `lvreduce`, `e2fsck`, and `resize2fs` commands. Although the `lvextend` and `lvreduce` commands can be used directly, both alter the size of the LV while mounted and this can increase the risk of data corruption.

[1] For more information about the Red Hat ext3 journaling file system, visit http://www.redhat.com/support/wpapers/redhat/ext3.

[2] Data test is a generic program that helps administrators verify proper operation of peripherals, file systems, device drivers, or any data stream supported by the operating system. For more information about the dt program, visit http://www.bit-net.com/~rmiller/dt.html.

```
umount /lvm2
e2fsadm -L +4096 /dev/vg_two/lv_two
mount /lvm2
```

Creating LVM snapshots

A snapshot is a point-in-time image of all the data that resides on the LV. Although a snapshot cannot be written to, it is highly useful for reporting purposes, backups, and disaster recovery. Administrators can also use a snapshot to provide temporary access to data while resizing the LV—a convenient feature that maintains storage uptime while allowing administrators to dynamically alter the subsystem.

A snapshot requires the same amount of space that the data occupies on the original LV. To create a snapshot, use the `lvmcreate` command:

```
lvmcreate -L 4096M -s -n lv_two_snapshot
   /dev/vg_two/lv_two
mount /dev/vg_two/lv_two_snapshot /lvm2_snapshot
```

To remove a snapshot, use the `lvremove` command:

```
umount /lvm2_snapshot
lvremove /dev/vg_two/lv_two_snapshot
```

Monitoring the storage

LVM includes commands that enable administrators to monitor all levels of the LVM disk subsystem. The `pvdisplay` command provides a view of physical disks in a PV, the `vgdisplay` command shows VG information, and the `lvdisplay` command displays information about each LV. This information can also be viewed from the /proc/lvm directory as needed.

The `pvdisplay` output shows how much space is available on the physical drive using physical extents.

```
pvdisplay /dev/sdc1

—- Physical volume —-
PV Name             /dev/sdc1
VG Name             vgtest
PV Size             251.02 MB [514080 secs] / NOT
                    usable 4.19 MB [LVM: 128 KB]
PV#                 1
PV Status           NOT available
Allocatable         yes (but full)
Cur LV              1
PE Size (KByte)     4096
Total PE            61
Free PE             0
Allocated PE        61
PV UUID             W4VuBU-ovTc-TF4E-a0dj-cdiz-
                    79MN-mQB0ZT
```

When the `vgdisplay` command is used, free space is shown in both physical extents and megabytes.

```
vgdisplay vg_one

—- Volume group —-
VG Name             vg_one
VG Access           read/write
VG Status           available/resizable
VG #                0
MAX LV              256
Cur LV              1
Open LV             0
MAX LV Size         255.99 GB
Max PV              256
Cur PV              2
Act PV              2
VG Size             496 MB
PE Size             4 MB
Total PE            124
Alloc PE / Size     100 / 400 MB
Free  PE / Size     24 / 96 MB
VG UUID             D9vqOH-6NYR-2eZ0-pgxb-nP6d-
                    Qafg-kDD5ZG
```

Finally, the `lvdisplay` tool shows the LV data sizes in both physical extents and megabytes.

```
lvdisplay /dev/vg_one/lv_one

—- Logical volume —-
LV Name             /dev/vg_one/lv_one
VG Name             vg_one
LV Write Access     read/write
LV Status           available
LV #                1
# open              0
LV Size             400 MB
Current LE          100
Allocated LE        100
Allocation          next free
Read ahead sectors  1024
Block device        58:0
```

Simplifying storage management

The inclusion of the 1.0.5 version of LVM into Red Hat Enterprise Linux products can equip administrators in data center and server environments with a useful tool to grow storage in concert with the needs of the organization. The capabilities to dynamically resize logical volumes, create snapshots for backups, and monitor all levels of the LVM disk subsystem help administrators make storage available when and where it is needed throughout the enterprise.

Matthew Wygant (matthew_wygant@dell.com) is a network engineer and consultant in the Enterprise Expert Center Network Operating System Support group at Dell. He is a Red Hat Certified Engineer® (RHCE®) and works with Red Hat and other Linux distributions.

Industry Standards for Managing the HPC Cluster Life Cycle

For organizations deploying high-performance computing (HPC) clusters, reducing total cost of ownership, maximizing cluster uptime, and expanding remote manageability are among the baseline IT requirements. This article discusses how LM sensor management, Wired for Management (WfM), Intelligent Platform Management Interface (IPMI), and other industry standards can help IT professionals manage the overall HPC cluster life cycle.

BY YUNG-CHIN FANG AND RIZWAN ALI

The design of a high-performance computing (HPC) cluster environment typically takes into account several requirements, including reduced total cost of ownership (TCO), maximized cluster uptime, optimized management efforts, and expanded remote manageability. Several industry-standard management specifications are designed to address these requirements, enabling many of the basic building blocks that IT organizations need to consider as they design an underlying HPC cluster environment. To create a cluster with the desired state of manageability, architects must understand these management specifications. This article discusses how industry standards, specifications, and implementations can affect the management of an HPC cluster throughout its life cycle.

Understanding key management specifications

To interface with server hardware, cluster monitoring and management applications use mainstream management standards and specifications such as LM sensor management,[1] Wired for Management (WfM),[2] Intelligent Platform Management Interface (IPMI),[3] and Extensible Firmware Interface (EFI).[4]

LM sensors manage hardware health

The core idea of LM sensor management revolves around using a dedicated management processor to monitor and manage hardware health conditions, so that the sensor monitoring and management effort will not consume host resources. When the management processor communicates with on-board sensors through its dedicated system management bus, the management traffic does not consume system (address, data, and control) bus bandwidth. The most common implementations of LM management features include monitoring CPU temperature and cooling fan speed.

LM sensor management was widely implemented in the 1990s and has since become a de facto standard, but has left room for independent implementation of this

[1] For more information about LM sensor management, see “Industry-Standard Specifications for Cluster Management” by Yung-Chin Fang; Jenwei Hsieh, Ph.D.; Victor Mashayekhi, Ph.D.; and Reza Rooholamini, Ph.D., in *Dell PowerSolutions*, Issue 3, 2001.

[2] For more information about the Wired for Management initiative, visit http://www.intel.com/labs/manage/wfm.

[3] For more information about the Intelligent Platform Management Interface, visit http://www.intel.com/design/servers/ipmi/index.htm.

[4] For more information about the Extensible Firmware Interface, visit http://www.intel.com/technology/efi/index.htm.

standard. Many varieties of LM-compliant management sensors and processors are available; more than ten semiconductor companies produce more than 40 unique implementations of LM chips. However, not all platforms can remotely monitor and manage hardware health conditions. As a result, cross–operating system (OS) and cross-platform interoperability with LM sensors can be problematic.

WfM provides remote management

The WfM specification defines services that enable remote server management in the deployment and operation phases of the cluster life cycle (see "Following the HPC cluster life cycle" for more information). WfM incorporates the Advanced Configuration and Power Interface (ACPI),[5] remote wakeup[6] (also known as Wake on LAN, or WOL), and Preboot Execution Environment (PXE)[7] specifications—which are typically helpful in the deployment phase—as well as interfaces to several popular management standards.

ACPI. The ACPI specification allows administrators to remotely power up a new cluster that does not have a resident OS, or to power cycle a hung node. ACPI also can be used with the OS to provide OS-directed power management (OSPM) functions. In the OSPM model, the OS determines when to provide power management and the BIOS determines how to provide it. Some manufacturers implement the ACPI specification on their platforms, whereas other manufacturers choose to implement the Advanced Power Management (APM) specification[8] or a proprietary remote power management specification.

Remote wakeup. The remote wakeup specification complements the ACPI power management feature. Remote wakeup enables cluster management tools to wake up a node by sending it a remote-wakeup packet. This process requires that the cluster node support either the ACPI or the APM specification and that the node's network interface card (NIC) support the remote-wakeup feature.

PXE. Typically, after a new node is powered up remotely through implementations of remote wakeup and ACPI, an OS is deployed to that node. PXE, which is usually implemented in the NIC option ROM and in the system BIOS, allows a server to boot remotely without a local OS. When a server boots, it loads the PXE code and sends out a Dynamic Host Configuration Protocol (DHCP) request to a remote boot server asking for an IP address. Once the server receives the IP address, its PXE routine interacts with the remote boot server to dynamically retrieve the requested boot image over the network. This functionality allows administrators to remotely install the OS and applications, and to remotely configure a new cluster without the presence of a technician.

PXE allows administrators to remotely install the OS and applications, and to remotely configure a new cluster without the presence of a technician.

One of the most popular cluster computing packages, Open Source Cluster Application Resources (OSCAR),[9] uses PXE as a building block for the main remote-system deployment mechanism. In fact, PXE is widely used by IT departments to deploy and update images for mid- to large-scale computers.

Other specifications. WfM also includes interfaces to other management standards such as Boot Integrity Services (BIS),[10] Common Information Model (CIM),[11] Desktop Management Interface (DMI),[12] Network PC System Design Guidelines,[13] Simple Network Management Protocol (SNMP),[14] System Management BIOS (SMBIOS),[15] Web-Based Enterprise Management (WBEM),[16] and Windows® Management Instrumentation (WMI).[17] These standards provide OS-level interoperability.

IPMI enables cross-platform management

Interoperability among management components is important when a cluster comprises nodes from multiple vendors. WfM is designed to facilitate hardware management within the in-band management fabric; however, the in-band management traffic can potentially degrade cluster performance for certain communication-intensive

[5] For more information about the Advanced Configuration and Power Interface, visit http://www.acpi.info.

[6] For more information about remote wakeup, download the Wired for Management Baseline Version 2.0 specification at ftp://download.intel.com/labs/manage/wfm/download/base20.zip.

[7] The Preboot Execution Environment Version 2.1 specification is available at ftp://download.intel.com/labs/manage/wfm/download/pxespec.pdf.

[8] For more information about the Advanced Power Management v. 1.2 interface, visit http://www.microsoft.com/whdc/archive/amp_12.mspx.

[9] For more information about OSCAR, see "Open Source Cluster Application Resources (OSCAR): Design, Implementation and Interest for the [Computer] Scientific Community" by Benoît des Ligneris, Stephen Scott, Thomas Naughton, and Neil Gorsuch. Presented at The 17th Annual International Symposium on High-Performance Computing Systems and Applications, http://hpcs2003.ccs.usherbrooke.ca/papers/desLigneris_01.pdf.

[10] For more information about Intel Boot Integrity Services, visit http://www.intel.com/labs/manage/wfm/tools/bis.

[11] For more information about Common Information Model standards, visit http://www.dmtf.org/standards/cim.

[12] For more information about Desktop Management Interface standards, visit http://www.dmtf.org/standards/dmi.

[13] The Network PC System Design Guidelines are available at http://sunsite.rediris.es/sites/download.intel.nl/design/netpc/NETPC.PDF.

[14] For RFC 1270—SNMP Communications Services, visit http://www.faqs.org/rfcs/rfc1270.html.

[15] For more information about the System Management BIOS specification, visit http://www.dmtf.org/standards/smbios.

[16] For more information about the Web-Based Enterprise Management initiative, visit http://www.dmtf.org/standards/wbem.

[17] For more information about Windows Management Instrumentation, visit http://msdn.microsoft.com/library/default.asp?url=/library/en-us/wmisdk/wmi/wmi_start_page.asp.

applications. To help solve that problem, a joint industry effort resulted in the development of IPMI—an out-of-band, cross-platform management specification.

IPMI defines the common commands, data structures, and message formats for all IPMI interfaces. IPMI also defines common management functions, including how the system event log (SEL) and sensor data records (SDRs) are managed and accessed; how the system interfaces work; how sensors operate; how control functions such as system power-up, power-down, and reset are initiated; and how the IPMI host-system watchdog timer operates.

IPMI has two supporting specifications: Intelligent Platform Management Bus (IPMB) and Intelligent Chassis Management Bus (ICMB). IPMB defines an internal management expansion bus that is typically used to link chassis management features with the motherboard management subsystem. ICMB defines the dedicated external management bus between IPMI-enabled platforms.

In the IPMI architecture, the central intelligence is provided by a microcontroller called the Baseboard Management Controller (BMC). The BMC operates on standby power and can automatically poll system health status. When the BMC detects any predefined exception or threshold violation, it can react to the violation condition based on preset rules—for example, log the event, generate alerts, and perform error recovery schemes such as power cycling. IPMI works by specifying common, abstracted message-based interfaces to the BMC. This abstraction isolates software from the hardware implementation and provides further interoperability.

The design of an HPC cluster environment typically takes into account several requirements, including reduced TCO, maximized cluster uptime, optimized management efforts, and expanded remote manageability.

