DELL™
FEBRUARY 2006 · $12.95

POWER SOLUTIONS

THE MAGAZINE FOR DIRECT ENTERPRISE SOLUTIONS

Architecting a Blueprint for Disaster Recovery

Also Inside This Issue:

Best Practices for Enterprise Testing

Multi-Node Scalability Practices for Oracle 10*g* RAC

Highly Available Cluster Services on SAP

Uptime expert.

It could be you.

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

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

Intel® PRO
Network Connections

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

Learn more: intel.com/go/adapters

intel®

DELL POWER SOLUTIONS
THE MAGAZINE FOR DIRECT ENTERPRISE SOLUTIONS
FEBRUARY 2006

COVER STORY | PAGE 8

Architecting a Blueprint for Disaster Recovery

By Rich Armour, Paul Eno, Michael Kimble, and Jesse Freund

Dell's top 10 rules for the disaster planning process can help organizations align recovery efforts with overall business objectives. This article provides guidance to minimize downtime of mission-critical systems and help reduce business consequences while organizations restore enterprise-wide operations in a logical manner.

TABLE OF CONTENTS

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

EDITORIAL

EDITOR-IN-CHIEF | Tom Kolnowski
MANAGING EDITOR | Debra McDonald
SENIOR EDITORS | Liza Graffeo, Bob Johnson

CONTRIBUTING AUTHORS | Mahmoud Ahmadian; Rich Armour; Matthew Brisse; Michael E. Brown; Ravikanth Chaganti; Balasubramanian Chandrasekaran; Babu Chandrasekhar; Shaun Connolly; David Detweiler; Mohammad Dhedhi; Jake Diner; Paul Eno; Yung-Chin Fang; Anthony Fernandez; Jesse Freund; Bhushan Gavankar; Steven Grigsby; Richard Goodwin; Manoj Gujarathi; Kevin Guinn; Aziz Gulbeden; Drew Habas; Kyon Holman; Bruce Holmes; Munira Hussain; Dave Jaffe, Ph.D.; Javier Jimenez; Michael Kimble; Florenz Kley; Pranaya Kondekar; Chethan Kumar; Ed Lawrence; Achim Lernhard; Kit Lou; Cynthia Lovin; Zafar Mahmood; David Mar; Tesfamariam Michael; Jatin N. Muddu; Todd Muirhead; Srikrishna Sridhar Murthy; Bob Ng; Cuong T. Nguyen; Dat Nguyen; Nam Nguyen; Ron Pepper; Ujjwal Rajbhandari; Nathan Rakoff; Drue Reeves; Norman Richards; Vishvesh Sahasrabudhe; Ananda Sankaran; Bert Scalzo, Ph.D.; John Sieber; Sanjeev Singh; Thorsten Staerk; Scott Stanford; Eric Szewczyk; Ahmad Tawil; Prathap Thathireddy; Catherine J. Tilton; Wolfgang Trenkle; Murali Vallath; Bharath Vasudevan; Erik Veerman; Kevin Winert; and Tony Yaptangco

ART

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

MARKETING

MARKETING MANAGER | Kathy White

ONLINE

WEB PRODUCTION | Brad Klenzendorf

SUBSCRIPTION AND EDITORIAL SERVICES

EDITORIAL ASSISTANT | Amy Hargraves

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

ABOUT DELL

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

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

February 2006

TALK BACK

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

EMC²
where information lives®

EMC's File System Archiving for Windows Solution won't leave you lost in a sea of servers.

EMC® Centera™ and EMC DiskXtender® provide a complete data archiving solution. Many organizations are attempting to solve their growing storage problems by continually adding production capacity. There is a better, simpler way. Introducing File System Archiving for Windows from EMC. The industry leader in advanced storage technology, EMC delivers a complete archiving solution at a price you can afford. Experience reduced overhead, faster backups and restores, and improved IT resource efficiencies.

It's not just about server capacity anymore.

- Automatically move inactive data to more cost-effective storage while freeing up primary server capacity
- Reduce backups and optimize quality
- Store infrequently accessed data in a self-managed, active archive that does not require backup
- Archive data transparently and restore upon user or application request

Let EMC's integrated archiving solutions help you better manage your data at your lowest total cost.

EMC® Centera™ and EMC DiskXtender® are included within EMC's File System Archiving for Windows Solution.

DELL | EMC²

Dell and EMC bring you solutions that are high on results—and simple to use. That's because we've made it easier than ever to put premium software, robust storage and world-class technical support to work solving your business' critical IT challenges.

For an assessment of your storage system, call us toll-free at 800-BUY-DELL, or visit us online at emc.com/centera.

TABLE OF CONTENTS

ADVERTISER INDEX

SEE IT HERE FIRST!

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

ONLINE EXTRA

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

WWW.DELL.COM/ POWERSOLUTIONS

Remote Cluster Management Using Dell OpenManage Tools

By Yung-Chin Fang, Ron Pepper, Munira Hussain, Vishvesh Sahasrabudhe, and Aziz Gulbeden

The Dell OpenManage suite can facilitate effective remote monitoring and management of high-performance computing clusters.

Upgrading to Microsoft Windows Server 2003 SP1 and Dell OpenManage 4.4

By Mohammad Dhedhi and Kit Lou

Release 4.4 of the Dell OpenManage suite—which enables proactive monitoring, diagnosis, notification, and remote-access capabilities for Dell PowerEdge servers—introduces support for Microsoft Windows Server 2003 Service Pack 1.

Deploying Microsoft Windows 2000 SP4 Update Rollup 1 on Dell PowerEdge Servers

By Bhushan Gavankar

To help administrators maintain the security and stability of Microsoft Windows 2000 systems, Microsoft has released Update Rollup 1 for Windows 2000 Service Pack 4.

The Role of Biometrics in Enterprise Security

By Catherine J. Tilton

Biometrics can be used as an authentication mechanism in the enterprise IT environment, replacing or augmenting a standard password.

OS Deployment Using Dell OpenManage Server Assistant and Preboot Execution Environment

By Michael E. Brown and Manoj Gujarathi

Dell OpenManage Server Assistant and the Preboot Execution Environment can be used together to perform standards-compliant network booting.

Efficient Bare-Metal Configuration of Dell PowerEdge Servers Using the Dell OpenManage Deployment Toolkit 2.0

By Pranaya Kondekar and Kevin Winert

The Dell OpenManage Deployment Toolkit 2.0 provides enhanced, automated, script-based provisioning tools that can help increase productivity and boost system reliability.

DELL AND COX ARE KEEPING

6.6 MILLION CUSTOMERS TUNED IN, IN TOUCH AND CONNECTED.

COX COMMUNICATIONS' DATACENTER OF THE FUTURE.

When Cox, an industry-leading broadband communications company, needed a new IT infrastructure to handle their mission-critical operations, they partnered with Dell. The Dell solution handles the most sensitive data Cox uses to run its business from core financial records, to supply chain management, to compliance and more. With Dell, Cox Communications is getting the technology and services they need to connect with their customers. Isn't it time we did the same for you?

Click www.dell.com/coxcommunications
Call (toll free) 1.866.212.9335

New Year, New Colors

It can be argued that contemporary black-and-white artwork is defined by the photography of Ansel Adams. Through his iconic landscapes of the American West, including Grand Canyon National Park and Yosemite National Park, Adams captured the power of a two-color world in carefully crafted shades of grey.

From its inception, *Dell Power Solutions* has paid homage to Adams in being a predominately black-and-white publication. With the exception of four-color covers and advertisements, the typical *Dell Power Solutions* editorial page was produced via a two-color process that interspersed blue inks to create the illusion of color.

In the February 2006 issue, however, we have parted ways with Ansel Adams. Besides introducing full-color photographs and screen captures, we have infused color into figures and tables to enhance the reading experience—both in the print edition and on our Web site at www.dell.com/powersolutions. To demonstrate the difference, Figure 1 portrays a typical technical diagram in two colors versus four colors. Look for more design changes later in 2006 as we fully embrace four-color.

While there has been a flurry of art activity in our editorial offices, we have continued to raise the bar for our core competency as a technical journal. Turn to page 52 for the first installment in a new series of articles on testing enterprise solutions, "Best Practices: Enterprise Testing Fundamentals." This inaugural article explores recommendations for a structured approach to component-, feature-, and system-level testing—encompassing the full spectrum of phases from alpha testing to out-of-the-box audits.

This issue also marks our first-ever cover story on the critical topic of enterprise disaster recovery. Starting on page 8, "Architecting a Blueprint for Disaster Recovery" leads the way with Dell's top 10 rules for disaster planning, discussing best practices for aligning disaster recovery and business continuity efforts with overall business objectives. Additional articles in the disaster recovery section delve into perspectives on disaster recovery and business continuity for Microsoft® Windows® environments as well as for Oracle® databases on Linux® platforms.

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

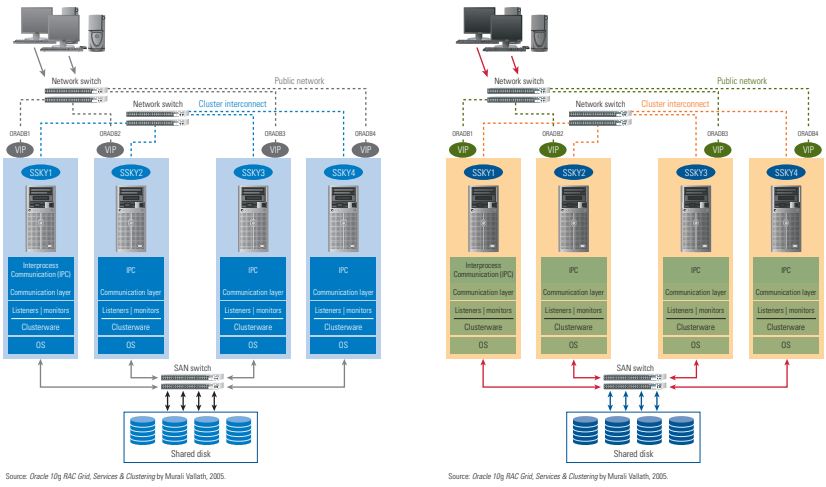


Figure 1. Transitioning from two-color to four-color diagrams

NO VIRUSES.
NO SPAM.
NO DOWNTIME.
EMAIL DONE RIGHT.

No one can promise complete email security and availability. We don't live in that kind of world. Yet one company has earned a worldwide reputation for making email as secure and available as it is important. A company that not only screens out viruses, spam and spyware, but also provides solutions for speedy recovery in case of system failure. A company that reduces storage costs by archiving to secondary storage and blocking unwanted emails. A company that provides management tools for efficient email retention and fast email discovery. A company that does email right. Symantec. Because we know it's not just email, it's your business. For more information visit www.symantec.com/esa. **BE FEARLESS.**

symantec™

Architecting a Blueprint for Disaster Recovery

Effective risk assessment and business continuity provisions enable organizations to minimize downtime and recover crucial applications quickly when disaster strikes. By analyzing business processes and functions and defining the impact of system downtime in financial terms, enterprises can effectively prioritize the enterprise-wide recovery process to help reduce business consequences. Based on internally honed best practices, Dell's top 10 rules for disaster planning can help enterprises align recovery efforts with overall business objectives.

BY RICH ARMOUR, PAUL ENO, MICHAEL KIMBLE, AND JESSE FREUND

Related Categories:

Backup

Business continuity

Data center technology

Dell PowerEdge servers

Dell PowerVault storage

Dell/EMC storage

Disaster recovery

Enterprise management

Planning

Storage

Tape backup

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

These days it is not enough merely to build great applications. Enterprises of all sizes have come to realize that essential components for success and longevity are high availability and easy recoverability for business-critical systems. This realization has led to a focus on business continuity planning, which can help organizations identify critical processes and technologies, maintain and recover functionality after a planned or unplanned event, and balance the risks and costs of disaster recovery.

Dell is no stranger to business continuity planning. Five years ago, Dell decided to overhaul its own business continuity and disaster recovery plan. Since that time, Dell has learned a number of lessons from its efforts to refine and improve the plan. While most organizations are aware of the need to classify and tier data and applications, Dell has explored various approaches that go beyond the basics. Dell's top 10 rules for disaster recovery planning can help enterprises draft their own blueprints for effective business continuity and disaster recovery decisions.

Rule 1: Articulate the need in financial terms

It is no secret that business decisions should drive business continuity planning. The lack of an effective business continuity and disaster recovery plan can expose organizations to considerable financial risk. Aside from lost income, interrupted service can besmirch a good brand name. Plus, the disruption of service in regulated industries—such as healthcare—increases the risk of legal liability.

It is essential that enterprises affix a monetary value to business continuity planning efforts to help ensure buy-in from leadership across the organization. Additionally, enterprises should track financial progress as they move through the implementation of the business continuity plan. After all, the benefits of business continuity planning do not arrive en masse the day the organization flips the switch on its new strategy. Many disaster recovery benefits are associated with incremental steps to reduce the lead time to recovery, and as organizations implement technologies and processes, this lead time begins to decrease. In the end, by viewing the

business continuity strategy in monetary terms and tracking financial progress over time, organizations can align business continuity efforts with smart business decisions.

Rule 2: Use hard data to create a risk profile

Because Dell's data centers are located in Texas, the Dell business continuity team might have assumed that the company's primary risk would be a catastrophic tornado. In fact, after analyzing hard data, the team determined that the primary threat to data centers is likely to be fire.

How did Dell construct its risk profile? The company invited its insurance carriers to bring hard data to the business continuity plan. Insurance carriers possess claims data that can tell organizations what the risks are in a given geographic area. Plus, insurance carriers naturally want to help clients reduce risk. In the end, the use of hard claims data from the insurance carriers offered some of the most valuable and surprising information as Dell developed its own business continuity plan.

Rule 3: Identify the critical resources

All data is not created equal, and the same holds true for applications. An effective business continuity planning process requires organizations to undergo a rigorous analysis of business processes and functions and to identify the critical resources that require redundancy, backup, and recovery. Before organizations can discuss the IT resources necessary to maintain business-critical processes, they must assess the business impact of losing systems—paying particular attention to interdependencies that exist among systems. For example, a business-critical system may rely on input from another system that is not deemed critical in and of itself. It is crucial that the business side of the organization lead the discussion of critical processes before the IT side can define the technologies necessary to enable business continuity.

Dell has developed a three-tier strategy for classifying data and applications. Class 1 systems support business-critical processes. For example, at Dell a business-critical process involves any service that directly interacts with the customer, which includes taking and processing orders as well as building, shipping, and servicing products. By contrast, Class 2 systems correlate to business-essential processes, where a 48-hour outage would begin to have a negative business impact. Class 3 systems enable business-support processes, for which a temporary loss of service is deemed noncritical. By classifying and defining processes, applications, and data along business-criticality demarcations, Dell helps ensure that the appropriate investment is made to recover the most crucial systems first.

Rule 4: Think beyond the data center

Many disaster recovery efforts are focused on keeping the data center up and running. However, effective business continuity planning must reach beyond applications, data, and infrastructure considerations. For example, it does no good to have the data center up and running if no provisions have been made to support people performing vital business functions such as shipping and receiving. Of course, applications must be available, data must be accessible, and the network must be working. But by focusing on the data center at the expense of essential business processes and infrastructure components, enterprises run the risk of turning robust data center functionality into little more than a paper tiger.

TOP 10 RULES FOR DISASTER PLANNING

1. *Articulate the need in financial terms.*
2. *Use hard data to create a risk profile.*
3. *Identify the critical resources.*
4. *Think beyond the data center.*
5. *Eliminate or mitigate single points of failure.*
6. *Assume that everything is going to fail.*
7. *Consider an active/active data center strategy.*
8. *Recognize potential vendor weaknesses.*
9. *Keep disaster recover capability up-to-date.*
10. *Perform tests on a regular basis.*

Rule 5: Eliminate or mitigate single points of failure

A single point of failure occurs when there is no redundancy to compensate for a missing application, data, or infrastructure component. It may be an application or a database server, a lone backup generator in a data center, or the long-haul network itself. Moreover, several single points of failure typically exist within an enterprise infrastructure. Consequently, organizations should perform a specific and detailed single-point-of-failure analysis across the entire infrastructure. Doing so may reveal that a key component was missed when a data center, or another form of disaster recovery system, was built. In the end, performing a single-point-of-failure analysis may help prevent an organization from having to entirely reconstitute business capabilities when a relatively minor component fails.

Rule 6: Assume that everything is going to fail

Oftentimes, when enterprises build a disaster recovery plan, they do so with the expectation that land lines, cell phones, and the network will be available. Or they take for granted that the roads to the data center will be accessible—assuming the data center itself is still standing and the employees are capable of getting there. The best-laid plans for business continuity include the consideration that every key piece of internal and external infrastructure may fail or become unavailable for extended periods.

At Dell, the operating assumption is that every vital piece of infrastructure is capable of failing, and all of them may go down at the same time. Along these lines, the Dell recovery plan itself is stored on CD. Copies of the CD are distributed across multiple teams. At least one copy resides in each data center, and another

INTELLIGENT DATA PROTECTION: THE DELL POWERVAULT ML6000 TAPE LIBRARY

The Dell™ PowerVault™ ML6000 modular tape library is designed to intelligently protect an organization's critical data. Through proactive diagnostics and flexible scalability, the PowerVault ML6000 enables organizations to prepare their storage environments for disaster recovery.

The built-in intelligence of the PowerVault ML6000 helps ensure that backups can execute as planned. The library's diagnostics are designed to predict failures in the library's environment, such as tape or drive malfunction, and send e-mail messages to warn administrators of potential issues. This proactive notification allows administrators to plan ahead and resolve problem conditions before failures occur—thus limiting unplanned downtime. If a failure occurs, administrators can use one of the library's simple troubleshooting wizards, which provide solutions to known issues ranging from cleaning or restarting drives to opening a service ticket. The wizards help administrators save time by resolving issues on-site. If administrators cannot resolve an issue using the wizards, detailed event logs and built-in relational diagnostics isolate failures at a subcomponent level, minimizing the time to repair of the PowerVault ML6000.

Figure A. Dell PowerVault ML6010 CM tape library

The tape library's modular and versatile scalability provides organizations with several capacity options, offering organizations the flexibility to pay as they grow without limiting their ability to add more drives or slots to the existing library. The PowerVault ML6000 can scale from 14.4 TB of native backup storage capacity to 51.2 TB (native) using 9U expansion modules to support the demands of workgroup and mid-range data centers; up to 161 TB of capacity is planned for future releases. The library scales from two to six Ultrium 3 Linear Tape-Open (LTO-3) SCSI or Fibre Channel drives; future support is planned for up to 18 LTO-3 SCSI or Fibre Channel drives. In addition, from 36 to 128 cartridge slots give organizations added backup performance and capacity; support for up to 404 cartridge slots is planned for future releases. To satisfy ever-shrinking backup windows, the PowerVault ML6000 is designed to provide a maximum native transfer rate of 1.7 GB/hour.

The PowerVault ML6000 control module can be placed anywhere in the expansion stack so that organizations can easily expand and customize their libraries. Each additional expansion module leverages the existing robotics and intelligence of the control module to reduce the overall number of moving parts, thus enhancing the library's reliability.

The tape library also offers a wide range of connectivity and compatibility options for protecting storage environments. The drive technology used in the PowerVault ML6000 can expand to support different connectivity standards, including SCSI and Fibre Channel. The library is also compatible with storage software from CommVault, EMC, and Symantec.

In addition, the PowerVault ML6000 is available with two Dell Services offerings. The Backup and Recovery Design and Deployment service provides a detailed backup and recovery plan that is designed to help organizations establish appropriate procedures to minimize or avoid data loss. Meanwhile, the Backup and Recovery Implementation service is designed to be a comprehensive approach for organizations implementing a backup and recovery process on a new Dell or Dell/EMC storage area network or network attached storage solution, and may include software from key third-party technology providers.

The drive and media technology used by the PowerVault ML6000 provides robust backup and restore functionality as well as effective, long-term data retention. For example, the library's use of LTO-3 drives and WORM (write once, read many) media aids in regulatory compliance by preventing data from being overwritten or digitally altered while providing excellent tape drive performance.

The Dell PowerVault ML6000 tape library is optimized for Dell PowerEdge™ servers and Dell's comprehensive storage portfolio. For more information, visit www.dell.com/storage.

copy is kept in the IT operations center. This way, if a disaster cripples phone service, Internet availability, or transportation infrastructure, Dell still has the capability to begin recovery.