The BMC also manages the storage of SDRs, the SEL, and field replaceable unit (FRU) information in nonvolatile memory. The SDRs describe the number and type of monitoring and control capabilities available in a given platform. These records allow software to discover and automatically adapt to the monitoring and control features offered by each platform. The FRU includes serial numbers and part numbers used to identify different serviceable or failed entities in a system.

IPMI 1.5 builds on the proven technology from IPMI 1.0 and exposes the same capabilities through new interfaces, allowing both local and remote software to automatically configure itself for multiple systems. This feature facilitates the creation of cross-platform management software. IPMI 1.5 added many features to the previous version, including:

- Serial over LAN (SOL), which uses IPMI messages encapsulated in User Datagram Protocol (UDP) packets
- IPMI over a serial/modem connection, including Basic Mode for highest speed with automated remote consoles, Point-To-Point Protocol (PPP) Mode, and Terminal Mode for limited access by dumb terminals in legacy environments
- Platform Event Filtering (PEF) to generate selectable actions when a new event matches a configurable set of event filters
- LAN alerting, which sends SNMP traps in the Platform Event Trap (PET) format to a specified destination
- Serial/modem alerting, which includes numerical paging, Telocator Alphanumeric Protocol (TAP) paging, and PPP alerting
- Alert policies that support alerts to multiple destinations
- Serial port sharing for enabling a single serial connector to be shared between the motherboard's serial controller and the serial connection to the BMC
- Boot option to direct the system boot process
- Support for a Peripheral Component Interconnect (PCI®) management bus, which defines commands and a protocol for sending and receiving IPMI messages over the proposed PCI management bus
- Definition of user privileges and authentication access to the serial/modem and LAN interfaces

More than 165 companies[18] have adopted or promoted IPMI, which was introduced in 1998. Designers can select a management processor for implementing IPMI in their platforms, and these platforms can provide varying levels of interoperability through the management framework and through different protocols.

EFI provides standard boot environment

All 64-bit Intel® architecture (IA-64) and some 32-bit Intel architecture (IA-32) platforms implement EFI. Defined by Intel, EFI creates a layer between the host OS and the platform firmware. It is implemented as a pre-OS boot environment and can be used to load EFI-level drivers. EFI provides boot and service calls to OS loaders and operating systems, so designers can implement an OS loader with little knowledge of hardware and firmware.

Following the HPC cluster life cycle

The HPC cluster life cycle consists of four phases: design, deployment, operation, and retirement (see Figure 1). Differently scaled HPC clusters have disparate needs and considerations in all

[18] For a list of IPMI industry promoters, adopters, and contributors, visit http://developer.intel.com/design/servers/ipmi/adopterlist.htm.

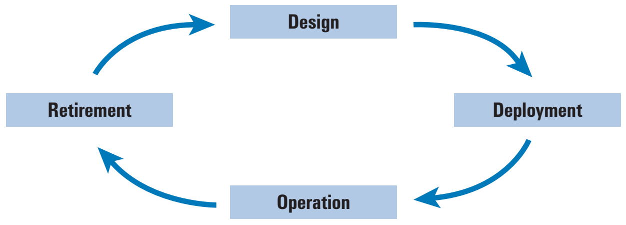


Figure 1. The HPC cluster life cycle

phases. For that reason, administrators must implement management specifications properly according to the size of their cluster installations.

Design phase: Investigating technologies and requirements

Designers must first understand the original problem that the cluster is required to solve, and then create the corresponding cluster based on that problem. Design phase tasks include understanding candidate platform architectures, understanding management specification implementation, and selecting the best-fit cluster architecture and components.

Designers can change the cluster software stack after a cluster is deployed, but they cannot change the implementation of hardware-based management specifications. Consequently, understanding industry management standards thoroughly and selecting the best-fit implementation are key tasks in this phase. Decisions made in this phase will affect all other phases.

Beyond knowing the specifications involved, designers must also understand the differences among implementations. For example, SOL can be implemented either using a shared NIC or using a dedicated NIC, and both implementations are IPMI compliant. For very large-scale clusters, the shared-NIC implementation has the potential to generate noticeable network traffic over the shared fabric. A dedicated-NIC implementation of SOL can avoid management traffic, but this implementation also increases the cost for a dedicated out-of-band management fabric.

The same results can be obtained using different specifications and implementations. For example, both EFI and IPMI implementations enable the examination of hardware health conditions without an OS. The following sections discuss the effects derived from selecting different management specifications and implementations.

Deployment phase: Constructing the cluster

The deployment phase includes racking and stacking hardware components, interconnecting all the in-band and out-of-band fabrics, enabling the cluster architecture to remotely verify the health condition of the bare-metal hardware, and deploying and configuring the cluster-computing software stack.

Certain specifications can obtain the SDRs and SEL remotely without a host OS; other specifications can enable remote firmware-level diagnostics without a host OS. When these types of specifications and an out-of-band implementation are selected, then the health of remote cluster hardware can be verified without the use of an OS. For mid- to large-scale clusters, this capability can effectively reduce cluster staging time.

During the deployment of the cluster-computing software stack, ACPI and remote wakeup can both be used to remotely power up a node without an OS. This implementation may require enabling the remote-wakeup bit on the NIC. IPMI can also be used to remotely power up a node by enabling the ACPI hardware implementation; this method does not depend on the host's NIC settings.

For the remote deployment of a cluster software stack, certain IA-32 platforms implement PXE, and other platforms provide EFI-level network boot. PXE has a dependency on the NIC option ROM or the system BIOS. The EFI-level network boot has an EFI firmware implementation dependency. In an IA-64 platform, EFI provides a network boot feature that is similar to the NIC-level PXE implementation. Both the PXE and EFI network boot implementations can help save deployment time.

The remote deployment process requires a remote console mechanism for monitoring deployment progress. IPMI defines OS-independent SOL and supports remote pre-OS serial console access over IP, providing Internet-level scalability. Another specification for remote consoles is the Intel Emergency Management Port (EMP). EMP is based on the serial-port connection and uses the standard RS-232 line interface. Pre-OS console access can reduce the difficulties of deploying the remote cluster-computing software stack.

Operation phase: Maintaining cluster operation

The operation phase includes tasks such as maintaining a cluster's operating status and optimizing resource utilization by distributing jobs properly (based on the node's health status acquired from remote management capabilities). An important goal in this phase is reducing overall operational costs by downtime reduction. Downtime can be reduced by preventing hardware failures and by implementing automatic failure recovery and unattended remote firmware provisioning.

Downtime can be reduced by preventing hardware failures and by implementing automatic failure recovery and unattended remote firmware provisioning.

In this phase, an operator can poll SDRs and the SEL from time to time to check the cluster health condition. SDRs list sensor readings and the SEL carries system events that exceed a preset threshold. When exceptional conditions such as memory bit error, fan failure, and high CPU temperature occur and are noticed by a system administrator, the administrator can either schedule a maintenance shutdown or perform a task migration to prevent system failure. Administrators can also perform periodic OS-level or EFI-level remote hardware diagnostics to examine hardware health conditions and react accordingly to prevent hardware failures.

For implementations that use clusters or cluster nodes from different generations and different vendors, interoperability that enables administrators to manage individual nodes through one centralized management console is desirable. In an ideal situation, administrators use a centralized management console to monitor and manage cluster hardware and software conditions, which is achievable using an IPMI and SNMP/CIM implementation. This implementation can save cluster management time and reduce management software costs.

In certain implementations, a node's BMC or management processor will send out a heartbeat signal periodically through the management fabric to notify the centralized management console that this node is alive. Administrators can also use the BMC to monitor the on-board watchdog for auto-recovery of a hung OS. The most common auto-recovery scheme is to automatically power cycle the hung node. Certain BMC implementations can send a PET to a centralized management console for notification if a preset threshold is met while power cycling the hung node. The centralized management console can then e-mail or page an administrator or even execute code to repair the software issue.

In the operation phase, a cluster will typically need to update its firmware for a bug fix or for new features. Updating one piece of firmware usually requires one reboot, resulting in downtime for system maintenance. To reduce system downtime, the remote firmware provisioning capability—using out-of-band fabric to update all firmware in one reboot—is a highly desired feature. Many manufacturers provide software tools for this purpose, but no standardized specification exists; so interoperability for remote firmware provisioning is not currently possible. A next-generation management specification may be able to address this need.

Retirement phase: Migrating jobs to a new cluster

The cluster retirement phase requires all runtime jobs and queued jobs to seamlessly migrate to a new cluster. By selecting the proper implementations of industry standards, administrators can help to reduce deployment time and cost, prevent node hardware failures, reduce cluster downtime, reduce operating cost, and perform smooth migrations of runtime tasks to new hardware.

Administrators can apply a runtime task migration feature in this phase to move live jobs from a retiring cluster to a new cluster. After all jobs migrate to a new cluster, a hardware management utility can be used to poll for a final hardware health status, and then ACPI can be used to shut down the retiring cluster and complete the cluster life cycle.

Achieving remote manageability and interoperability

Industry standards can help to reduce TCO during the deployment and operation phases of a cluster's life cycle. Management specifications defined by industry standards can help provide ever-increasing levels of remote manageability and interoperability, and the hardware industry is continually enhancing these specifications. However, OS-level node and centralized console management specifications are not yet standardized enough to provide full interoperability and scalability.

Management specifications defined by industry standards can help provide ever-increasing levels of remote manageability and interoperability, and the hardware industry is continually enhancing these specifications.

Microsoft has built OS-level hardware management specifications into the Windows OS. These specifications include WMI and integrated management solutions such as Microsoft® Operations Manager (MOM), Automated Deployment Services (ADS), and Systems Management Server (SMS), all of which can be effective tools for cluster management. The Linux® domain does not have a comparable, interoperable common hardware-management specification. To effectively monitor and manage a large Linux cluster, administrators may need to run several stand-alone utilities to acquire needed information. However, many such open source utilities exist.

By investing time during the design phase to better understand the various management specifications and implementations, HPC cluster designers and administrators can properly build a system with the desired level of manageability.

Yung-Chin Fang (yung-chin_fang@dell.com) is a senior consultant of 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 and an M.S. in Computer Science from Utah State University.

Rizwan Ali (rizwan_ali@dell.com) 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.

Exploring the Performance of the Dell PowerEdge 3250 for HPC Cluster Environments

The Dell™ PowerEdge™ 3250 server uses Intel® Itanium® 2 processors to deliver performance advantages for high-performance computing clusters. This article examines the strengths of the Itanium 2 architecture and presents results of industry-standard benchmark performance testing.

BY RAMESH RADHAKRISHNAN, PH.D., AND DAVID ARD

The concept of high-performance computing (HPC) clusters—also named Beowulf clusters after the project name used by the original designers—originated in 1993 at the Center of Excellence in Space Data and Information Sciences (CESDIS), which is located at the NASA Goddard Space Flight Center in Maryland. The goal of the project was to design a cost-effective, parallel computing cluster built from standards-based components to satisfy the computational requirements of researchers in the fields of earth and space sciences. Since then, HPC clusters have gained wide acceptance as a platform for solving complex scientific and technical computing problems. In fact, because HPC clusters are often more economical to purchase than proprietary systems, they are an attractive alternative to supercomputers in the role of high-performance computing.

Organizations evaluate potential HPC compute nodes based on server capabilities and price/performance ratios. The Dell™ PowerEdge™ 3250 offers power and performance geared toward HPC environments.[1] The rack-optimized 2U server supports up to two 64-bit Intel® Itanium® 2 processors, 16 GB of double data rate (DDR) memory, and two internal hard drives with a maximum capacity of up to 292 GB. In addition, the PowerEdge 3250 is compatible with Red Hat® Enterprise Linux® 2.1, Red Hat Enterprise Linux 3, and Microsoft® Windows Server™ 2003 operating systems.

Intel Itanium architecture enables the PowerEdge 3250 to achieve significant performance gains compared to 32-bit and proprietary technologies, specifically in floating-point workloads.[2] Most HPC applications are floating-point intensive, making the PowerEdge 3250 an excellent choice for HPC clusters, as the benchmark results described in this article indicate.[3]

Using the Linpack benchmark to evaluate the PowerEdge 3250

Linpack is a benchmark used to solve a dense system of linear equations. A version of the benchmark that allows

[1] For more information on the PowerEdge 3250 server, visit http://www1.us.dell.com/content/products/productdetails.aspx/pedge_3250?c=us&cs=555&l=en&s=biz.

[2] The Intel Itanium 2 processor holds the top position as of June 14, 2004 for the SPECfp2000 benchmark, which is an industry-standard benchmark to measure the floating-point performance of a processor. Visit http://www.spec.org/osg/cpu2000/results for the latest SPECfp2000 benchmark results.

[3] For more information, visit http://www.dell.com/downloads/global/products/pedge/en/3250_wpapr.pdf.

engineers to scale the size of the problem and optimize the software to achieve the best performance for a given machine has been used since 1993 to rank the world's top supercomputers.[4] The results of this benchmark are measured in floating-point operations per second (FLOPS), which reflect the floating-point capabilities of the processor, the compiler performance, and the capabilities of the software libraries.

Linpack performance on a single server

To demonstrate the floating-point capabilities of Dell PowerEdge 3250 servers, the Dell Enterprise Performance Team ran the Linpack benchmark on two different PowerEdge 3250 server configurations in November 2003 and compared those results to an M&A Technology™ 64-bit Patriot 4400 server configured with AMD™ Opteron™ processors.[5] The server configuration details were as follows:

- A PowerEdge 3250 was configured with dual Intel Itanium 2 processors at 1.5 GHz using 6 MB of level 3 (L3) cache, 8 GB of DDR 200 memory, a 36 GB Ultra320 SCSI controller, and a 15,000 rpm hard drive. The server ran the Red Hat Enterprise Linux AS 2.1 operating system (OS).
- Another PowerEdge 3250 was configured with dual Intel Itanium 2 processors at 1.4 GHz using 1.5 MB of L3 cache, 8 GB DDR 200 memory, a 36 GB Ultra320 SCSI controller, and a 15,000 rpm hard drive. The server ran the Red Hat Enterprise Linux AS 2.1 OS.
- An M&A Technology 64-bit Patriot 4400 server was configured with dual model 244 AMD Opteron processors at 1.8 GHz using 1 MB of level 2 (L2) cache and 4 GB of DDR 333 memory. The server ran the 64-bit SUSE® Linux Professional Edition 8.1 OS with the NUMA kernel.

Figure 1 presents comparative test results running the Linpack benchmark. Performance is shown along the x-axis in gigaflops (1 GFLOPS = 10^9 FLOPS). Theoretical peak performance for each of the processors, which is also shown in Figure 1, indicates the maximum number of floating-point operations that the processor can perform. The efficiency of the processor is defined as actual peak gigaflops divided by theoretical peak gigaflops. Figure 1 indicates these important results:

- The PowerEdge 3250 system configured with dual Intel Itanium 2 processors at 1.5 GHz using 6 MB of L3 cache achieved 94 percent efficiency using the Intel Math Kernel Libraries (MKL) for a peak gigaflop rating of 11.24 GFLOPS.[6]
- The PowerEdge 3250 system configured with dual Intel Itanium 2 processors at 1.4 GHz using 1.5 MB of L3 cache achieved 91 percent efficiency using the MKL for a peak gigaflop rating of 10.18 GFLOPS, as compared to the 6.22 GFLOPS score on the 1.8 GHz Opteron processor.

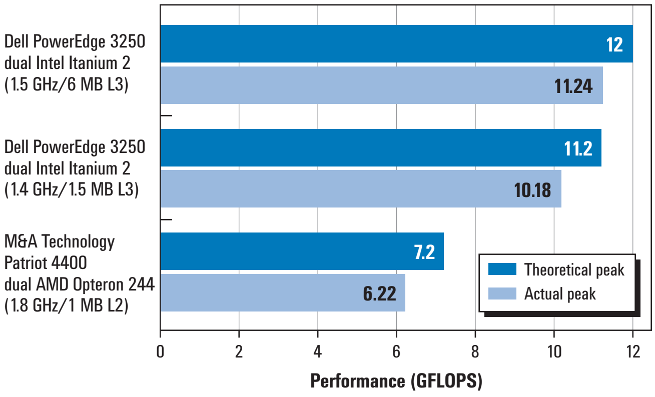


Figure 1. Single-server Linpack performance results

Linpack performance on an HPC cluster

The Linpack benchmark is also used to classify the Top500 Supercomputer Sites ranking of the 500 most powerful computer systems.[7] In November 2003, the Dell High-Performance Computing Cluster (HPCC) Lab tested a 16-node PowerEdge 3250 HPC cluster that comprised 32 Intel Itanium 2 processors at 1.5 GHz with 6 MB of L3 cache and 4 GB of DDR 266 memory. The cluster nodes ran Red Hat Enterprise Linux AS 2.1.

The Dell HPCC Lab achieved a performance result of 155.7 GFLOPS. This result exceeded the RISC-based result on a 32-processor 1.7 GHz IBM® eServer® pSeries® 690 server using POWER4™ technology that achieved a result of 143.3 GFLOPS that was published in a report dated February 2004.[8]

In addition to providing superior performance results, a PowerEdge 3250 HPC cluster can provide excellent price/performance for organizations seeking optimal floating-point performance. These impressive results position the PowerEdge 3250 as a leader in floating-point application performance, and further demonstrate the suitability of high-performance Intel architecture–based Dell servers for standards-based HPC cluster environments.

[4] For more information about the Linpack benchmark, visit http://www.top500.org/lists/linpack.php.

[5] The Opteron test was published by AMD in July 2003. For test results, visit http://www.amd.com/us-en/Processors/ProductInformation/0,,30_118_8796_8807~72905,00.html. Test results current as of June 14, 2004.

[6] For more information about the Intel Math Kernel Libraries, visit http://www.intel.com/software/products/mkl.

[7] For more information, visit http://www.top500.org.

[8] The IBM eServer pSeries test was published in a test report on February 2004. For test results, visit http://www-1.ibm.com/servers/eserver/pseries/hardware/system_perf.pdf. Test report current as of June 14, 2004.

Evaluating PowerEdge servers based on the SPEC CPU2000 benchmark

SPEC® CPU2000 is a CPU-intensive set of benchmarks from the Standard Performance Evaluation Corporation (SPEC).[9] Industry-standard SPEC CPU2000 benchmarks are designed to provide a comparative measure of compute-intensive performance across different hardware platforms. The benchmarks are developed from actual user applications, and they measure the performance of the processor, memory, and compiler on the tested system.

SPEC CPU2000 contains two benchmark suites, including CFP2000. The CFP2000 suite comprises 14 CPU-intensive benchmarks (six written in FORTRAN 77, four in FORTRAN 90, and four in C). The suite measures system performance when running compute-intensive floating-point applications.

SPEC CFP2000 provides performance measurements for system speed and throughput. Within the CFP2000 suite, the SPECfp®_peak2000 speed metric measures how fast a server completes running all of the floating-point benchmarks. The throughput metric, SPECfp_rate2000, measures how many tasks a computer can complete in a given amount of time. CFP2000 has been designed to measure throughput for both single-processor servers and symmetric multiprocessing (SMP) systems.

CFP2000 speed test results for the PowerEdge 3250

To test the floating-point capabilities of Dell PowerEdge 3250 servers, the Dell Enterprise Performance Team compared the CFP2000 test results published for the PowerEdge 3250 server in August 2003 to the fastest Microsoft Windows®–based Opteron processor–based server result (as of June 14, 2004).[10] The CFP2000 tests demonstrated superior floating-point performance for the Dell PowerEdge 3250 compared to the Opteron processor–based server when all servers were configured with one processor. The server configuration details were as follows:

- A PowerEdge 3250 was configured with a single Intel Itanium 2 processor at 1.5 GHz using 6 MB of L3 cache and 8 GB of DDR 200 memory. The server ran the Microsoft Windows Server 2003 Enterprise Edition OS. The Dell test team used Intel C++ and Fortran 7.1 compilers to compile the benchmark suite. This result was published in August 2003.
- An AMD server using an ASUS® SK8V motherboard was configured with a single model 150 AMD Opteron processor at 2.4 GHz using 1 MB L2 cache and 1 GB of DDR 400 memory. The server ran the Microsoft Windows XP Professional OS. AMD used Intel C++ and Fortran 8.0 compilers and the PGI Fortran 5.1 compiler to compile the benchmark suite.

Figure 2 indicates the results of SPECfp_peak2000—the speed test for the CFP2000 suite—for the Dell PowerEdge 3250 and AMD Opteron server configurations just described. Figure 2 indicates this important result:

- The PowerEdge 3250 configured with a single Itanium 2 processor at 1.5 GHz using 6 MB of L3 cache performed 22.7 percent better than the AMD Opteron processor–based server configured with a model 150 AMD Opteron processor at 2.4 GHz.

CFP2000 throughput test results for the PowerEdge 3250

Figure 3 indicates the results of SPECfp_rate2000—the throughput test for the CFP2000 suite—for the Dell PowerEdge 3250 and AMD Opteron processor–based servers. As in the speed tests, the Dell Enterprise Performance Team compared the PowerEdge 3250 test results published in August 2003 to the fastest Microsoft Windows–based Opteron dual processor–based server result (as of June 14, 2004).[11] The throughput test configuration for the PowerEdge 3250 was identical to that in the speed test (see "CPF2000 speed test results for the PowerEdge 3250") except that the PowerEdge 3250 used dual processors, not a single processor,

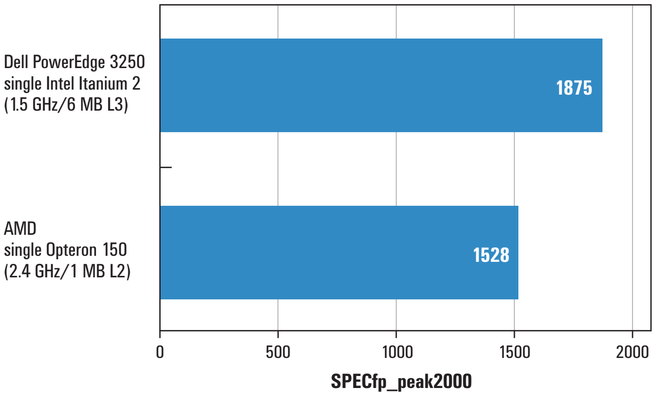


Figure 2. SPECfp_peak2000 performance using a single processor

[9] For more information about the SPEC CPU2000 benchmark, visit http://www.spec.org/cpu2000.

[10] The Opteron test results were published by AMD in May 2004. For test results, visit http://www.spec.org/osg/cpu2000/results. Competitive numbers reflect published results on http://www.spec.org as of June 14, 2004.

[11] The Opteron test results were published by AMD in May 2004. For test results, visit http://www.spec.org/osg/cpu2000/results. Competitive numbers reflect published results on http://www.spec.org as of June 14, 2004.

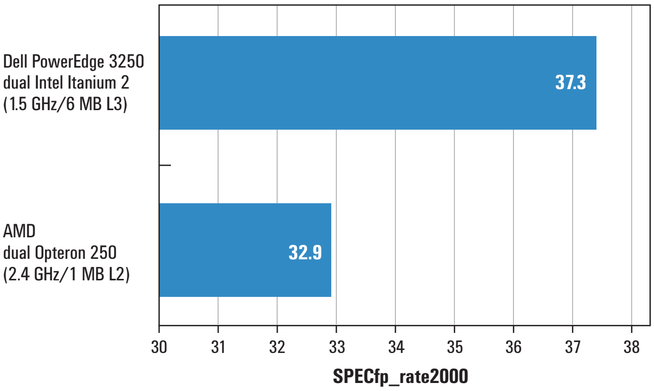


Figure 3. SPECfp_rate2000 throughput using dual processors

to measure multiprocessor system performance. The configuration for the AMD Opteron processor–based server was as follows:

- An AMD server using a Tyan® Thunder K8W motherboard was configured with dual model 250 AMD Opteron processors at 2.4 GHz using 1 MB L2 cache and 2 GB of DDR 400 memory. The server ran the Microsoft Windows Server 2003 Enterprise Edition OS. AMD used Intel C++ and Fortran 8.0 compilers and the PGI Fortran 5.1 compiler to compile the benchmark suite.

Figure 3 indicates this important result:

- The PowerEdge 3250 system configured with dual Itanium 2 processors at 1.5 GHz using 6 MB of L3 cache performed 13.3 percent better than the AMD Opteron processor–based system configured with dual model 250 AMD Opteron processors at 2.4 GHz using 1 MB of L2 cache and 2 GB of DDR 400 memory.

CFP2000 32-bit versus 64-bit performance comparison

The Dell team also saw significant improvements in floating-point performance when comparing the performance of the 64-bit PowerEdge 3250 configured with Itanium 2 processors to the performance of a Dell server configured with 32-bit Intel Xeon™ processors running the SPECfp2000 benchmark. Figure 4 shows the performance improvements of the PowerEdge 3250 system configured with a single Intel Itanium processor at 1.4 GHz using 1.5 MB of L3 cache and a PowerEdge 3250 system configured with dual Intel Itanium processors at 1.4 GHz using 1.5 MB of L3 cache over a PowerEdge 2650 system configured with a single Intel Xeon processor at 3.2 GHz using 2 MB of L3 cache and a PowerEdge 2650 system configured with dual Intel Xeon processors at 3.2 GHz using 2 MB of L3 cache.