Rule 7: Consider an active/active data center strategy

One way to know that a recovery plan will work is to make it a part of the load-balancing activities. Along these lines, Dell relies on an active/active data center strategy as part of its everyday operations. To do so, Dell provisions more than 100 percent capacity for each application so that it can split application load balancing across multiple data centers. Each application has 75 percent of required capacity in each data center—lending each application 150 percent of its nominal capacity requirement. Not only does this load-balancing strategy translate to high-performance applications, but it also helps ensure that disaster recovery and failover capabilities are being tested every moment of every day. This way, when Dell needs to implement its disaster recovery plan, the company knows it will work because it is already part of the existing load-balancing strategy.

Rule 8: Recognize potential vendor weaknesses

Critical vendors can significantly affect an enterprise's capacity for disaster recovery. When putting together a business continuity plan, organizations must evaluate a vendor's own disaster recovery capabilities to understand how their potential weaknesses might hinder the enterprise. For example, after the 9/11 terrorist attack, many customers turned to Dell to rapidly reconstitute systems that had been destroyed or damaged. Thanks to its size and scalability, Dell was able to respond quickly to get these customers back online. Had Dell been smaller and less scalable, hardware procurement might have become a significant bottleneck in its customers' disaster recovery capabilities.

Rule 9: Keep disaster recovery capability up-to-date

Disaster recovery capability can quickly become outdated. It must be maintained by a strong set of procedures and processes, so it becomes part of the everyday, every project, and every implementation culture. As each new project or application is initiated, enterprises need to perform an analysis of where it fits in the criticality matrix. If, for instance, a new application is deemed to support a Class 1 business process, then the application must be engineered with the appropriate recoverability. Plus, that capability must be maintained going forward. As changes are made to applications, databases, and data centers, disaster recovery capabilities should be updated as well.

Rule 10: Perform tests on a regular basis

Enterprises can never assume that their disaster recovery capability is actually working. Dell tests its failover and recovery processes on a quarterly basis. Aside from validating that Dell does indeed have failover and recovery capabilities, these quarterly tests help keep the business continuity plan in front of the infrastructure and application teams, which helps encourage future development with business continuity in mind. The quarterly tests represent an important part of the effort to make disaster recovery part of the everyday, every project, and every implementation culture at Dell.

Align recovery efforts with business objectives

Effective disaster recovery and business continuity planning depends on an enterprise's ability to identify critical processes and technologies, maintain and recover functionality after a planned or unplanned event, and balance the risks with the costs of continuity efforts. In turn, this effectiveness requires an alignment of business continuity planning with articulated business goals. To align business objectives with continuity efforts, enterprises must develop a risk profile based on hard data. Furthermore, the business side of the organization should guide the development of the risk profile. By basing disaster recovery and business continuity efforts on business objectives and by refining these practices over time, organizations can develop a plan that not only pays dividends in the event of an unfortunate event, but also helps organizations realize efficiencies in their day-to-day operations.

Rich Armour is a director on the Dell Information Technology team. He has a B.S. in Computer Science and Mathematics from Eastern New Mexico University and an M.B.A. from George Washington University.

Paul Eno is a senior manager on the Dell Information Technology team. He has a B.S. in Engineering from the U.S. Military Academy at West Point, New York, and an M.B.A in Financial Information Management from The University of Texas at Austin.

Michael Kimble is an enterprise technologist in the Advanced Systems Group at Dell. He focuses on storage solutions for business continuity and disaster recovery.

Jesse Freund is a business and technology writer based in San Francisco. He has written about business and technology for leading publications, corporations, and organizations, including *Business 2.0* and *Wired* magazines. Jesse has a B.A. in History from the University of California, Berkeley.

FOR MORE INFORMATION

Dell business continuity:
www.dell.com/disasterrecovery
www.dell.com/enterprise
www.dell.com/storage
www.dell.com/services

Site-Wide Disaster Recovery and Business Continuity Solutions

Enterprises need an effective disaster recovery and business continuity plan to safeguard critical business processes. This article presents a survey of site-wide business continuity and disaster recovery solutions.

BY ANANDA SANKARAN, KEVIN GUINN, AND BHARATH VASUDEVAN

Related Categories:

Business continuity

Disaster recovery

Storage

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

Information systems that execute critical business processes are prone to failures from system malfunctions, human error, and disasters. Excessive downtime of such systems can result in lost revenue and productivity. An effective disaster recovery and business continuity plan is essential for mission-critical systems. Such a plan typically involves identifying and analyzing risks and implementing solutions to mitigate them.

Business continuity solutions can vary in their complexity, cost, and recovery time. They may recover only business data or business applications and data. Also, they may provide recovery only from local data center failures or from site-wide disasters.

The focus of this article is site-wide business continuity implementations. In this context, local failure is considered to be failure of entities within a single data center or campus (up to 10 km). Site failures occur when an entire data center or campus becomes unavailable. Such failures are caused by fire, flood, severe weather, long-term power outage, or any other natural or man-made disaster. Site failures are inherently more complicated to recover from than local failures, because there is a need to use or enable equipment at an alternate site.

Site-wide data recovery

Many hardware and software products exist to help ensure that data is preserved during a site-wide failure. Site-wide data recovery requires the backing up or replication of business data from live production systems onto secondary media such as magnetic tapes or disks located at a remote site. In most cases, manual intervention is required to restore the backed-up data to a new production environment given that system reconstruction and data restoration can take significant time.

Site-wide data recovery solutions provide varying levels of complexity, automation, and recovery time. Some of these solutions are outlined in the following sections.

Off-site data storage

One method of retaining data in the event of a site failure is to store a copy of the data off-site. The data can be

When information comes together, business just keeps getting better.

DELL | EMC²

ALL THE RIGHT CONSOLIDATION, BACKUP AND ARCHIVE SOLUTIONS

Whether you need fast backup and complete protection or scalable and easy-to-manage storage consolidation for your mid-size enterprise, Dell|EMC brings you solutions that are high on results—and simple to use. That's because it's easier than ever to put premium software, robust storage, and world-class technical support to work solving your business's critical IT challenges.

Entry SAN Solution

- Dell|EMC AX100 Storage Platform
- iSCSI or Fibre Channel Connectivity
- EMC® Navisphere® SAN Management Software

SAN Windows Backup Solution

- Dell|EMC CX300 Storage Platform
- EMC® Navisphere® SAN Management Software
- EMC Snapview™ and EMC Replication Manager SE Software
- EMC SAN Copy™ Software

Data Archiving Solution

- EMC Centera™ Storage Platform
- Windows File System Archive Edition with EMC DiskXtender® Software
- Governance Edition with EMC EmailXtender® and EMC DiskXtender® Software

BUSINESS SOLUTIONS FOR MIDSIZE ENTERPRISES

CALL 800.BUY.DELL toll free **www.dell.com/storage**

backed up locally before being stored off-site, or the backup process and storage both can take place off-site.

Local data backup with off-site media storage. This solution involves copying the production data onto a local tape library or low-cost secondary storage disks. The duplication process is performed periodically at consistent intervals, and the backup media is transferred to an off-site storage location (see Figure 1). These data backup solutions can range from simple file-copying software to enterprise backup systems with application-specific features.

This approach is useful when recovery time is not critical. For recovery, suitable hardware must be acquired and deployed. Furthermore, the media must be transported to the hardware location before the data can be used. Several data management companies provide media pickup, storage, and delivery services.

Off-site data backup with off-site media storage. This solution employs similar techniques and technologies as those described in the preceding section, except the data is backed up remotely using a network. The data may be moved across a host-based network or a storage-based network.

For instance, backup agents at the primary site might communicate with a remote backup server at the alternate site using TCP/IP over a metropolitan area network (MAN) or wide area network (WAN). In that case, the remote backup server must have a backup device and media pool available for use. Even if the remote backup server, backup device, and media are all available at the alternate site, additional replacement hardware will likely still be required before the data can be recovered and used. Therefore, this solution is best suited for conditions where recovery time is not critical.

In a dynamic environment, many applications must be taken offline or paused to help ensure that a consistent view of the data is available for backup. Hardware- or software-based snapshot tools such as EMC® SnapView™ software or Microsoft® Volume Shadow Copy Service can be used to produce a consistent image of the data with minimal application downtime, and the backup system can then use this image as its source.

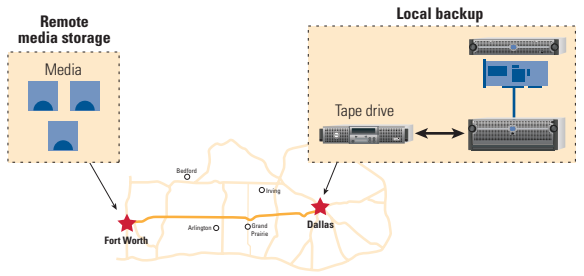


Figure 1. Off-site storage of backup data

Remote data replication

Remote data replication solutions can employ software-based or hardware-based mechanisms to create replicas of data volumes on storage devices at an alternate site. In the event of a site failure, the replicated volumes can be disassociated from the replication process, and then mounted by servers at the alternate site. As with off-site data backup, suitable server hardware must be acquired, set up, and configured before the replicated data can be used. However, recovery time using remote data replication is faster compared to either local or off-site backups because the data volumes are present in a usable form on storage devices at the alternate site.

Remote data replication solutions can employ software-based or hardware-based mechanisms to create replicas of data volumes on storage devices at an alternate site.

In many cases, businesses will have one alternate site that combines remote data replication and off-site data backup. Remote data replication provides the time-to-recover benefits of having the data in a usable form, while backup jobs are performed from the replicated data to provide an additional tier of data protection.

Software-based remote data replication. When the replication process is controlled by host-based software that copies data from a host at the primary site to another host at the alternate site via TCP/IP over a MAN or WAN, the solution is considered to be software based. Legato® RepliStor® software and NSI DoubleTake are examples of software-based remote data-replication products.

Hardware-based remote data replication (mirroring). When the replication process is controlled by storage-based features that copy data from a storage system at the primary site to a storage system at the alternate site over a storage area network (SAN), the solution is considered to be hardware based (see Figure 2).

Solutions are differentiated by the types of storage systems, the allowable distance between sites, and the replication process (synchronous or asynchronous). Many hardware-based remote data replication products are qualified for use with distance-extension solutions that enable the primary and alternate sites to be separated by more than 10 km. These distance-extension solutions include dense wavelength division multiplexing (DWDM) and Fibre Channel–to-IP gateways. EMC MirrorView™, EMC MirrorView/Asynchronous, and EMC SAN Copy™ software support hardware-based remote data replication or mirroring.

Site-wide application recovery

With some critical applications, the ability to reliably back up and recover business data alone is not sufficient because recovery from

backup data can be time-consuming and involves manual repair or reconstruction of the failed system. Automated recovery of the entire system state, including the applications, is required.