As in the CFP2000 tests previously described, Dell engineers ran the SPECfp_peak2000 speed test using a single processor in both the PowerEdge 2650 and PowerEdge 3250 systems. They ran the SPECfp_rate2000 throughput test using dual processors in both systems. The PowerEdge 2650 tests were run in February 2004 for both single- and dual-processor configurations.

The performance gains of the Dell PowerEdge 3250 system configured with Itanium 2 processors at 1.4 GHz using 1.5 MB of L3 cache compared to the PowerEdge 2650 system as follows:

- The PowerEdge 3250 system configured with a single Intel Itanium 2 processor at 1.4 GHz using 1.5 MB of L3 cache performed 7.2 percent faster on the SPECfp_peak2000 speed test compared to the PowerEdge 2650 configured with a single Intel Xeon processor at 3.2 GHz using 2 MB of L3 cache.
- The PowerEdge 3250 system configured with dual Intel Itanium 2 processors at 1.4 GHz using 1.5 MB of L3 cache performed 14.4 percent faster on the SPECfp_rate2000 throughput test compared to the PowerEdge 2650 configured with dual Intel Xeon processors at 3.2 GHz using 2 MB of L3 cache.

The performance gains were more significant when the PowerEdge 3250 system was configured with Itanium 2 processors at 1.5 GHz using 6 MB of L3 cache:

- The PowerEdge 3250 system configured with a single Intel Itanium 2 processor at 1.5 GHz using 6 MB of L3 cache performed 39.1 percent faster on the SPECfp_peak2000 speed test compared to the PowerEdge 2650 configured with a single Intel Xeon processor at 3.2 GHz using 2 MB of L3 cache.
- The PowerEdge 3250 system configured with dual Intel Itanium 2 processors at 1.5 GHz using 6 MB of L3 cache performed 53.5 percent faster on the SPECfp_rate2000

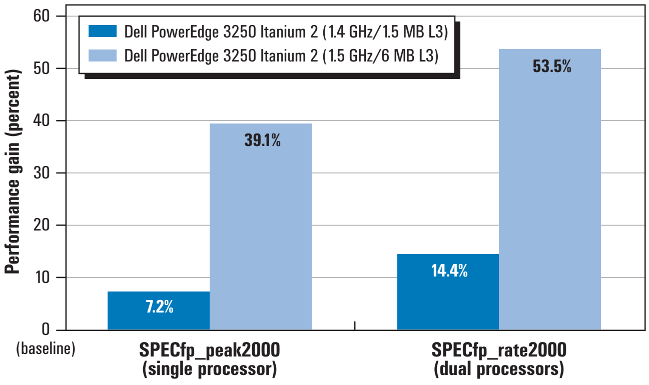


Figure 4. SPECfp2000 performance improvements for the 64-bit PowerEdge 3250 over the 32-bit PowerEdge 2650 (baseline)

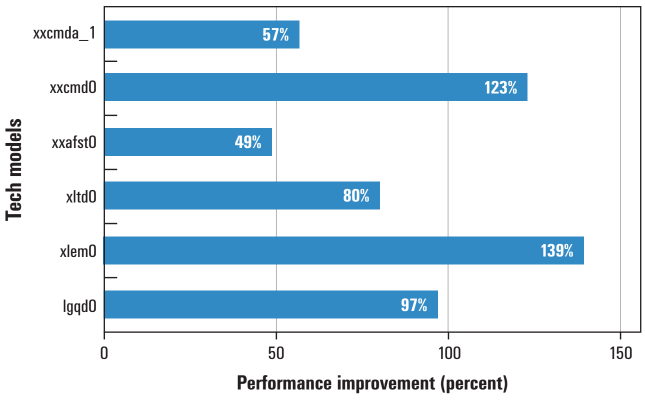


Figure 5. 32-bit versus 64-bit MSC.Nastran benchmark performance results

throughput test compared to the PowerEdge 2650 system configured with dual Intel Xeon processors at 3.2 GHz using 2 MB of L3 cache.

Examining MSC.Nastran benchmark performance

The MSC.Nastran™ benchmark is a widely used computer-aided engineering (CAE) program for simple to complex linear and nonlinear analyses of structural, fluid, thermal, and coupled systems. MSC.Nastran is a floating-point intensive program that exercises the processor, memory, and I/O subsystem of the server. The benchmark results illustrated in Figure 5 are based on three models from MSC.Nastran: cube, car body, and propeller housing.[12]

In November 2003, the Dell Enterprise Performance Team submitted results for the latest MSC.Nastran version 2004 benchmark on the PowerEdge 3250 and PowerEdge 2650 to MSC.Software Corporation. The configurations used were as follows:

- A PowerEdge 3250 was configured with dual Itanium 2 processors at 1.5 GHz using 6 MB of L3 cache and 16 GB of DDR 200 memory. The server ran the 64-bit Red Hat Enterprise Linux AS 2.1 OS. The configuration included external storage on a Dell PowerVault™ 200S enclosure using five 36 GB Ultra320 SCSI controllers and five 15,000 rpm software-striped hard drives (one for the OS and four for scratch directories).
- A PowerEdge 2650 was configured with dual Intel Xeon processors at 3.2 GHz using 1 MB of L3 cache and 12 GB of DDR 266 memory. The server ran the 32-bit Red Hat Enterprise Linux AS 2.1 OS. The configuration included external storage on a Dell PowerVault 200S enclosure using five 36 GB Ultra320 SCSI controllers and five software-striped hard drives (one for the OS and four for scratch directories).

Figure 5 compares the performance of the PowerEdge 3250 server to the 32-bit PowerEdge 2650 server. The PowerEdge 3250 server configured with dual Itanium 2 processors at 1.5 GHz using 6 MB of L3 cache outperformed the Intel Xeon processor–based server on all the models. The advanced floating-point capabilities of the Itanium 2 processor, along with its ability to handle large data sets, enabled the Itanium 2 to achieve impressive performance gains versus the PowerEdge 2650. These gains ranged from 49 percent to 139 percent across the different MSC.Nastran models.

Gaining performance advantages in HPC clusters using Itanium 2 processors

Both the architectural benefits and performance capabilities of the 64-bit Itanium 2 processor make transitioning to Itanium 2 processors a compelling option for organizations running HPC clusters. As discussed in this article, 64-bit systems offer distinct advantages over 32-bit systems for high-performance clusters. The performance of the PowerEdge 3250 on industry-standard benchmarks such as Linpack, SPEC CPU2000, and MSC.Nastran—in addition to considerations such as a dense form factor and excellent price/performance—suggest that the PowerEdge 3250 configured with Itanium 2 processors can be an optimal compute node for HPC cluster environments.

Ramesh Radhakrishnan, Ph.D. (ramesh_radhakrishnan@dell.com) is a design engineer consultant with the Dell System Performance and Analysis Lab. His responsibilities include performance analysis of Dell servers and characterization of enterprise-level benchmarks. Ramesh has a Ph.D. in Computer Engineering from The University of Texas at Austin.

David Ard (david_ard@dell.com) is a product marketing manager in the Enterprise Product Group at Dell Inc. His responsibilities include product marketing for Dell 2U rack server products. David has a B.A. in Technical Communication from Texas Tech University.

FOR MORE INFORMATION

Intel Itanium migration:
http://www.intel.com/products/server/processors/server/itanium2/itanium_wp.htm

Intel Itanium performance:
http://intel.com/products/server/docs/risc_epic.pdf

Intel Itanium architecture:
http://csdl.computer.org/comp/mags/mi/2003/02/m2toc.htm

[12] For more details on MSC.Nastran, visit http://www.mscsoftware.com/products/?Q=132 and http://www.mscsoftware.com/products/products_detail.cfm?PI=7.

Defending Networks with Intrusion Detection Systems

Securing an enterprise network requires significant technical skills as well as an ongoing effort to keep up with the ever-expanding universe of security exploits, threats, software, methodologies, and tools. This article explains how to increase the level of network security proactively by integrating a network intrusion detection system.

BY JOSE MARIA GONZALEZ

An early-warning system that alerts IT organizations to the presence of intruders can help prevent security breaches on the corporate network and help protect servers from being compromised. To help minimize unauthorized access attempts from both inside and outside the enterprise firewall, administrators can install an intrusion detection system (IDS)—which essentially acts as a network surveillance "camera." This article explores how administrators can thwart break-in attempts using an IDS, such as the Sourcefire® Snort™ product, and Dell™ PowerConnect™ Ethernet switches with port mirroring enabled.[1] In addition, the information in this article can help IT organizations raise corporate awareness of security threats and vulnerabilities.[2]

Understanding intrusion detection systems

While security vulnerabilities have been a topic of increasing concern over the last several years, no effective gauge exists to evaluate the risk that enterprise networks actually face. In spite of using the latest and most powerful security tools and encryption algorithms, companies are still being hacked and Web sites are still being put offline. Best practices require that administrators gather and analyze the data generated by an attacker's unsuccessful access attempts before a successful network breach occurs. This proactive approach to security threats can be far more cost-effective and useful than patching systems after a break-in.

Like a surveillance camera, an IDS enables security teams in an IT environment to capture every action an intruder makes. In this way, an IDS allows administrators to identify how and when a network is being scanned as well as how and when to observe an intruder without being noticed. By using a network adapter in surveillance—or *promiscuous*—mode, an IDS can monitor and analyze in real time every suspicious frame that travels into and out of a network. This setup allows administrators to monitor for attack signatures—specific patterns that usually indicate an unauthorized attempt to gain access to systems. Administrators can download the latest attack signatures from the Internet in the same way that they download virus definition files to keep anti-virus software up-to-date with new security protections.[3]

[1] For more information about Snort, visit http://www.snort.org. Snort was rated the best IDS open source product of 2003 by *Information Security Magazine* (see "The Best: Celebrating Security's Best People, Policy, Process, Products" in *Information Security Magazine*, December 2003, http://infosecuritymag.techtarget.com/ss/0,295796,sid6_iss288_art517,00.html).

[2] For more information about network security, visit the Honeynet Project at http://project.honeynet.org. Honeynet is a nonprofit research organization of security professionals dedicated to learning the tools, tactics, and motives of the hacker, or *black hat*, community and sharing lessons learned with administrators and other IT workers responsible for network security.

[3] To download the latest attack signatures, visit http://www.snort.org/snort-db.

Two main types of IDS implementation exist:

- **Host-based intrusion detection system (HIDS):** A host-based system monitors activity on a specific server. An HIDS is a software-based solution that continuously scans the system log on a single host.
- **Network-based intrusion detection system (NIDS):** A network-based system monitors and analyzes traffic on the network, instead of on a single host, and provides a reporting mechanism that enables real-time detection and response. For instance, this capability can help an NIDS to stop a distributed denial of service attack (DDoS) before the targeted host crashes.

This article focuses on network-based intrusion detection systems. One of the key decisions for administrators when deploying an NIDS is where to place it. For some organizations a pair of NIDSs is best—one outside the enterprise firewall and one behind it. This type of implementation can report which threats are filtered out by the firewall and which attacks pass through the firewall into the organization's network. By examining the alert logs on both NIDSs, administrators can determine which traffic has been filtered and which has not. *Note:* NIDS system placement can vary widely, and individual networks may use different implementations.[4]

Setting up the NIDS and configuring rules

Figure 1 shows one NIDS server outside the firewall with port mirroring enabled on a Dell PowerConnect Ethernet switch positioned between the firewall and the router. The switch mirrors all frames traveling inbound and outbound between the router and firewall to the port where the NIDS server is connected to the switch.

Setting up the Snort NIDS is easy. Simply download the latest tar file from http://www.snort.org and untar it. The tar file comes with an installation shell script to facilitate installation. A Microsoft® Windows® version of Snort is also available.[5]

Administrators should avoid running any services on the NIDS server except the NIDS software so that invading viruses, worms, and other security threats will find little or nothing to exploit. The NIDS server should be configured only with a hardened operating system and the NIDS software. In addition, the latest security patches to the operating system should be installed. As another precaution, the server should be configured without an IP setup or TCP/IP stack to ensure that the NIDS will be undetectable to would-be intruders. *Note:* If this server were to be compromised, the NIDS would also be compromised and hackers could make system administrators see only what they wanted the administrators to see.

Snort has three main modes: sniffer, packet logger, and network intrusion detection. The third is the most complex, configurable mode and is the main focus of this article. Network intrusion detection mode analyzes network traffic for matches against a defined rule set stored in a configuration file (snort.conf) and triggers actions based on what the NIDS has captured.