Several site-wide application recovery solutions, including remote system backup, remote standby systems, and geoclustering are outlined in the sections that follow. As with data recovery solutions, site-wide application recovery solutions are available in varying levels of complexity and offer vastly different recovery times and automation capabilities.

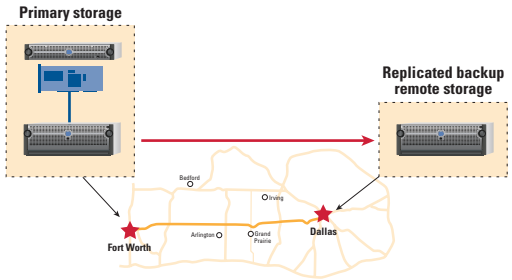
Remote system backup

To recover from site-wide failures, remote system backup solutions enable backing up application data, the OS, and the application states to an off-site location. In the absence of such solutions, system administrators must reconfigure hardware to the required settings, reinstall the OS and applications, and restore application data for complete recovery. System backup solutions can significantly automate the otherwise-lengthy system recovery process.

Many backup tools have features to back up system and application states along with data. During a system backup, all critical data reflecting the state of the OS and the applications is remotely backed up to external media such as tape or low-cost disk storage. For example, the OS-critical disks (that is, those containing the boot and system volumes) are backed up.

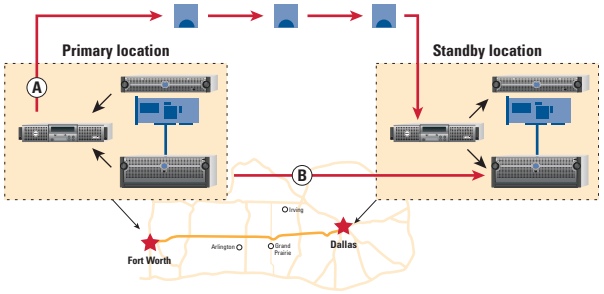
The recovery process usually involves reinstalling and reloading system data from backup media—whether locally restored or remotely restored—resulting in a fully functional replacement system. In these implementations, the target system hardware should be identical to the original hardware.

Remote system backup solutions typically incur some downtime for data and application recovery in the event of system failure. To minimize recovery time, certain recovery steps may be taken prior to a failure—for example, having already reinstalled the OS and applications. Certain manual steps still cannot be avoided, such as reloading the system image and restoring application data to the recovered system.



Dallas and Fort Worth, Texas, used as example sites
(50 km/35 mi separation)

Figure 2. Hardware-based remote data replication



Dallas and Fort Worth, Texas, used as example sites
(50 km/35 mi separation)

Figure 3. Hardware-based remote standby recovery system

Remote system backup implementations can range from low-end system-imaging software to high-end backup systems that offer customized application-specific features. Dell-supported solutions include applications such as EMC MirrorView, which can continuously perform synchronous real-time backups or scheduled backups run in asynchronous mode. Oracle® Data Guard performs similar functions for Oracle databases.

Remote standby system

Remote standby recovery solutions require an identical spare server housed in a secure off-site location that is configured with the same OS and applications as the production server (see Figure 3). If the primary system fails, the spare system assumes the functions of the primary system. To do so, the spare system must have access to the same application data that the primary system was using before the failure.

With some critical applications, the ability to reliably back up and recover business data alone is not sufficient because recovery from backup data can be time-consuming and involves manual repair or reconstruction of the failed system.

Remote standby solutions can be differentiated by the data recovery methods that enable the spare system to access the primary server's data in a consistent manner. One method is to back up the primary server's data at regular intervals to media such as off-site tapes, and then use the backed-up data to recover the spare system after a primary server failure. This process has a lower deployment cost than the geographically distributed configuration, but it is manual and involves restoring from tape, thereby necessitating some downtime. And as with local recovery using tape, the data may not be completely up-to-date.

A similar solution involves replicating the production storage and application configuration information at regular intervals to a remote location. In this implementation, a remote standby system already has the OS and applications installed. Once a failure is detected, the OS and application states are placed on the standby system.

This process requires manual reconfiguration of the spare server's interface to the external storage, such as the host bus adapter or the RAID controller. This method can help ensure that the spare system has access to the latest data—provided that the primary server flushed the data to a clean state before the failure.

Geoclusters present a small risk of producing inconsistent data. Most enterprises find this trade-off acceptable because geoclusters allow nearly instantaneous recovery.

Compared to remote system backup, remote standby solutions typically incur less downtime because there is no requirement for recovering the OS and applications. The downtime incurred with remote standby solutions depends on the data recovery approach used, as discussed in the "Site-wide data recovery" section in this article—more downtime is needed for remote backup than for remote replication. Remote standby solutions also incur additional costs for the hardware and the secondary location to enable recovery from a site-wide failure.

Geographically distributed cluster

In the local context, high-availability clustering involves two or more servers and a shared storage device. A geographically distributed cluster, or *geocluster*, comprises server nodes with independent replicated storage separated by a distance that exceeds the limitations of a shared-storage interconnect, such as a deployment that stretches across continents. The geocluster replicates real-time changes in data from one system's storage to the other system's storage. As is true with local high-availability clustering, all applications must be installed on each cluster node.

Because storage is mirrored rather than shared, geocluster installations offer more flexible configurations than local cluster implementations. A geocluster may employ replicated internal or external storage. In external storage configurations, logical disk volumes connected to the source server over Fibre Channel or SCSI are replicated over IP to the external storage connected to the target system.

Geoclusters present a small risk of producing inconsistent data. Most enterprises find this trade-off acceptable because geoclusters allow nearly instantaneous recovery. In case of a site-wide disaster, the downtime required to restore tape backups and the potential for multiple days of data loss can justify the cost and complexity of implementing geoclusters.

A balance between uptime needs and disaster recovery costs

Site-wide business continuity solutions can help recover business data alone or data in conjunction with business applications. Recovery time, solution complexity, maintenance, and costs vary widely, and choosing a business continuity implementation requires a clear understanding of these factors. Enterprises should define their requirements for business continuity clearly and then select appropriate products. Most enterprises will use several of the techniques discussed in this article to properly balance uptime needs against the costs of implementing a business continuity or disaster recovery plan.

Ananda Sankaran is a systems engineer in the High-Availability Cluster Development Group at Dell. His current interests related to high-availability clustering include storage systems, application performance, business continuity, and cluster management. Ananda has a master's degree in Computer Science from Texas A&M University.

Kevin Guinn is a systems engineer in the High-Availability Cluster Development Group at Dell. His current interests include storage management and business continuity. Kevin is a Microsoft Certified Systems Engineer (MCSE) and has a B.S. in Mechanical Engineering from The University of Texas at Austin.

Bharath Vasudevan currently manages the High-Availability Cluster Group at Dell. He has previously designed server hardware and served as a lead for multiple cluster releases. His current interests include application performance characterization and storage technologies. He has a master's degree in Electrical and Computer Engineering from Carnegie Mellon University.

FOR MORE INFORMATION

Dell business continuity solutions:
www.dell.com/businesscontinuity

Dell/SunGard Disaster Recovery Service:
www1.us.dell.com/content/topics/global.aspx/services/en/dell_sungard?c=us&cs=555&l=en&s=biz

Dell backup and recovery services:
www1.us.dell.com/content/topics/global.aspx/services/en/dps_bus_cont?c=us&cs=555&l=en&s=biz

Using Oracle Recovery Manager and Dell/EMC Storage

Oracle® Recovery Manager (RMAN) can be used to create backups and to restore and recover an Oracle database. With Dell/EMC storage, RMAN can provide a foundation for enterprise-class data protection and recovery.

BY TESFAMARIAM MICHAEL, MAHMOUD AHMADIAN, AND BOB NG

Related Categories:

Backup

Business continuity

Clustering

Database

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

Many IT departments have implemented user-managed backups for database as well as enterprise infrastructure software and data. During the user-managed backup process, database administrators (DBAs) use raw data backup tools to directly back up the disk content holding Oracle data. The disk content is first conditioned by putting the Oracle database (or tablespace files) into hot backup. Putting the database into hot backup causes the Oracle server to perform checkpoints on different tablespaces as part of the disk data conditioning. These checkpoints can affect the production performance level. Use of fast, storage-based, point-in-time logical replication helps to minimize the time required to copy the conditioned Oracle disk data. Integrating this type of replication into the user-managed backup process can help to minimize the duration of the backup and its impact on applications service.

Oracle® Database 10*g* includes Recovery Manager (RMAN), a utility for creating backups and for restoring and recovering an Oracle database. RMAN works directly through the database management system's engine to secure the appropriate data for use in its backup process. As such, RMAN does not require the preconditioning of the Oracle data stored on disk through mechanisms such as hot backup. However, RMAN competes for system resources with production user activities, which can impact production performance for the entire duration that RMAN needs to complete the backup task.

The combination of the two approaches, user-managed backup and RMAN, offers an option that can help minimize the service-level impact as well as the duration of the disruption. Starting with user-managed backup of the Oracle database that is facilitated by storage-based, logical, near-instantaneous replication, DBAs can minimize the length of time during which the production service level is disturbed. The captured disk copy can then be mounted onto a separate set of resources, such as backup servers, where the RMAN task can be run in the background. This technique enables DBAs to benefit from an RMAN backup with no further impact on the production service level. It can be an effective way to leverage additional hardware and software resources to enhance the overall quality of service of the production environment.

Offering enhanced functionality

Given the importance of backup, Oracle has recently enhanced RMAN in Oracle Database 10*g*. By design, RMAN works intimately with the database server processes. As a result, RMAN enables many features that may not be present in user-managed backup alternatives. These features include block-level detection of data corruption during backup and restore; optimized backup using

process parallelism; reduced space consumption using compression and block change tracking; a common user interface across all supported operating systems; a catalog containing detailed information of backup history; compatibility with most of the leading backup software applications through RMAN's Media Management Library application programming interface (API); incremental backup; and image copy backup.

Comparing online and RMAN backups

Online backups, also referred to as hot backups, are performed while the database is mounted and open. Hot backups do not require database downtime and allow higher availability than offline, or cold, backups. Oracle best practices recommend using RMAN for all Oracle database backup and recovery operations.

For hot backups, redo logs must be archived regardless of whether RMAN is used. Recovering a database from a hot backup without archived redo logs is not possible. Oracle requires a minimum of two redo log files. However, the best practice is to configure several redo logs. Redo logs also may be multiplexed to enhance redundancy. These files switch roles from online redo log to offline redo log. While the online redo log is being updated by Oracle log writer, the offline redo log is archived by the log archive process. The role switch happens when the online redo log is full, after verifying that the offline redo log is archived.