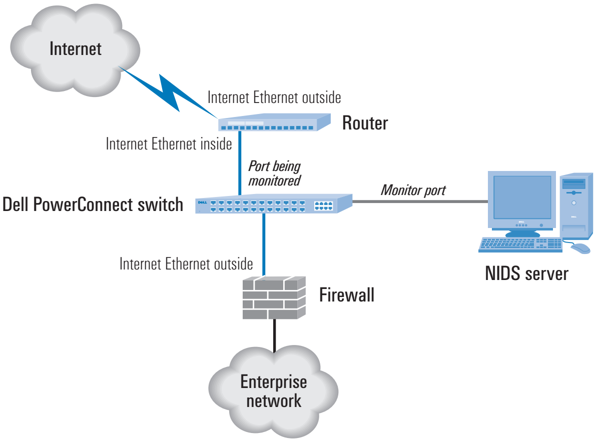


Figure 1. A typical IDS setup

Once Snort is installed, administrators can enable network intrusion detection mode simply by typing the following command line:

```
./snort -d -b -A full -i eth0 172.168.1.0/24
   -l /var/log/snort -c snort.conf
```

Snort.conf is the name of the rules file and the IP address is the network's IP range. Fictitious network address 172.168.1.0/24 is used to represent the Internet IP address pool discussed throughout this article.

Putting an unprotected server on the network

To assess the potential level of threat to unprotected systems, in February 2004 a Dell engineer at the Dell Application Solution Centre in Limerick, Ireland, placed an isolated, off-the-shelf server—running with no security patches and some security weaknesses—outside an enterprise firewall (see Figure 2). Of course, in a real-world environment, administrators should set up enterprise servers behind a firewall to protect them from the public network and contain and control outbound connections if the servers become compromised.

In this exercise, a firewall was not used to protect the target server. However, in the final analysis the target server would have been compromised even if it had been placed behind a firewall (see "Analyzing the captured data"). Certain security exploits—for example, the so-called port 80 problem—trespass firewalls because they are carried out through ports that can get through filters.

[4] For more information about NIDS placement, visit http://www.snort.org/docs/#deploy.

[5] For more information, visit http://www.snort.org/dl/binaries/win32.

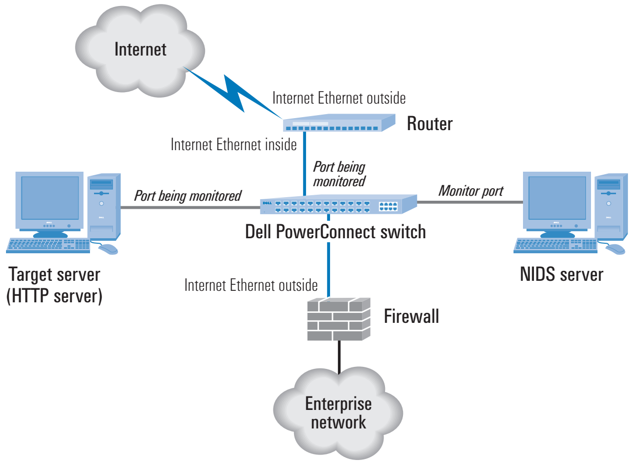


Figure 2. Test environment with target system installed

Analyzing the captured data

The target server was online for five days. In the first 15 minutes, it was randomly scanned 20 times from different locations on the Internet. Over the course of five days, the server received 400 random scans from would-be attackers scanning servers for security vulnerabilities. For a representative sample of the attack data, visit *Dell Power Solutions* online at http://www.dell.com/magazines_extras.

Analyzing captured data requires time, knowledge, and various tools. For this analysis, the following tools were used: the Snort NIDS; Ethereal network protocol analyzer; Argus open source network utility; Tripwire® HIDS file system security tool (installed on the target server); and The Coroner's Toolkit (TCT), a forensic analysis tool. They are all available on the Internet at no cost.[6] Each tool provides different pieces of information and when used together, these utilities can help administrators resolve how, when, and by whom a system was compromised.

Fortunately, the target server was not compromised during the test because hackers did not try to exploit the particular services running on the server. However, the analysis tools found evidence of several attempts to gain unauthorized access to the system using exploits such as directory traversal, Nimda, CodeRed II, MS-SQL, and a Simple Network Management Protocol (SNMP) worm—nothing serious because the system was not running any of the services that these attempts tried to exploit. Further analysis using forensic techniques supported by Tripwire and TCT showed no added, removed, or modified binary files on the target system.

Had a successful attack occurred, hackers could have installed a rootkit utility—which deletes any evidence of unauthorized attempts from the log files—or one or more back doors, which are modified binaries of trust utilities[7] that enable hackers to repeatedly log on to a system without arousing suspicion. Fortunately, the use of an NIDS and forensic analysis tools can help administrators detect such hacker ploys—and therefore help prevent a hacker who gains access to one system from applying the same techniques to compromise other systems and networks.

Taking a layered approach to network security

Network security is one of the most compelling concerns facing IT organizations today. As the security exercise in this article demonstrated, unknown attackers scan networks randomly to exploit security vulnerabilities on a daily basis. Unfortunately, there is no silver bullet or single product that can protect enterprise data against malicious intent. Even a well-configured firewall cannot provide adequate protection because many threats exploit weaknesses such as ports that are generally open at the firewall.

Diverse hardware and software tools—including Snort, Tripwire, and port mirroring on Dell PowerConnect switches—are available to help administrators detect security breaches and protect enterprise networks. Using layers of protection, such as those listed here, can improve the chances of keeping networks and systems secure. Some of the techniques that network security staff should consider include:

- Scanning to assess network vulnerabilities
- Conducting penetration testing and log audits
- Setting up security policies within the enterprise
- Installing the latest security patches as soon as they are available
- Educating IT staff about social engineering attacks
- Reviewing newsletters, mailing lists, and other security information on a daily basis to learn of new threats
- Using the scanning and detecting capabilities of an NIDS to find out what is traveling into and out of the network

Today, the level of threat to mission-critical enterprise data and assets is high. Effective network security requires an ongoing effort to defend against attack on a 24/7 basis, and an IDS can provide valuable assistance in helping to achieve network security goals.

Acknowledgments

The author would like to thank Kevin Libert, director of EMEA Enterprise Systems Marketing, and Anthony Quigney, Dell Application Solution Centre manager, for valuable feedback.

Jose Maria Gonzalez (jose_maria_gonzalez@dell.com) is a system consultant on Microsoft technologies and security at the Dell Application Solution Centre in Limerick, Ireland. He has a B.S. in Computer Science from the University of Madrid, and is a Microsoft Certified Systems Engineer (MCSE) and Red Hat Certified Engineer® (RHCE®).

[6] For more information, visit http://www.ethereal.com, http://www.qosient.com/argus, http://www.tripwire.org, and http://www.porcupine.org/forensics/tct.html.

[7] On UNIX® or Linux® servers, trust utilities include bin, ls, top, and ps.

ONLINE-EXTRA CONTENT

Building Distributed Microsoft SQL Server 2000
Database Applications on Dell PowerEdge 6650 Servers

By Dave Jaffe, Ph.D.; Todd Muirhead; and Will Claxton

Database build script for online DVD store

```
-- Copyright Dell Inc. 2004

-- Database
IF EXISTS (SELECT * FROM SYSDATABASES WHERE NAME='DS')
DROP DATABASE DS
GO

CREATE DATABASE DS ON
  PRIMARY
    (
    NAME = 'primary',
    FILENAME = 'c:\sql\dbfiles\ds.mdf'
    ),
  FILEGROUP DS_MISC_FG
    (
    NAME = 'ds_misc',
    FILENAME = 'e:\ds_misc.ndf',
    SIZE = 1GB
    ),
  FILEGROUP DS_CUST_FG
    (
    NAME = 'cust1',
    FILENAME = 'e:\cust1.ndf',
    SIZE = 25GB
    ),
    (
    NAME = 'cust2',
    FILENAME = 'e:\cust2.ndf',
    SIZE = 25GB
    ),
  FILEGROUP DS_ORDERS_FG
    (
    NAME = 'orders1',
    FILENAME = 'e:\orders1.ndf',
    SIZE = 20GB
    ),
    (
    NAME = 'orders2',
    FILENAME = 'e:\orders2.ndf',
    SIZE = 20GB
    ),
  FILEGROUP DS_IND_FG
    (
    NAME = 'ind1',
    FILENAME = 'e:\ind1.ndf',
    SIZE = 20GB
    ),
    (
    NAME = 'ind2',
    FILENAME = 'e:\ind2.ndf',
    SIZE = 20GB
    )
  LOG ON
    (
    NAME = 'ds_log',
    FILENAME = 'e:\ds_log.ldf',
    SIZE = 20GB
    )
GO

USE DS
GO

-- Tables

CREATE TABLE CUSTOMERS
  (
  CUSTOMERID INT IDENTITY NOT NULL,
  FIRSTNAME VARCHAR(50) NOT NULL,
  LASTNAME VARCHAR(50) NOT NULL,
  ADDRESS1 VARCHAR(50) NOT NULL,
  ADDRESS2 VARCHAR(50),
  CITY VARCHAR(50) NOT NULL,
  STATE VARCHAR(50),
  ZIP INT,
  COUNTRY VARCHAR(50) NOT NULL,
  REGION TINYINT NOT NULL,
  EMAIL VARCHAR(50),
  PHONE VARCHAR(50),
  CREDITCARD VARCHAR(50) NOT NULL,
  CREDITCARDEXPIRATION VARCHAR(50) NOT NULL,
  USERNAME VARCHAR(50) NOT NULL,
  PASSWORD VARCHAR(50) NOT NULL,
  AGE TINYINT,
  INCOME INT,
  GENDER VARCHAR(1)
  )
  ON DS_CUST_FG
GO

CREATE TABLE ORDERS
  (
  ORDERID INT IDENTITY NOT NULL,
  ORDERDATE DATETIME NOT NULL,
  CUSTOMERID INT NOT NULL,
  NETAMOUNT MONEY NOT NULL,
  TAX MONEY NOT NULL,
  TOTALAMOUNT MONEY NOT NULL
  )
  ON DS_ORDERS_FG
GO
```

```
CREATE TABLE ORDERLINES
  (
  ORDERLINEID SMALLINT NOT NULL,
  ORDERID INT NOT NULL,
  PROD_ID INT NOT NULL,
  QUANTITY SMALLINT NOT NULL,
  ORDERDATE DATETIME NOT NULL
  )
  ON DS_ORDERS_FG
GO

CREATE TABLE PRODUCTS
  (
  PROD_ID INT IDENTITY NOT NULL,
  CATEGORY TINYINT NOT NULL,
  TITLE VARCHAR(50) NOT NULL,
  ACTOR VARCHAR(50) NOT NULL,
  PRICE MONEY NOT NULL,
  QUAN_IN_STOCK INT NOT NULL,
  SPECIAL TINYINT
  )
  ON DS_MISC_FG
GO

CREATE TABLE CATEGORIES
  (
  CATEGORY TINYINT IDENTITY NOT NULL,
  CATEGORYNAME VARCHAR(50) NOT NULL,
  )
  ON DS_MISC_FG
GO

  SET IDENTITY_INSERT CATEGORIES ON
  INSERT INTO CATEGORIES (CATEGORY, CATEGORYNAME)
    VALUES (1,'Action')
  INSERT INTO CATEGORIES (CATEGORY, CATEGORYNAME)
    VALUES (2,'Animation')
  INSERT INTO CATEGORIES (CATEGORY, CATEGORYNAME)
    VALUES (3,'Children')
  INSERT INTO CATEGORIES (CATEGORY, CATEGORYNAME)
    VALUES (4,'Classics')
  INSERT INTO CATEGORIES (CATEGORY, CATEGORYNAME)
    VALUES (5,'Comedy')
  INSERT INTO CATEGORIES (CATEGORY, CATEGORYNAME)
    VALUES (6,'Documentary')
  INSERT INTO CATEGORIES (CATEGORY, CATEGORYNAME)
    VALUES (7,'Drama')
  INSERT INTO CATEGORIES (CATEGORY, CATEGORYNAME)
    VALUES (8,'Family')
  INSERT INTO CATEGORIES (CATEGORY, CATEGORYNAME)
    VALUES (9,'Foreign')
  INSERT INTO CATEGORIES (CATEGORY, CATEGORYNAME)
    VALUES (10,'Games')
  INSERT INTO CATEGORIES (CATEGORY, CATEGORYNAME)
    VALUES (11,'Horror')
  INSERT INTO CATEGORIES (CATEGORY, CATEGORYNAME)
    VALUES (12,'Music')
  INSERT INTO CATEGORIES (CATEGORY, CATEGORYNAME)
    VALUES (13,'New')
  INSERT INTO CATEGORIES (CATEGORY, CATEGORYNAME)
    VALUES (14,'Sci-Fi')
  INSERT INTO CATEGORIES (CATEGORY, CATEGORYNAME)
    VALUES (15,'Sports')
  INSERT INTO CATEGORIES (CATEGORY, CATEGORYNAME)
    VALUES (16,'Travel')
  GO
```