If RMAN is not used, the database must be put in hot-backup mode. The System Change Number in the data file headers ceases to increment with checkpoints, and the full image of the changed database blocks are written to redo logs. This results in "extra work" if extensive changes are made to the database during a hot backup. Once the backup is complete, the database is taken out of hot-backup mode and reverts to normal logging.

Advantages of RMAN backups. The advantage of RMAN backups is that the database does not need to be put into hot-backup mode because the backup will be performed by Oracle processes that employ the read-consistency mechanism used by SQL statements. Another advantage is ease of backup management. When using RMAN to perform backups, RMAN can be directed to perform validation of backup data created. RMAN also maintains information about the created backup set in a catalog. This catalog information can greatly facilitate the ability to perform subsequent database restores using the most optimal backup data set alternatives.

However, running an RMAN backup directly against the production database can have drawbacks. Because the RMAN session runs concurrently with other production user requests, it competes for service from the database engine against those user processes.

Using storage-based replication to support Oracle backups

To leverage storage and database features for better performance, DBAs can use the storage software's fast replication features to create

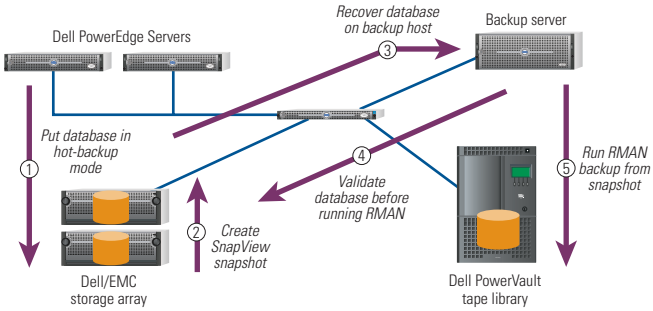


Figure 1. Using SnapView snapshots to offload RMAN backups

a logical or physical copy of the database. Rather than running RMAN on the production database, in this scenario the DBA puts the Oracle database into hot-backup mode. The storage system's near-instantaneous replication technology provides the ability to create a second copy of the database. The production database is therefore required to be kept in the hot-backup state just long enough to allow the storage replica to be made, which typically takes just a few seconds. Once the copy is available and accessible, it is mounted onto a separate server dedicated to the task of performing the RMAN backup. The logical copy, with the archived logs, is restored and the database is recovered onto the backup server node using RMAN. The recovered database can now be used as the target for RMAN to create the final backup to a Dell PowerVault™ tape library. The actual process of creating the RMAN backup may take substantially longer (minutes or hours), but that activity occurs without affecting the service of the production database for the entire backup duration.

Dell/EMC CX series storage arrays are available with the optional EMC® SnapView™ point-in-time copy software, which DBAs can use to create physical or logical copies of an Oracle database.

Physical point-in-time clones. Using SnapView clones for backups alleviates the overhead associated with database backups from the production volume. A clone is an actual copy of a logical unit (LUN) or volume, and it is created by mirroring (copying) the changes from the source (production volume) to the clone—a process referred to as cloned mirror write (CMW). As of the 3Q05 Dell/EMC CX series array release, up to eight clones can be assigned to one production volume.

DBAs can use the Oracle database residing on the clone as the target for RMAN. In this case, the database is switched to hot-backup mode so that the database will be in a consistent state when the clone is fractured. Once the clone is fractured, the database is returned to normal operation mode. The fractured clone is then used by RMAN to create a backup of the database onto backup media. This method offloads all the backup reads from the production volume, assuming that the SnapView clone is placed on spindles other than the ones they mirror, which is the recommended practice. Additionally, the load of CMW ceases as soon as the clone is fractured for backup. The same methodology applies even if the Oracle database resides on multiple LUNs. As long as the Oracle database stays in the hot-backup state, while all the clones associated with each of the production database LUNs are being fractured, the clone set is a validated database LUN set that can be used to run the RMAN backup task.

The advantage of this method is that the source volume is not touched by the backup processes because the point-in-time copy of the source resides in the clone. However, with this technique the fractured clone(s) must be resynchronized with the production volume once the backup is complete. Because up to eight clones can be supported, users who need an up-to-date clone can be accommodated by keeping at least one of the clones in the group synchronized with the production volume.

Virtual point-in-time snapshots. Another method is to use SnapView snapshots to make a virtual point-in-time copy of the LUN, which is designed to complete in a matter of seconds. First the database is put in hot-backup mode. Then the SnapView software is used to create a snapshot of the LUN. The database is set back to normal mode and the snapshot is backed up. The advantage of this method is that the LUN is not fractured and no resynchronization to the production volume is required. However, because the snapshot is the point-in-time logical replica of the source LUN, the backup processing will affect the source LUN write I/O performance while the snapshot is used to drive the RMAN backup (see Figure 1).

Providing effective support for Oracle backups

The Dell/EMC CX300, CX500, and CX700 storage arrays are designed to provide reliable and scalable storage for Oracle databases. Storage instance replication features of these arrays, such as EMC SnapView software, can be valuable resources for Oracle DBAs. Leveraging the array-based software to support Oracle backups can help to optimize the service levels of the production environment during backup operations. Both SnapView clones and snapshots are options on the CX arrays that can enable DBAs to back up an Oracle database, incurring low performance impact during the backup process. Using RMAN in conjunction with SnapView can provide an effective method of storage and server offloading for Oracle backups.

Tesfamariam Michael is a software engineer in the Dell Database and Application Engineering Department of the Dell Product Group. Tesfamariam has an M.S. in Computer Science and a B.S. in Mathematics from Clark Atlanta University, and a B.S. in Electrical Engineering from Georgia Institute of Technology.

Mahmoud Ahmadian is an engineering consultant in the Database and Applications team of the Dell Product Group. Mahmoud has an M.S. in Computer Science from The University of Houston, Clear Lake.

Bob Ng is an engineering consultant in the EMC® CLARiiON® Application Solution Integration team at EMC. Bob has an M.S. in Electrical Engineering and Computer Science from the University of California, Berkeley.

Enhancing Windows Data Recovery

with Symantec Backup Exec Continuous Protection Server 10d

Symantec® Backup Exec™ 10d *for Windows Servers*—incorporating the Backup Exec Continuous Protection Server 10d—offers controlled, rapid backup and recovery of critical enterprise files through continuous disk-based data protection.

BY RICHARD GOODWIN AND KYON HOLMAN

Related Categories:

Backup

Continuous data protection (CDP)

Disaster recovery

Storage

Symantec

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

Fighting the explosive growth of data across the Windows enterprise can be a difficult battle. Backup windows are often shrinking at the same time that the demands enterprises place on protection and recoverability of their data are expanding. Although still a critical component of a business continuance strategy, traditional tape-based backup configurations are typically slower and can be less reliable than disk-based methods—including continuous disk-based data protection.

Symantec Backup Exec 10d *for Windows Servers*—with its implementation of a continuous disk-based data protection feature known as the Backup Exec Continuous Protection Server 10d—can enhance reliability, provide fast backup and recovery, and help eliminate problems with backup windows altogether. Most important, this software can enable end-user restores without IT administrator intervention.

Traditional tape backup processes

To understand the power of continuous disk-based data protection, administrators first need to understand the way today's traditional backups generally take place. Most environments employ tape for short- and long-term archiving, and many have begun to use limited disk resources as well. Data storage in these scenarios typically follows the normal data-archiving sequence:

1. Storing a full backup of the system data on tape or disk
2. Storing incremental or differential backups to tape or disk

These data-archiving processes are staggered across calendar dates and times, or they follow certain usage and retention models, such as grandfather, father, son rotation. These traditional storage methods can protect system data, but with the advent of inexpensive, readily available disk resources, they may not be the best-suited storage method for many organizations.

Besides having to manage the physical resources of tape, administrators have to manipulate rotation and backup schemes to help ensure that they have sufficient data protection for a given period. When a large or full system restore is needed, administrators must first locate the correct full backup and then locate any associated incremental or differential backups before they can complete the operation. This process can be lengthy and error-prone.

Tools that have become available recently, such as the Synthetic Backup feature, available since Backup Exec 10.0, can help alleviate some of these problems. Synthetic Backup works by allowing administrators to create a new full backup from previously gathered full and incremental

backups. This provides a seamless restore for a given point in time to facilitate restore operations. However, such tools typically work to overcome the traditional limitations of tape and often use disk storage to mimic tape's behavior. Thus, they may not fit the usage paradigm of a given environment.

Advantages of continuous data protection

In most data centers, backups are managed by IT administrators, and the backup application is needed for identifying and restoring files. Typically, the backup and restore process is inaccessible to an enterprise's end users. Continuous data protection harnesses the disk resources available in today's IT environments and provides power and flexibility not only to IT administrators but also to an enterprise's end users—enhancing the processes for protecting critical business information.

As discussed earlier in this article, traditional backup schemes generally center on the requirements of the tape devices associated with them. This means that a protection scheme does not always map to the activity of the live business data. With the advent of disk-based backups, administrators do not have to wait to protect data only during backup cycles—it can be protected continuously. The Backup Exec Continuous Protection Server 10d is specifically designed for disk-based backups, enabling a complete disk-to-disk-to-tape protection strategy.

Real-time file protection

The Continuous Protection Server protects files in real time—continuously. Whenever a file is created or changed, it is protected immediately. By doing so, administrators can eliminate the full, incremental, and differential backups inherent in tape-based backups.

The Continuous Protection Server works by installing the Continuous Protection Agent (CPA) on one or more file servers and directing them to a Continuous Protection Server (CPS). The CPS can be installed on an existing Backup Exec media server or on a separate server, allowing for maximum flexibility. The CPA continuously tracks and records changes to files as they are created or changed. When a file is created or a change occurs, the CPA immediately transmits only the changed data in the file to the Continuous Protection Server. By pairing the Continuous Protection Server with the internal storage on a Dell™ PowerEdge™ server or an external storage device such as the Dell PowerVault™ 220S enclosure, administrators can help ensure that files are protected continuously. Backup windows—the limited period during which backup operations run—can be eliminated.