Stored procedures for online DVD store

```
-- copyright Dell Inc. 2004

-- NEW_CUSTOMER

USE DS
IF EXISTS (SELECT name FROM sysobjects WHERE name =
     'NEW_CUSTOMER' AND type = 'P')
  DROP PROCEDURE NEW_CUSTOMER
GO

USE DS
GO

CREATE PROCEDURE NEW_CUSTOMER
  (
  @firstname_in              VARCHAR(50),
  @lastname_in               VARCHAR(50),
  @address1_in               VARCHAR(50),
  @address2_in               VARCHAR(50),
  @city_in                   VARCHAR(50),
  @state_in                  VARCHAR(50),
  @zip_in                    INT,
  @country_in                VARCHAR(50),
  @region_in                 TINYINT,
  @email_in                  VARCHAR(50),
  @phone_in                  VARCHAR(50),
  @creditcard_in             VARCHAR(50),
  @creditcardexpiration_in   VARCHAR(50),
  @username_in               VARCHAR(50),
  @password_in               VARCHAR(50),
  @age_in                    TINYINT,
  @income_in                 INT,
  @gender_in                 VARCHAR(1)
  )
  AS
  IF (SELECT COUNT(*) FROM CUSTOMERS WHERE
     USERNAME=@username_in) = 0
  BEGIN
    INSERT INTO CUSTOMERS
      (
      FIRSTNAME,
      LASTNAME,
      ADDRESS1,
      ADDRESS2,
      CITY,
      STATE,
      ZIP,
      COUNTRY,
      REGION,
      EMAIL,
      PHONE,
      CREDITCARD,
      CREDITCARDEXPIRATION,
      USERNAME,
      PASSWORD,
      AGE,
      INCOME,
      GENDER
      )
    VALUES
      (
      @firstname_in,
      @lastname_in,
      @address1_in,
      @address2_in,
      @city_in,
      @state_in,
      @zip_in,
      @country_in,
      @region_in,
      @email_in,
      @phone_in,
      @creditcard_in,
      @creditcardexpiration_in,
      @username_in,
      @password_in,
      @age_in,
      @income_in,
      @gender_in
      )
    SELECT @@IDENTITY
  END
  ELSE
    SELECT 0
GO
```

```
-- LOGIN

USE DS
IF EXISTS (SELECT name FROM sysobjects WHERE
     name = 'LOGIN' AND type = 'P')
  DROP PROCEDURE LOGIN
GO

USE DS
GO

CREATE PROCEDURE LOGIN
  (
  @username_in               VARCHAR(50),
  @password_in               VARCHAR(50)
  )

  AS
  DECLARE @customerid_out           INT

  SELECT @customerid_out=CUSTOMERID FROM
     CUSTOMERS WHERE USERNAME=@username_in AND
     PASSWORD=@password_in

  IF (@@ROWCOUNT > 0)
    SELECT @customerid_out
  ELSE
    SELECT 0
GO

USE DS
IF EXISTS (SELECT name FROM sysobjects WHERE
     name = 'BROWSE_BY_CATEGORY' AND type = 'P')
  DROP PROCEDURE BROWSE_BY_CATEGORY
GO

USE DS
GO

CREATE PROCEDURE BROWSE_BY_CATEGORY
  (
  @batch_size_in             INT,
  @category_in               INT
  )

  AS
  SET ROWCOUNT @batch_size_in
  SELECT * FROM PRODUCTS WHERE CATEGORY=@category_in
  SET ROWCOUNT 0
GO

USE DS
IF EXISTS (SELECT name FROM sysobjects WHERE
     name = 'BROWSE_BY_ACTOR' AND type = 'P')
  DROP PROCEDURE BROWSE_BY_ACTOR
GO

USE DS
GO

CREATE PROCEDURE BROWSE_BY_ACTOR
  (
  @batch_size_in             INT,
  @actor_in                  VARCHAR(50)
  )

  AS

  SET ROWCOUNT @batch_size_in
  SELECT * FROM PRODUCTS WHERE ACTOR=@actor_in
  SET ROWCOUNT 0
GO

USE DS
IF EXISTS (SELECT name FROM sysobjects WHERE
    name = 'BROWSE_BY_TITLE' AND type = 'P')
  DROP PROCEDURE BROWSE_BY_TITLE
GO

USE DS
GO

CREATE PROCEDURE BROWSE_BY_TITLE
  (
  @batch_size_in             INT,
  @title_in                  VARCHAR(50)
  )

  AS

  SET ROWCOUNT @batch_size_in
  SELECT * FROM PRODUCTS WHERE TITLE=@title_in
  SET ROWCOUNT 0
GO

USE DS
IF EXISTS (SELECT name FROM sysobjects WHERE
    name = 'PURCHASE' AND type = 'P')
  DROP PROCEDURE PURCHASE
GO

USE DS
GO

CREATE PROCEDURE PURCHASE
  (
  @customerid_in     INT,
  @number_items      INT,
```

```
  @netamount_in      MONEY,
  @taxamount_in      MONEY,
  @totalamount_in    MONEY,
  @prod_id_in0       INT = 0,   @qty_in0   INT = 0,
  @prod_id_in1       INT = 0,   @qty_in1   INT = 0,
  @prod_id_in2       INT = 0,   @qty_in2   INT = 0,
  @prod_id_in3       INT = 0,   @qty_in3   INT = 0,
  @prod_id_in4       INT = 0,   @qty_in4   INT = 0,
  @prod_id_in5       INT = 0,   @qty_in5   INT = 0,
  @prod_id_in6       INT = 0,   @qty_in6   INT = 0,
  @prod_id_in7       INT = 0,   @qty_in7   INT = 0,
  @prod_id_in8       INT = 0,   @qty_in8   INT = 0,
  @prod_id_in9       INT = 0,   @qty_in9   INT = 0
  )

  AS

  DECLARE
  @date_in           DATETIME,
  @neworderid        INT,
  @item_id           INT,
  @prod_id           INT,
  @qty               INT

  SET DATEFORMAT ymd

  SET @date_in = GETDATE()
--SET @date_in = '2003/10/31'


  -- CREATE NEW ENTRY IN ORDERS TABLE
  INSERT INTO ORDERS
    (
    ORDERDATE,
    CUSTOMERID,
    NETAMOUNT,
    TAX,
    TOTALAMOUNT
    )
  VALUES
    (
    @date_in,
    @customerid_in,
    @netamount_in,
    @taxamount_in,
    @totalamount_in
    )

  SET @neworderid = @@IDENTITY


  -- ADD LINE ITEMS TO ORDERLINES

  SET @item_id = 0
  WHILE (@item_id < @number_items)
  BEGIN
    SELECT @prod_id = CASE @item_id
       WHEN 0 THEN @prod_id_in0
       WHEN 1 THEN @prod_id_in1
       WHEN 2 THEN @prod_id_in2
       WHEN 3 THEN @prod_id_in3
       WHEN 4 THEN @prod_id_in4
       WHEN 5 THEN @prod_id_in5
       WHEN 6 THEN @prod_id_in6
       WHEN 7 THEN @prod_id_in7
       WHEN 8 THEN @prod_id_in8
       WHEN 9 THEN @prod_id_in9
    END

    SELECT @qty = CASE @item_id
       WHEN 0 THEN @qty_in0
       WHEN 1 THEN @qty_in1
       WHEN 2 THEN @qty_in2
       WHEN 3 THEN @qty_in3
       WHEN 4 THEN @qty_in4
       WHEN 5 THEN @qty_in5
       WHEN 6 THEN @qty_in6
       WHEN 7 THEN @qty_in7
       WHEN 8 THEN @qty_in8
       WHEN 9 THEN @qty_in9
    END

    INSERT INTO ORDERLINES
      (
      ORDERLINEID,
      ORDERID,
      PROD_ID,
      QUANTITY,
      ORDERDATE
      )
    VALUES
      (
      @item_id,
      @neworderid,
      @prod_id,
      @qty,
      @date_in
      )
    SET @item_id = @item_id + 1
  END

  SELECT @neworderid
GO

USE DS
IF EXISTS (SELECT name FROM sysobjects WHERE
    name = 'ROLLUP_BY_CATEGORY_FULL' AND type = 'P')
  DROP PROCEDURE ROLLUP_BY_CATEGORY_FULL
GO

USE DS
GO

CREATE PROCEDURE ROLLUP_BY_CATEGORY_FULL
  (
  @category_in               INT
  )

  AS

  SET DATEFORMAT ymd

  SELECT SUM(OL.QUANTITY * P.PRICE) FROM
    ORDERLINES OL, PRODUCTS P WHERE
    OL.PROD_ID = P.PROD_ID AND
    P.CATEGORY = @category_in AND
    '2003-06-01' <= OL.ORDERDATE AND
    '2003-07-01' > OL.ORDERDATE;

  SELECT SUM(OL.QUANTITY * P.PRICE) FROM
    ORDERLINES OL, PRODUCTS P WHERE
    OL.PROD_ID = P.PROD_ID AND
    P.CATEGORY = @category_in AND
    '2003-04-01' <= OL.ORDERDATE AND
    '2003-07-01' > OL.ORDERDATE;

  SELECT SUM(OL.QUANTITY * P.PRICE) FROM
    ORDERLINES OL, PRODUCTS P WHERE
    OL.PROD_ID = P.PROD_ID AND
    P.CATEGORY = @category_in AND
    '2003-07-01' > OL.ORDERDATE;
GO
```

Using WMI Scripting for System Administration

By Sudhir Shetty

Scripting remote power control

A system administrator can remotely reboot a system using the script shown in Figure A. To run this script, save it in a file called reboot.vbs and run the following command, passing the remote machine name on the command line:

```
C:> cscript reboot.vbs remoteMachineName
```

```
strComputer = WScript.Arguments.Item(0)
Set objWMIService = GetObject("winmgmts:" _
  & "{impersonationLevel=impersonate,(Shutdown)}!\\" & _
  strComputer & "\root\cimv2")
Set colOS = objWMIService.ExecQuery _
  ("SELECT * FROM Win32_OperatingSystem")

'* Reboot the remote machine
For each objOS in colOS
  objOS.Reboot()
Next
```

Figure A. Sample script for remote power control

Scripting service control

Services are applications that can communicate and be administered by the Service Control Manager (SCM). Administrators can use scripts to access and modify service information. Figure B provides an example of how administrators can access service information and start and stop services.

Windows® Management Instrumentation (WMI) scripts allow administrators to perform a variety of functions using this basic model, such as starting, stopping, pausing, and resuming services; enumerating dependent and antecedent services; configuring service start options; installing and removing services; and managing service accounts and service account passwords. WMI scripts can be useful to system administrators in several scenarios:

- An administrator needs to alter the service account information on several computers running in the enterprise. A WMI script can enable the administrator to change the service account information for these services.
- An administrator can write a WMI script to monitor changes in service status using a temporary event subscriber. The script enables the administrator to automatically set up a notification process when a change occurs. The WMI notification scheme can also be leveraged to notify the administrator of other events, such as when free disk space on a server falls below a specific value or a CPU is overloaded.
- An administrator needs to monitor the event log for particular events. A sample script is presented in Figure C, which monitors the event log on the local computer for Dell™ OpenManage™ Server Administrator events.

```
strComputer = "."
Set objWMIService = GetObject("winmgmts:" _
  & "{impersonationLevel=impersonate}!\\" & strComputer & _
  "\root\cimv2")
  Set colServices = objWMIService.ExecQuery _
  ("SELECT * FROM Win32_Service")


'* List all the services on the machine
For each objService in colServices
  WScript.Echo objService.DisplayName & "," & _
  objService.StartName & "," & objService.State
Next


'* Stop and start the Task Scheduler Service
Set colServices = objWMIService.ExecQuery _
  ("SELECT * FROM Win32_Service WHERE Name = 'Schedule'")
For each objService in colServices
  WScript.Echo "Stopping the service:" & objService.Name & _
  "," & objService.DisplayName
  objService.StopService()
  WScript.Echo "Starting the service:" & objService.Name & _
  "," & objService.DisplayName
  objService.StartService()
Next
```

Figure B. Sample script for controlling services

```
strComputer = "."
Set objWMIService = GetObject("winmgmts:" _
  & "{impersonationLevel=impersonate}!\\" & strComputer & _
  "\root\cimv2")

'* Monitor for Server Administrator events
strWQL = "SELECT * FROM __InstanceCreationEvent WITHIN 15 WHERE TargetInstance ISA 'Win32_NTLogEvent' " & _
  " AND TargetInstance.SourceName = 'Server Administrator' AND TargetInstance.Logfile = 'System'"
Set objEventSource = objWMIService.ExecNotificationQuery(strWQL)

'* Continuously poll for those events
While True
  Set objEvent = objEventSource.NextEvent()
  Set NTLogEvent = objEvent.TargetInstance
  Wscript.Echo "Category: " & NTLogEvent.Category
  Wscript.Echo "Computer Name: " & NTLogEvent.ComputerName
  Wscript.Echo "Event Code: " & NTLogEvent.EventCode
  Wscript.Echo "Message: " & NTLogEvent.Message
  Wscript.Echo "Record Number: " & NTLogEvent.RecordNumber
  Wscript.Echo "Source Name: " & NTLogEvent.SourceName
  Wscript.Echo "Time Written: " & NTLogEvent.TimeWritten
  Wscript.Echo "Event Type: " & NTLogEvent.Type
  Wscript.Echo "User: " & NTLogEvent.User
Wend
```