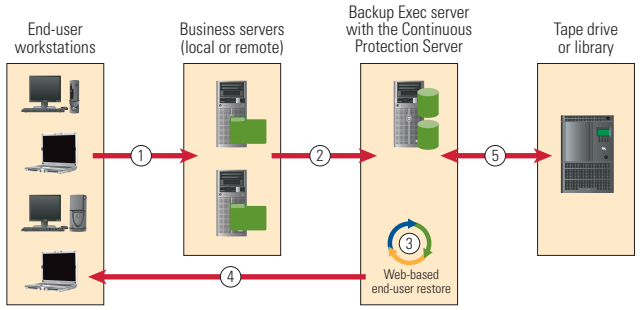


Figure 1. Backup and restoration of data using the Backup Exec Continuous Protection Server 10d

Point-in-time snapshots

In addition to always maintaining the current version of protected data, the Continuous Protection Server offers the ability to take point-in-time snapshots. By implementing a snapshot schedule—for example, one snapshot taken every hour during business hours—the Continuous Protection Server can retain multiple versions of data as it existed at that particular point in time. If a file is overwritten or corrupted at 5 P.M., an administrator can recover that file as it existed when any of the day's snapshots were created.

SmartLink technology

The Continuous Protection Server is integrated with the Backup Exec product using Symantec Backup Exec SmartLink technology, letting administrators track the status of traditional and continuous data-protection jobs. Of course, tape is still an integral part of an overall data protection strategy, and the integration of the Continuous Protection Server within Backup Exec lets administrators back up important data to tape for long-term storage or off-site disaster recovery.

Figure 1 illustrates how the Continuous Protection Server functions within a Symantec Backup Exec 10d *for Windows Servers* environment. As shown in Figure 1, the backup and restore process incorporates the following steps:

- Users save files to business file servers.
- The CPA streams file changes to the Continuous Protection Server.
- Microsoft® Volume Shadow Copy Service (VSS) snapshots provide versioning and granular point-in-time recovery of files. The Continuous Protection Server offers smart management of VSS snapshots, allowing administrators maximum flexibility in their data protection.
- Users retrieve previous versions of files using simple, Web-based recovery.
- The Backup Exec media server provides backup to tape for long-term retention or disaster recovery.

End-user restores

The Backup Exec Continuous Protection Server can help IT administrative staff ease the burden of traditional backup and restore

operations. However, administrators also often receive requests from end users to restore files that have been lost, corrupted, or deleted. To fulfill such requests, administrators are diverted from their primary responsibilities and must perform the following steps:

1. Locate the backup session in question on the media that was employed.
2. Mount the media.
3. Locate the file(s) in question.
4. Restore the file(s) to the end user's desktop or file share.
5. Verify with the end user that the correct file or version was restored.

This is a long process that can often take hours or days, depending on the situation. It also can be slow and inefficient and may only result in end users having to re-create their work. With Backup Exec Continuous Protection Server, end users can locate and restore their own data directly from an easy-to-use Web-based interface called Backup Exec Retrieve (see Figure 2).

Simple and safe. Backup Exec Retrieve is a simple, safe method to help end users restore their own files. It uses a standard Web browser; no software needs to be installed on end-user workstations. End users are allowed only to browse and recover files, and therefore cannot alter backup data or disrupt protection operations—and they cannot delete a file. Backup Exec Retrieve leverages standard Microsoft Windows® authentication, so users have access only to the files they created or for which they have permissions.

Search options. After bringing up the Backup Exec Retrieve Web page, end users are given various familiar options to locate and restore their data. A standard search view is available for locating files with certain file-name characteristics. Files are indexed by title or file type. A history view allows end users to browse by periods of recent activity. A more advanced browse view lets end users manually browse the data that has been protected.

Retrieval procedures. Files are displayed in the Web interface as standard Internet links. End users simply click on the link,

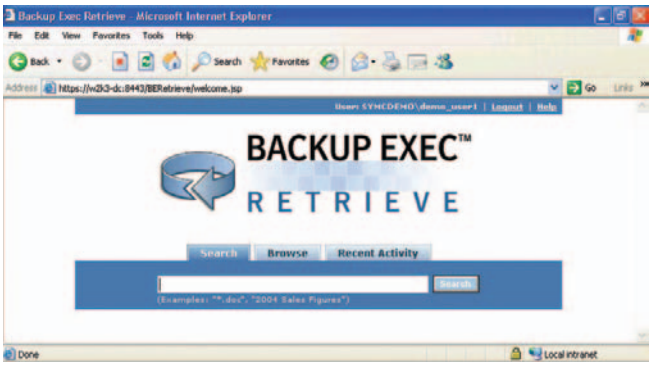


Figure 2. Backup Exec Retrieve interface screen

Retrieval using traditional tape backup	Symantec Backup Exec Retrieve
1. Administrator locates the backup session in question, probably stored on a tape or set of tapes.	1. End user searches for files.
2. Administrator mounts the tape(s).	2. End user clicks on Web link to download the files.
3. Administrator locates the files in question.	
4. Administrator restores the files to the end user's desktop or file share.	
5. Administrator verifies with the end user that the correct files or version was restored.	

Figure 3. Data retrieval times and procedures: traditional tape backup versus Symantec Backup Exec Retrieve

and that action generates a prompt to download the file. This familiar, streamlined process requires no additional training and no additional software to be deployed to the end user's system. Symantec Backup Exec Retrieve is supported on Microsoft Internet Explorer 6.0 and later.

Figure 3 summarizes the differences in time and procedures between traditional backup-from-tape methods and the Symantec Backup Exec Retrieve approach.

Continuous data protection for the enterprise

Continued data growth poses a problem for most enterprises. Continuous data protection provides a ready answer by helping to eliminate backup windows and improve reliability. With Symantec Backup Exec 10d *for Windows Servers,* Microsoft Windows–based data centers can leverage both continuous data protection and traditional tape-based technologies in one powerful configuration. Continuous disk-based data protection enabled by Backup Exec Continuous Protection Server 10d eliminates backup windows while simultaneously improving reliability. Best of all, though, it allows end users to retrieve their own data, eliminating the burden placed on IT staff and promoting efficiency in day-to-day operations. In most cases, end users can locate and retrieve their own files in less time than it takes the help desk to pick up the phone, meaning IT can improve its service levels without increasing administrative costs or adding resources.

Richard Goodwin is a technical product manager for the Backup Exec series of products at Symantec, and is responsible for incorporating customer and partner requirements into current and future releases. His experience in the storage industry spans almost a decade with Dell, VERITAS, and Symantec.

Kyon Holman is a lead software engineer on the Tape Storage team in the Dell Enterprise Product Group. He has a B.S. in Computer Science from the University of Michigan at Ann Arbor, an M.S. in Software Engineering from The University of Texas at Austin, and is currently pursuing an Executive M.B.A. from The University of Texas at Austin.

Dell and SunGard:

Disaster Recovery Made Simple

Natural disasters and equipment failure can place mission-critical data and systems at risk. To provide users of Dell™ PowerEdge™ servers with a simple process for disaster recovery, Dell and SunGard have created the Dell/SunGard Disaster Recovery Service. This article describes how this service can help with processing recovery.

BY ED LAWRENCE

Related Categories:

Business continuity

Data center technology

Dell Services

Disaster recovery

Planning

SunGard

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

An effective disaster recovery plan has two important components: data protection and processing recovery. Simply put, administrators must perform regular backups of critical data, store those backups at an alternate site to help ensure that the data survives a disaster, and then be able to quickly restore the data after a disaster.

Many enterprises focus on data protection. They create regular backup tapes and store those tapes off-site. Few small- or medium-size enterprises attend to the equally important element of processing recovery, which involves pre-arranging IT resources to restore processing and user access to data quickly after a disruptive event.

Disasters can compromise not only IT equipment but also the data processing facility. Besides replacing servers, small- and medium-size enterprises may find themselves needing an alternate facility where they can recover the data center.

Pre-arranging servers and facilities can be prohibitively expensive when performed internally or through a hot-site vendor. Recognizing this, Dell teamed up with SunGard Availability Services—one of the world's largest providers of Information Availability services—to provide the Dell/SunGard Disaster Recovery Service, an effective and economical option for small- and medium-size enterprises.

With the Dell/SunGard Disaster Recovery Service, enterprises can simply add coverage for their Dell PowerEdge servers. SunGard supports this coverage by quickly providing temporary recovery servers that are configured as similarly as possible to the failed servers. In addition, SunGard maintains recovery facilities in most major metropolitan areas throughout the United States to serve as temporary office locations for displaced workers—complete with desktop PCs, voice phones, and Internet access. The replacement server and facility access are both bundled into the add-on service.

The Dell/SunGard Disaster Recovery Service can provide rapid, cost-effective processing recovery, complete with alternate office space. In addition, enterprises can easily subscribe to this service on a server-by-server basis.

Ed Lawrence is a senior director of market development for SunGard Availability Services. He has worked at SunGard for 19 years in various positions, all focused on providing effective and economical disaster recovery and business continuity solutions. Ed has a bachelor's degree and a master's degree from The Pennsylvania State University and is a Certified Business Continuity Professional.

FOR MORE INFORMATION

Dell/SunGard Disaster Recovery Service:
www.dell.com/sungard

YOUR JOB IS TO KEEP SYSTEMS AND APPLICATIONS RUNNING. OUR MISSION IS TO KEEP PEOPLE AND INFORMATION CONNECTED. LET'S WORK TOGETHER.

Continuous access to information no matter what. That's Information Availability. It's what your employees, suppliers and customers demand every minute of every day. But to deliver it flawlessly, you need a massive global infrastructure, redundant systems and diverse networks being monitored and supported by skilled technical experts at secure facilities. That's exactly what SunGard provides.

As a result, we can offer you a higher level of availability and save your company, on average, 25%* versus building the infrastructure yourself. Plus, it's a vendor neutral solution that lets you control your data, applications and network while giving you the flexibility to adjust to the changing needs of your business. But best of all, it lets you spend more time solving business problems and less time solving technical problems.

For years, companies around the world have turned to SunGard to restore their systems when something went wrong. So, it's not surprising that they're now turning to us to mitigate risk and make sure they never go down in the first place.

You want your network and systems to always be up and running. We want the same thing. Let's get together. To learn more, contact us at 1-800-WWW-DELL or go to www.dell.com/sungard.

SUNGARD® Availability Services | Keeping **People** and **Information** *Connected.*™

*Potential savings based on IDC white paper sponsored by SunGard: "Ensuring Information Availability: Aligning Customer Needs with an Optimal Investment Strategy." September 2004.

Exploring High-Availability Features in Microsoft SQL Server 2005

Microsoft® SQL Server™ 2005 offers failover clustering, database mirroring, log shipping, replication, and database snapshot features that can help safeguard an enterprise's critical data and operations. This article discusses these features and their associated trade-offs.