Figure C. Sample script for monitoring the event log

Executing a remote batch file

Administrators can run a sequence of commands on a remote managed node in several ways, such as:

- Creating a batch file on the middle tier, which contains a sequence of remote execution commands that are executed on the middle-tier system.
- Copying a batch file to the remote node and executing it on the remote node. To run a batch file called esmlog.bat on remote node machineA, administrators would use the following command:

  ```
  C:> cscript remotebatch.vbs
     machineA esmlog.bat
  ```

  where esmlog.bat is stored in the current working directory. The script for remotebatch.vbs is listed in Figure D.

```
strComputer = WScript.Arguments.Item(0)
strBatchFile = WScript.Arguments.Item(1)

'* First copy the batch file to the remote system
Set objFSO=CreateObject("Scripting.FileSystemObject")
ObjFSO.CopyFile strBatchFile, "\\" & strComputer & "\C$\tmp\" & _
  strBatchFile

'* Launch the process on the remote system
Set objWMIService = GetObject("winmgmts:" _
  & "{impersonationLevel=impersonate}!\\" & strComputer & _
  "\root\cimv2:Win32_Process")

strCommand = "C:\tmp\" & strBatchFile
errReturn = objWMIService.Create(strCommand,null,null,intProcessID)
if errReturn = 0 Then
  Wscript.Echo strCommand & " was started with a process ID of " _
  & intProcessID & "."
Else
  Wscript.Echo strCommand & " could not be started due to error " & _
  errReturn & "."
End If
```

Figure D. Sample WMI script for executing a remote batch file

ONLINE-EXTRA CONTENT

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

Key functions in Dell OpenManage Server Assistant 8.x

Dell™ OpenManage™ Server Assistant 8.x helps support the installation of the following operating systems[1] on Dell PowerEdge™ servers:

- Microsoft® Windows® 2000 Server and Advanced Server, and Small Business Server 2000
- Microsoft Windows Server™ 2003 (Enterprise, Standard, and Web editions) and Small Business Server 2003
- Red Hat® Enterprise Linux® AS 2.1
- Red Hat Enterprise Linux 3
- Novell® NetWare® 5.1
- Novell NetWare 6.5

Server Assistant also supports the following localized languages for operating system (OS) installation: English, French, German, Japanese, Simplified Chinese, and Spanish. When booting the system with Server Assistant, administrators can select the desired language and the preferred keyboard type in which the program will operate.

Configuration of RAID for storage devices

Dell PowerEdge Expandable RAID Controllers (PERCs) are designed to be high-performance, intelligent Peripheral Component Interconnect (PCI®)–to–SCSI host adapters with RAID-control capabilities. The RAID approach is designed to improve I/O performance or data integrity compared to a computer using a single disk drive. A RAID configuration can enable data to be read from multiple hard drives at the same time. RAID also can help improve data storage, reliability, and fault tolerance because data lost when one disk drive fails can be reconstructed from the data that exists on other disk drives.

However, configuration can be the most significant challenge administrators encounter with RAID. Traditionally, administrators have employed hard-to-use BIOS or easy-to-lose floppy diskettes to configure RAID. Server Assistant provides a graphical user interface (GUI) that allows administrators to quickly and easily set up the most common RAID configurations. Administrators simply choose the RAID level—from full redundancy to maximum performance—and Server Assistant is designed to configure the system accordingly. Figure A shows the express mode RAID configuration page. Dell OpenManage Server Assistant 8.x supports the following RAID levels:

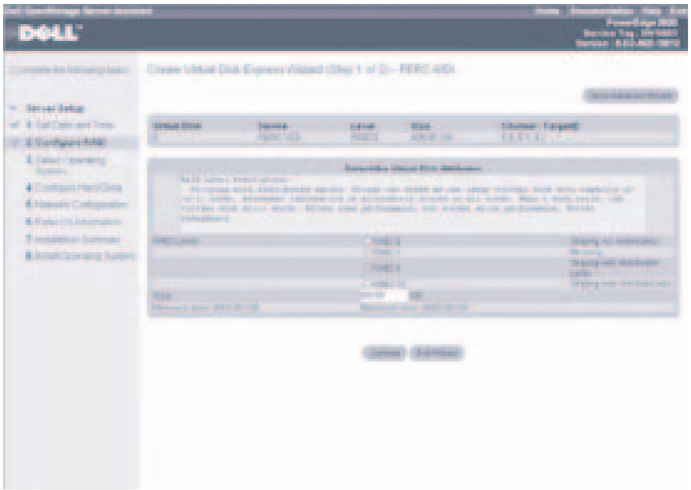


Figure A. Dell OpenManage Server Assistant 8.x express mode RAID configuration page

- **RAID-0 (striping):** Groups selected disks together as one large virtual disk having a total capacity of all the disks. Data is divided into blocks and distributed sequentially among the selected disks. RAID-0 improves read and write performance but does not provide fault tolerance.
- **RAID-1 (mirroring):** Groups two disks together as one virtual disk having the capacity of a single disk; data written to one disk is duplicated on the other disk. RAID-1 provides fault tolerance, and when compared to a single disk, it offers higher read performance but slightly slower write performance.
- **RAID-5 (striping with distributed parity):** Groups three or more disks together as one large virtual disk with a total capacity of *n*–1 disks (where *n* is the total number of disks). For example, in a three-drive configuration, data is distributed among the selected disks, and parity bits from two disks are distributed on a third disk to help provide fault tolerance.
- **RAID-10 (striping over mirrored arrays):** Offers a combination of RAID-1 and RAID-0. Data is striped across the replicated and mirrored-pair disks. RAID-10 helps provide fault tolerance through RAID-1 and enhanced I/O performance through RAID-0.

System device detection and unattended driver installation

To help reduce complexity and save time when setting up a system, Dell OpenManage Server Assistant is designed to report the correct system model for the PowerEdge server, and also detect and report the hardware devices installed in the server; the detection result is reported in the View Hardware page. After OS selection and configuration, Server Assistant is designed to automatically install the most up-to-date, compatible drivers for RAID controllers, network adapters, video adapters, Dell remote access controllers (RACs), and other devices, such as modems, shipped with PowerEdge servers. This feature is designed to

[1] The list of supported operating systems varies with each Dell PowerEdge server model. Please check the product user's guide of the specific PowerEdge server or http://www.dell.com. Server Assistant is designed to automatically present the supported OS.

prevent administrators from installing an incompatible driver, which may degrade OS performance or cause system downtime.

Configuration of common OS features

Dell OpenManage Server Assistant is designed to help administrators quickly configure common server settings. It prompts administrators to provide pertinent information at the beginning of the setup process, to help prevent the installation phase from being interrupted.

Hard drive partitioning. Server Assistant allows administrators to select the file system type and partitioning details for the OS installation. In a Linux installation, Server Assistant is designed to let administrators resize each standard partition. In a Windows installation, administrators can choose the type and size of the C: drive. Server Assistant supports Windows installation on either the NT file system (NTFS) or FAT32. When installing NetWare, administrators can size the DOS partition and the SYS volume.

Network settings. Server Assistant allows administrators to preconfigure all network settings for the server. Administrators can configure IP settings, including IP address, gateway, Domain Name System (DNS), and other OS-specific addresses—such as Windows Internet Naming Service (WINS) addresses for Windows installations and Internetwork Packet Exchange™ (IPX™) addresses for NetWare installations. Server Assistant also enables administrators to add the server into the current network domain or workgroup during OS installation.

Windows network services installation. In addition to the base OS installation, the advanced mode of Server Assistant allows administrators to install optional network service components. For example, on the interview pages for Windows 2000 installation, Server Assistant lets administrators select Microsoft Cluster Service, Terminal Services, Microsoft Internet Information Services (IIS), DNS, Dynamic Host Configuration Protocol (DHCP), and WINS. By combining all separate installation options into one process, Server Assistant is designed to reduce the overall setup time for systems that provide the major services in a network.

SNMP configuration. The advanced mode network configuration page of Server Assistant also allows administrators to preconfigure Simple Network Management Protocol (SNMP) settings. Given the proper SNMP settings, the server can start sending SNMP traps to the network as soon as the OS installation is complete. These traps can carry alert information for the system status, and they can be monitored with management applications such as Dell OpenManage IT Assistant.

User and licensing information. Administrators can enter user names, organization names, and product ID numbers for Microsoft Windows installations. If administrators possess a Windows CD provided by Dell, which has a built-in product ID (Windows Server 2003 only), Server Assistant can automatically obtain it.

Sample Linux and Windows Custom Install Scripts

```
mkdir /mnt/net_share
mount -t smbfs -o username=user1,password=user1_password
   //10.1.1.5/server_log /mnt/net_share
mkdir /mnt/net_share/profiles/$HOSTNAME
cp /root/replication/* /mnt/netshare/profiles/$HOSTNAME/.
umount /mnt/net_share

mkdir /mnt/apps
mount 10.1.1.3:/shared_apps /mnt/apps
mkdir /repository
cp /mnt/apps/utilities/utility1.3.4-5.rpm /repository/.
rpm -ivh /repository/utility1.3.4-5.rpm

adduser -g root -p user1_passwd user1
```

Figure B. Sample Custom Install Script for a Linux platform

```
@echo Setting up user downloads share
net use z: \\user_share\downloads user1_password
   /user:CORP\user1 /persistent:y
mkdir c:\user1\downloads
cp z:\profiles\* c:\user1\downloads

@echo Downloading applications to be installed
net use y: \\data_server\\apps user1_password
   /user:CORP\user1 /persistent:y
cp y:\network_apps\*.exe C:\Documents and
   Settings\Administrator\Desktop
cp y:\user_apps\*.exe C:\Documents and Settings\
   Administrator\Desktop

@echo Adding entry to routing table
route -p ADD 192.168.1.1 MASK 255.255.0.0
   192.168.254.254 METRIC 2 IF 2
```

Figure C. Sample Custom Install Script for a Windows platform

Installing the .NET Client for Dell OpenManage IT Assistant 6.5

By Terry Schroeder and Mary Jean Raatz

Silent installation options script

```
[InstallShield Silent]
Version=v5.00.000
File=Response File
[File Transfer]
OverwriteReadOnly=NoToAll
[DlgOrder]
Dlg0=SdShowInfoList-0
Count=16
Dlg1=SdLicense-0
Dlg2=SdComponentDialogAdv-0
Dlg3=SdLicense-1
Dlg4=SdShowInfoList-1
Dlg5=SdLicense-2
Dlg6=SdComponentDialogAdv-1
Dlg7=SdOptionsButtons-0
Dlg8=SdShowDlgEdit2-0
Dlg9=SdOptionsButtons-1
Dlg10=SdComponentDialogAdv-2
Dlg11=SdOptionsButtons-2
Dlg12=SdShowDlgEdit2-1
Dlg13=SdStartCopy-0
Dlg14=SdFinish-0
Dlg15=RebootDialog-0
[SdShowInfoList-0]
Result=1
[SdLicense-0]
Result=1
[SdComponentDialogAdv-0]
szDir=C:\Program Files\Dell\OpenManage\IT Assistant
Web Browser User Interface Component-type=string
Web Browser User Interface Component-count=1
Web Browser User Interface Component-0=Web Browser
     User Interface Component\Dell Support Files
Component-type=string
Component-count=2
Component-0=OCX_INST
Component-1=Web Browser User Interface Component
Result=12
[SdLicense-1]
Result=12
[SdShowInfoList-1]
Result=1
[SdLicense-2]
Result=1
[SdComponentDialogAdv-1]
szDir=C:\Program Files\Dell\OpenManage\IT Assistant
Web Browser User Interface Component-type=string
Web Browser User Interface Component-count=1
Web Browser User Interface Component-0=Web Browser
     User Interface Component\Dell Support Files
Component-type=string
Component-count=2
Component-0=OCX_INST
Component-1=Web Browser User Interface Component
Result=1
[SdOptionsButtons-0]
Result=102
[SdShowDlgEdit2-0]
szEdit1=localhost
szEdit2=2607
Result=12
[SdOptionsButtons-1]
Result=12
[SdComponentDialogAdv-2]
szDir=C:\Program Files\Dell\OpenManage\IT Assistant
Web Browser User Interface Component-type=string
Web Browser User Interface Component-count=1
Web Browser User Interface Component-0=Web Browser
     User Interface Component\Dell Support Files
Component-type=string
Component-count=2
Component-0=OCX_INST
Component-1=Web Browser User Interface Component
Result=1
[SdOptionsButtons-2]
Result=102
[SdShowDlgEdit2-1]
szEdit1=localhost
szEdit2=2607
Result=1
[SdStartCopy-0]
Result=1
[Application]
Name=Dell OpenManage IT Assistant
Version=6.5.0.61
Company=Dell Computer Corporation
Lang=0009
[SdFinish-0]
Result=1
bOpt1=1
bOpt2=0
[RebootDialog-0]
Result=0
Choice=0
```

ONLINE-EXTRA CONTENT

Defending Networks with Intrusion Detection Systems

By Jose Maria Gonzalez

Sample attack data

```
[**] [1:483:2] ICMP PING CyberKit 2.2 Windows [**]
[Classification: Misc activity] [Priority: 3]
02/04-13:00:19.837833 172.168.1.0/24 -> 192.168.1.0/24
ICMP TTL:118 TOS:0x0 ID:12687 IpLen:20 DgmLen:92
Type:8  Code:0  ID:512   Seq:57484  ECHO
[Xref => http://www.whitehats.com/info/IDS154]

[**] [1:2003:2] MS-SQL Worm propagation attempt [**]
[Classification: Misc Attack] [Priority: 2]
02/04-13:07:04.007822 172.168.1.0/24:2753 ->
   192.168.1.0/24:1434
UDP TTL:111 TOS:0x0 ID:665 IpLen:20 DgmLen:404
Len: 376
[Xref => http://vil.nai.com/vil/content/v_99992.htm][Xref =>
   http://www.securityfocus.com/bid/5311][Xref =>
   http://www.securityfocus.com/bid/5310]

[**] [112:1:1] (spp_arpspoof) Unicast ARP request [**]
02/04-13:34:10.848523

[**] [111:2:1] (spp_stream4) possible EVASIVE RST detection [**]
02/04-13:43:45.287900 172.168.1.0/24:35704 ->
   192.168.1.0/24:135
TCP TTL:15 TOS:0x0 ID:0 IpLen:20 DgmLen:43
***A*R** Seq: 0x0  Ack: 0x0  Win: 0x0  TcpLen: 20

[**] [119:12:1] (http_inspect) APACHE WHITESPACE (TAB) [**]
02/04-13:49:28.028102 172.168.1.0/24:49973 ->
   192.168.1.0/24:80
TCP TTL:43 TOS:0x0 ID:3738 IpLen:20 DgmLen:1420 DF
***A*R** Seq: 0x517F0E61  Ack: 0x5913BF36  Win: 0x7BFC
   TcpLen: 20

[**] [111:17:1] (spp_stream4) TCP TOO FAST RETRANSMISSION
   WITH DIFFERENT DATA SIZE (possible fragroute) detection [**]
02/04-13:49:43.918191 172.168.1.0/24:50046 ->
   192.168.1.0/24:80
TCP TTL:43 TOS:0x0 ID:14874 IpLen:20 DgmLen:1204 DF
***A*R** Seq: 0x3ADF9942  Ack: 0x6723D51E  Win: 0x7BFC
   TcpLen: 20

[**] [111:16:1] (spp_stream4) TCP CHECKSUM CHANGED ON
   RETRANSMISSION (possible fragroute) detection [**]
02/04-13:55:41.158090 172.168.1.0/24:50598 ->
   192.168.1.0/24:80
TCP TTL:50 TOS:0x0 ID:59039 IpLen:20 DgmLen:1420 DF
***A*R** Seq: 0x72D92F9D  Ack: 0x8BE49CA1  Win: 0x3F67
   TcpLen: 20

[**] [1:469:1] ICMP PING NMAP [**]
[Classification: Attempted Information Leak] [Priority: 2]
02/04-14:00:44.077813 172.168.1.0/24 -> 192.168.1.0/24
ICMP TTL:245 TOS:0x0 ID:40032 IpLen:20 DgmLen:28
Type:8  Code:0  ID:768   Seq:565  ECHO
[Xref => http://www.whitehats.com/info/IDS162]

[**] [1:2003:2] MS-SQL Worm propagation attempt [**]
[Classification: Misc Attack] [Priority: 2]
02/04-14:14:56.297813 172.168.1.0/24:3264 ->
   192.168.1.0/24:1434
UDP TTL:108 TOS:0x0 ID:51795 IpLen:20 DgmLen:404
Len: 376
[Xref => http://vil.nai.com/vil/content/v_99992.htm][Xref =>
   http://www.securityfocus.com/bid/5311][Xref =>
   http://www.securityfocus.com/bid/5310]

[**] [1:1149:9] WEB-CGI count.cgi access [**]
[Classification: access to a potentially vulnerable web
   application] [Priority: 2]
02/04-15:50:37.339193 172.168.1.0/24:50863 ->
   192.168.1.0/24:80
TCP TTL:128 TOS:0x0 ID:20578 IpLen:20 DgmLen:502 DF
***AP*** Seq: 0x18E14DF8  Ack: 0x15F575B3  Win: 0xFC00
   TcpLen: 20
[Xref => http://cgi.nessus.org/plugins/dump.php3?id=10049]
   [Xref => http://cve.mitre.org/cgi-bin/cvename.cgi?name=
   CVE-1999-0021][Xref => http://www.securityfocus.com/bid/128]

[**] [1:1413:2] SNMP private access udp [**]
[Classification: Attempted Information Leak] [Priority: 2]
02/04-17:00:27.017884 172.168.1.0/24:38523 ->
   192.168.1.0/24:161
UDP TTL:55 TOS:0x0 ID:33923 IpLen:20 DgmLen:70 DF
Len: 42
[Xref => http://cve.mitre.org/cgi-bin/cvename.cgi?name=
   CAN-2002-0013][Xref => http://cve.mitre.org/cgi-bin/
   cvename.cgi?name=CAN-2002-0012]

[**] [1:1417:2] SNMP request udp [**]
[Classification: Attempted Information Leak] [Priority: 2]
02/04-17:00:27.018405 172.168.1.0/24:38525 ->
   192.168.1.0/24:161
UDP TTL:55 TOS:0x0 ID:33923 IpLen:20 DgmLen:67 DF
Len: 39
[Xref => http://cve.mitre.org/cgi-bin/cvename.cgi?name=
   CAN-2002-0013][Xref => http://cve.mitre.org/cgi-bin/
   cvename.cgi?name=CAN-2002-0012]

[**] [1:620:5] SCAN Proxy Port 8080 attempt [**]
[Classification: Attempted Information Leak] [Priority: 2]
02/04-17:01:57.957803 172.168.1.0/24:57619 ->
   192.168.1.0/24:8080
TCP TTL:55 TOS:0x0 ID:22607 IpLen:20 DgmLen:60 DF
******S* Seq: 0xA3A3F047  Ack: 0x0  Win: 0x16D0  TcpLen: 40
TCP Options (5) => MSS: 1460 SackOK TS: 397453320 0 NOP WS: 0

[**] [1:2049:1] MS-SQL ping attempt [**]
[Classification: Misc activity] [Priority: 3]
02/04-17:02:10.577703 172.168.1.0/24:41991 ->
   192.168.1.0/24:1434
UDP TTL:55 TOS:0x0 ID:43977 IpLen:20 DgmLen:29 DF
Len: 1
[Xref => http://cgi.nessus.org/plugins/dump.php3?id=10674]

[**] [1:1616:4] DNS named version attempt [**]
[Classification: Attempted Information Leak] [Priority: 2]
02/04-17:02:44.687917 172.168.1.0/24:41995 ->
   192.168.1.0/24:53
UDP TTL:55 TOS:0x0 ID:47690 IpLen:20 DgmLen:58 DF
Len: 30
[Xref => http://www.whitehats.com/info/IDS278][Xref =>
   http://cgi.nessus.org/plugins/dump.php3?id=10028]
```

```
[**] [1:524:6] BAD-TRAFFIC tcp port 0 traffic [**]
[Classification: Misc activity] [Priority: 3]
02/04-17:03:37.397808 172.168.1.0/24:58422 ->
   192.168.1.0/24:0
TCP TTL:55 TOS:0x0 ID:55667 IpLen:20 DgmLen:60 DF
******S* Seq: 0xA97A1792  Ack: 0x0  Win: 0x16D0  TcpLen: 40
TCP Options (5) => MSS: 1460 SackOK TS: 397463265 0 NOP WS: 0

[**] [1:1867:1] MISC xdmcp info query [**]
[Classification: Attempted Information Leak] [Priority: 2]
02/04-17:04:00.807816 172.168.1.0/24:42001 ->
   192.168.1.0/24:177
UDP TTL:55 TOS:0x0 ID:55300 IpLen:20 DgmLen:36 DF
Len: 8
[Xref => http://cgi.nessus.org/plugins/dump.php3?id=10891]

[**] [1:1893:1] SNMP missing community string attempt [**]
[Classification: Misc Attack] [Priority: 2]
02/04-17:04:11.687820 172.168.1.0/24:42002 ->
   192.168.1.0/24:161
UDP TTL:55 TOS:0x0 ID:56390 IpLen:20 DgmLen:70 DF
Len: 42
[Xref => http://cve.mitre.org/cgi-bin/cvename.cgi?name=
   CAN-1999-0517]

[**] [111:9:1] (spp_stream4) STEALTH ACTIVITY (NULL scan)
   detection [**]
02/04-17:07:57.017823 172.168.1.0/24:137 ->
   192.168.1.0/24:137
TCP TTL:244 TOS:0x0 ID:47626 IpLen:20 DgmLen:40
******** Seq: 0xF1C  Ack: 0x0  Win: 0x200  TcpLen: 20

[**] [1:1384:3] MISC UPnP malformed advertisement [**]
[Classification: Misc Attack] [Priority: 2]
02/04-17:12:33.167839 172.168.1.0/24:1900 ->
   192.168.1.0/24:1900
UDP TTL:55 TOS:0x0 ID:27002 IpLen:20 DgmLen:282
Len: 254
[Xref => http://cve.mitre.org/cgi-bin/cvename.cgi?name=
   CAN-2001-0877][Xref => http://cve.mitre.org/cgi-bin/
   cvename.cgi?name=CAN-2001-0876]

[**] [116:46:1] (snort_decoder) WARNING: TCP Data Offset is
   less than 5! [**]
02/04-17:12:33.168303 172.168.1.0/24:0 -> 192.168.1.0/24:0
TCP TTL:55 TOS:0x0 ID:27258 IpLen:20 DgmLen:40
*****R** Seq: 0xF90  Ack: 0x0  Win: 0x2000  TcpLen: 0

[**] [1:1384:3] MISC UPnP malformed advertisement [**]
[Classification: Misc Attack] [Priority: 2]
02/04-17:12:34.227892 172.168.1.0/24:1900 ->
   192.168.1.0/24:1900
UDP TTL:55 TOS:0x0 ID:27002 IpLen:20 DgmLen:282
Len: 254
[Xref => http://cve.mitre.org/cgi-bin/cvename.cgi?name=
   CAN-2001-0877][Xref => http://cve.mitre.org/cgi-bin/
   cvename.cgi?name=CAN-2001-0876]

[**] [1:634:2] SCAN Amanda client version request [**]
[Classification: Attempted Information Leak] [Priority: 2]
02/04-17:24:38.567703 172.168.1.0/24:42214 ->
   192.168.1.0/24:10080
UDP TTL:55 TOS:0x0 ID:48007 IpLen:20 DgmLen:97 DF
Len: 69

[**] [105:1:1] spp_bo: Back Orifice Traffic detected (key:
   31337) [**]
02/04-17:48:31.487823 172.168.1.0/24:42906 ->
   192.168.1.0/24:31337
UDP TTL:55 TOS:0x0 ID:59626 IpLen:20 DgmLen:46 DF
Len: 18

[**] [1:236:3] DDOS Stacheldraht client check gag [**]
[Classification: Attempted Denial of Service] [Priority: 2]
02/04-17:51:12.167810 172.168.1.0/24 -> 192.168.1.0/24
ICMP TTL:55 TOS:0x0 ID:13330 IpLen:20 DgmLen:39
Type:0  Code:0  ID:668  Seq:0  ECHO REPLY
[Xref => http://www.whitehats.com/info/IDS194]

[**] [1:239:1] DDOS shaft handler to agent [**]
[Classification: Attempted Denial of Service] [Priority: 2]
02/04-17:54:33.387701 172.168.1.0/24:1024 ->
   192.168.1.0/24:18753
UDP TTL:244 TOS:0x0 ID:2304 IpLen:20 DgmLen:49
Len: 21
[Xref => http://www.whitehats.com/info/IDS255]
```