BY ANANDA SANKARAN, DAT NGUYEN, AND NAM NGUYEN

Related Categories:

Clustering

Database

High availability (HA)

Microsoft SQL Server 2005

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

Enterprise data centers that host business-critical applications and databases can be at risk from failures and disasters. Availability and recoverability become crucial for business operations. Current database availability technologies vary in their complexity, cost, level of automation, incurred downtime, and supported distance. A major challenge for an enterprise is determining the most cost-effective and least complex solution to help ensure database availability and meet business needs. Microsoft SQL Server 2005 introduces several high-availability features and feature enhancements for the SQL Server database platform.

Failover clustering

Failover clustering with Microsoft Cluster Service (MSCS) is a popular way to achieve high availability for SQL Server. With MSCS, multiple servers or nodes are linked to function as a single system and provide an automatic failover solution. Enterprises are connected to virtual servers and not physical servers; each active physical server may host one or more virtual servers.

If one of the physical servers fails because of either a hardware or software problem, MSCS detects the failure and moves resources that reside on the failing server—including the virtual server(s)—to one or more remaining physical servers. End users connected to the failed server observe only a momentary delay in accessing its resources while MSCS restarts SQL Server services on the remaining server(s) and remaps the virtual server connections.

Installing a failover cluster requires the following:

- Microsoft Windows® 2000 Server (Advanced Server or Datacenter Server) or Microsoft Windows Server™ 2003 (Enterprise Edition, Datacenter Edition, Enterprise x64 Edition, or Datacenter x64 Edition)
- Microsoft SQL Server 2005 (Standard Edition or Enterprise Edition)
- Shared storage based on SCSI, Fibre Channel, or Serial Attached SCSI
- All components (server, host bus adapter, and storage system) and the cluster solution listed in the Microsoft Windows Server Catalog

Figure 1 shows two typical two-node failover cluster configurations for SQL Server 2005, one in active/passive

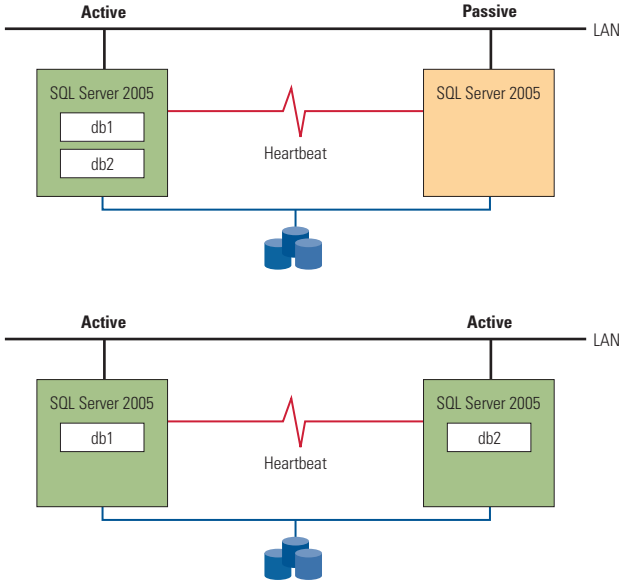


Figure 1. Failover cluster configurations for SQL Server 2005

mode and one in active/active mode. In a failover cluster environment, SQL Server 2005 offers several additional features and improvements compared to SQL Server 2000:

- **Support for multinode clustering:** Standard Edition allows up to two-node clustering and Enterprise Edition up to eight-node clustering.
- **Support for SQL services:** Analysis Services feature offers multi-instance support and full-text indexing.
- **Easier installation:** System Configuration Check automatically detects and installs SQL Server as virtual servers across the nodes in the cluster.

Database mirroring

Database mirroring is designed to augment the availability of SQL Server 2005. With this feature, transactions to a database hosted on a SQL Server instance, referred to as the principal, can be continuously duplicated in real time onto a copy hosted on another instance, referred to as the mirror. The principal and mirror servers are considered partners in a database mirroring session. Transactions (such as insertions, updates, and deletions) on the principal database are duplicated by continuously sending transaction log records to the mirror over the network (see Figure 2).

The mirror receives the log records continuously and restores them. If the principal or mirror server incurs a failure, the mirroring session is disconnected. Before starting the mirroring session, administrators must initialize the mirror database from a principal database full-restore with the NORECOVERY option.

Besides the primary and mirror, a mirroring session may include an optional third server, referred to as the witness. The witness server enables automatic failover after a principal server failure by promoting the mirror to function as the principal. Using a witness server helps achieve a quorum and prevents accidental promotion of the mirror database that results from communication failures between the principal and mirror.

Database mirroring requires the following configuration:

- The primary and mirror databases must be hosted on separate SQL Server 2005 instances.
- The principal database must be set to the Full Recovery model, not the Bulk-Logged Recovery or Simple Recovery model.

Database mirroring can be configured in any of the following three operating modes based on transaction safety level and the presence of a witness server:

- **High-availability mode:** In this mode, log records are transferred synchronously from the principal database to the mirror. The principal waits for an acknowledgment from the mirror before "hardening" log records to its disk. In addition to providing synchronous transfer, this mode uses a witness server to enable automatic failover.
- **High-protection mode:** This mode is similar to the high-availability mode except that it does not support a witness server, and thus automatic failover is not possible. However, failover can be performed manually to promote the mirror database. Because the log records are transferred synchronously in this mode, the mirror database is synchronous and consistent with the principal.
- **High-performance mode:** In this mode, log records are transferred asynchronously from the principal to the mirror database. The principal does not wait for an acknowledgment from the mirror before "hardening" its logs to the disk,

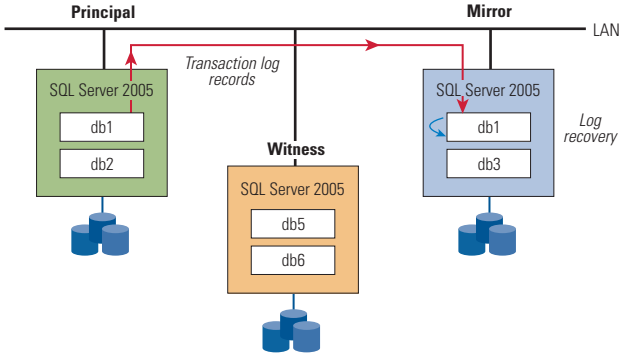


Figure 2. Database mirroring for database db1 in high-availability mode

and the mirror is not guaranteed to be synchronous with the principal at any point. This mode does not support a witness server, and thus neither automatic nor manual failover is possible. However, a forced failover can be performed to recover the mirror, with some data loss occurring because mirroring is asynchronous.

A key advantage of database mirroring is that it provides quick failover and high availability with minimal hardware cost and complexity. Formerly, such SQL Server mirroring solutions were often feasible only with expensive, proprietary hardware-based mechanisms. With acceptable latencies and bandwidth, mirroring can be deployed over a long-distance wide area network (WAN) as part of a disaster recovery solution.

SQL Server 2005 provides three types of replication technologies suitable for different needs: transactional, merge, and snapshot. SQL Server 2005 adds enhancements to these replication methods to help improve scalability, performance, and monitoring capabilities.

During the mirroring session, the mirror database is not available for client access because it is in a recovery state, but a database snapshot can be created on the mirror for limited reporting. Only one mirror database can be established for a principal database. However, the participating SQL Server instances can assume different roles for different mirroring sessions. For example, a SQL Server instance can serve as a principal for one database, as a mirror for another database, and as a witness for yet another mirroring session. The notion of a virtual server, as is used in failover clustering, does not exist. However, Microsoft ADO.NET provides an application programming interface (API) for clients to automatically reconnect to the mirror server during failures.

Log shipping

Log shipping is similar to database mirroring. It allows a database hosted on a primary server to automatically send transaction logs to a secondary database for duplication. The log records are not transferred continuously as in database mirroring, but rather in intervals.

Log shipping comprises backing up the transaction logs on the primary server instance, copying the backed-up logs to the secondary server instance, and restoring the copied logs on the secondary server instance periodically. The primary database should be configured for the Full Recovery or Bulk-Logged Recovery model. The secondary database should be initialized by a full recovery of the primary database with the NORECOVERY or STANDBY option.

Log shipping can optionally include a third server named "monitor" to record the history of log-shipping operations and to raise alerts during failures. Log-shipping operations are carried out by four Microsoft SQL Server agent jobs: backup job, copy job, restore job, and alert job.

During log shipping, the secondary database is not completely synchronized with the primary database. Also, log shipping does not provide automatic failover to the secondary server during a failure. Log shipping supports multiple secondary servers for a primary server. In addition, a SQL Server instance can function as a secondary or monitor server for multiple primary servers. The secondary server can be used directly for limited query processing. Log shipping can be used in scenarios where multiple destination servers are needed and where delays in restoring logs on the duplicate are acceptable.

Replication

Replication is a set of technologies for distributing data and database objects across Microsoft SQL Server databases over a network. Replication involves a publisher database instance that makes data available for copying, a distributor instance for copying data and maintaining metadata, and subscriber instances that can receive replicated data.

SQL Server 2005 provides three types of replication technologies suitable for different needs: transactional, merge, and snapshot. SQL Server 2005 adds enhancements to these replication methods to help improve scalability, performance, and monitoring capabilities. Some of the improvements include additional replication data types, support for partitioned tables and full-text index, direct Data Definition Language (DDL) replication for schema changes, and replication from Oracle® databases.

In transactional replication, SQL Server 2005 adds an important feature called peer-to-peer replication, which enhances the existing bidirectional replication option. Advantages of replicated databases include data load balancing and disaster recovery. With the peer-to-peer replication feature, administrators can set up multiple peer-to-peer transactional replication pairs among different data centers. That way, if one data center is down, it can be failed over to its peer center(s). Although there is no limit to the number of nodes in a peer-to-peer topology, manageability of the peer-to-peer relationships between nodes can become a constraint.

Database snapshots

Administrative and application errors can be a major cause of database downtime. SQL Server 2005 introduces a feature called Database Snapshots to help protect a database from erroneous user operations and to augment database backup and restore operations. The Database Snapshots feature allows recovery from user errors by allowing the database state to revert to a point in time before the error(s) occurred. It works in stand-alone, failover clustering, and database mirroring environments. However, it does not work with a log-shipping secondary database.

	Granularity	Data loss	Automatic failover	Downtime	Solution-specific certification requirements
Failover clustering	System and database	No	Yes	Approximately 20 seconds plus database recovery time	Cluster solution listed in Microsoft Windows Server Catalog
Database mirroring	Database	Depends on operating mode	Yes	Less than 3 seconds	No
Peer-to-peer replication	Table or view	Some	Optional	Variable (little to none)	No
Log shipping	Database	Some	No	Variable	No
Database snapshots	Table or view	Some	No	Variable	No

Figure 4. High-availability options and features in SQL Server 2005

A database snapshot is a snapshot of an entire database at a given time. The snapshot is created instantly and can be used for read-only tasks such as reporting. A snapshot must have been created before the error occurred for an administrator to restore the database to a consistent state. Snapshot creation does not impose restrictions on the base database operations. Administrators can create multiple snapshots from a database as well.

The SQL Server 2005 Database Snapshot feature employs a copy-on-write mechanism. Only changes to the base database—not the entire base database—are recorded after the snapshot has been taken. The snapshot is a pointer image to the base database. It shares the unchanged pages of the base database and requires only extra storage for changed pages. The Database Snapshot mechanism is illustrated in Figure 3.

The Database Snapshot feature is designed to be extremely space-efficient. Any modification such as a write (step 1 in Figure 3) to the base database after a snapshot is taken will be recorded (step 2 in Figure 3) as part of the target snapshot. If I/O read requests (step 3 in Figure 3) are to blocks that have not changed since the snapshot was created, then the request will read from the base database. If I/O requests are to blocks that have changed since snapshot creation, then the request will be read from the snapshot.

Figure 4 summarizes the high-availability features and options available in Microsoft SQL Server 2005.

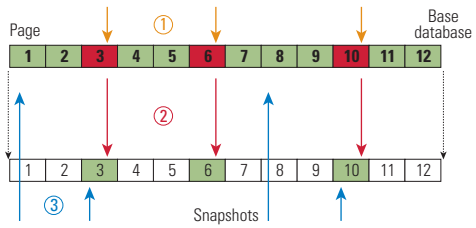


Figure 3. Database Snapshot mechanism in SQL Server 2005

A high-availability database platform

Microsoft SQL Server 2005 offers several features that are designed to help augment availability of data. Dell™ PowerEdge™ servers, Dell PowerVault™ storage, and Dell/EMC storage are also designed with a broad array of redundant hardware and software features to maximize hardware availability. The features in Dell hardware and SQL Server 2005 provide an integrated, industry-standard approach to high database availability. A combination of these components can greatly benefit enterprises with high-availability database requirements.

Ananda Sankaran is a systems engineer in the High-Availability Cluster Development Group at Dell. His current interests related to high-availability clustering include storage systems, application performance, business continuity, and cluster management. Ananda has a master's degree in Computer Science from Texas A&M University.

Dat Nguyen is a systems engineer in the High-Availability Cluster Development Group at Dell. His responsibilities include developing storage area network (SAN)–based high-availability clustering products. His current interests are in enterprise storage products and technologies. Dat has a B.S. in Electrical Engineering from the University of Houston.

Nam Nguyen is a senior consultant in the High-Availability Cluster Development Group at Dell, and the lead engineer for Dell Fibre Channel PowerEdge Cluster products. His current interests include business continuity, clustering, and storage technologies. He has a B.S. and an M.S. in Electrical Engineering from The University of Texas at Austin.

FOR MORE INFORMATION

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

★★★★

PERFORMANCE. HASSLE-FREE MULTI-CORE SERVERS.

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

DELL'S EASY TO DEPLOY MULTI-CORE TECHNOLOGY.

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

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

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

SQL Server 2005:

Preparing for a Smooth Upgrade

To meet a new generation of data-management needs, Microsoft® SQL Server™ 2005 has been reworked extensively to enhance performance and application programmability. Originally published by *SQL Server Magazine* as part of its "SQL Server 2005 Upgrade Handbook," this article explores how administrators can help ensure a successful transition to SQL Server 2005 by planning, testing, and using the Upgrade Advisor.

BY ERIK VEERMAN

Related Categories:

Data Transformation Services (DTS)

Database administration

Database

Microsoft SQL Server 2005

SQL Server 2005 Integration Services (SSIS)

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

Each Microsoft SQL Server 2005 component is designed to have a unique architecture and life cycle—the two primary areas that can affect an upgrade path. Some SQL Server 2005 components build on a solid foundation to augment, optimize, and help stabilize existing functionality. Microsoft has performed an extensive reworking of other SQL Server features to enhance performance and application programmability. SQL Server 2005 also incorporates completely overhauled components and additions designed to meet a new generation of data-management needs.

Preparing for a SQL Server 2005 upgrade involves understanding some basic principles that enable administrators to make appropriate decisions and help ensure success. As with any upgrade, the keys to success are appropriate planning and testing for the needs of the specific environment. This article explores the overall upgrade path for SQL Server 2005 components and how the SQL Server 2005 Upgrade Advisor tool can help identify areas that require special attention. Specific upgrade considerations for certain SQL Server 2005 components—the database engine, Integration Services, Analysis Services, and Reporting Services—are also examined.

Upgrade mechanism

For all components, SQL Server 2005 provides an upgrade from SQL Server 2000 or SQL Server 7.0. Note that Microsoft distinguishes between a SQL Server 2005 upgrade and a migration.

An upgrade is an automated process in which the upgrade tool, called Setup, moves an old instance of SQL Server to a new instance while maintaining the data and metadata of the old instance. At the end of

SQL Server 2005	SQL Server 2000 or SQL Server 7.0
Database engine	*Upgrade tool:* Setup *Migration method:* Administrators perform side-by-side installation and then database backup/restore or detach/attach
Analysis Services	*Upgrade tool:* Setup *Migration tool:* Migration Wizard *Migration method:* Migration Wizard migrates objects, but optimization and client-access upgrades are required
Integration Services	*Upgrade tool:* None *Migration tool:* DTS Migration Wizard *Migration method:* DTS Migration Wizard converts 50 to 70 percent of the tasks, but some manual migration is required; runtime DTS DLLs are available in SSIS; package re-architecture is recommended
Reporting Services	*Upgrade tool:* Setup *Migration method:* Administrators perform side-by-side installation, and reports are deployed on the new instance
Notification Services	*Upgrade tool:* None *Migration method:* Upgrade of Notification Services instances occurs during installation

Figure 1. Upgrade path for Microsoft SQL Server 2005 components

the upgrade, the old instance is no longer available and the new instance has the same name as the old instance. Alternatively, migration is a manual process in which the database administrator installs a new instance of SQL Server and copies the metadata and data from an old instance of SQL Server to the new instance. Migration provides access to two instances of the system, letting administrators verify and compare the two systems. During migration, both the old and new systems remain online until migration to the new instance is complete. At the end of the migration, all applications are directed to access the new instance and the old instance is manually removed.

Although the SQL Server 2005 database engine introduces many features, administrators can easily upgrade databases on SQL Server 2000 or SQL Server 7.0 to SQL Server 2005 by using the Setup wizard or by performing a database restore or attach/reattach. Moving from Data Transformation Services (DTS) to SQL Server 2005 Integration Services (SSIS), however, requires a migration assisted by an out-of-the-box migration tool to help move data processing to the SQL Server 2005 architecture. Figure 1 summarizes the upgrade path for each SQL Server component.

Using the compiled knowledge from its product team, internal lab testing, and extensive SQL Server 2005 early-adopter experience, Microsoft has developed an essential tool for upgrade preparation called the Upgrade Advisor. Figure 2 shows the Welcome screen for the Upgrade Advisor, which analyzes the configuration of the existing database server, services, and applications and provides reports that identify changes within the SQL Server product that can affect the upgrade. These changes include security enhancements, closer adherence to the SQL standard compared to previous SQL Server versions, and architectural changes. The Upgrade Advisor also provides links to documentation that describe these changes and necessary steps to complete the upgrade process. The Upgrade Advisor can help administrators manage the changes between releases, improve upgrade planning, and minimize any problems after the upgrade has completed. Whether enterprises are running Analysis Services for business intelligence, DTS for data processing, Notification Services for alerting, Reporting Services for enterprise reporting, or a combination of components, the Upgrade Advisor can help.

The Upgrade Advisor, built on a rules-based engine, is easy to install and run—even on remote servers. When administrators execute the tool, a simple wizard prompts them to select components on a local or remote server, as shown in Figure 3. Based on the selection, the wizard prompts administrators to identify details about each component. For the database engine, they can pick all the databases on the server or select them separately. The Upgrade Advisor analyzes all stored procedures and embedded Transact-SQL (T-SQL) programs. Furthermore, administrators can point to a SQL trace file that can analyze the T-SQL program running against the databases (an important

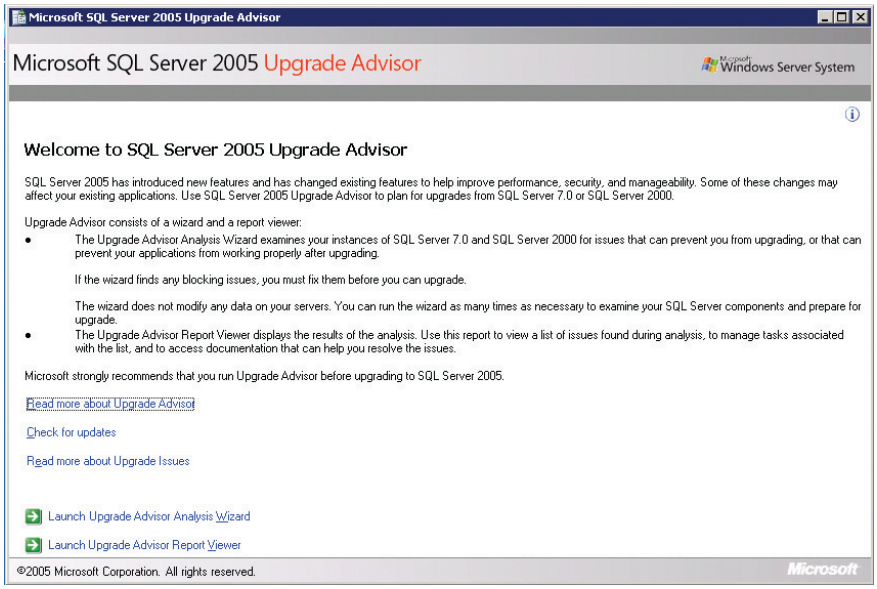


Figure 2. Microsoft SQL Server 2005 Upgrade Advisor Welcome screen

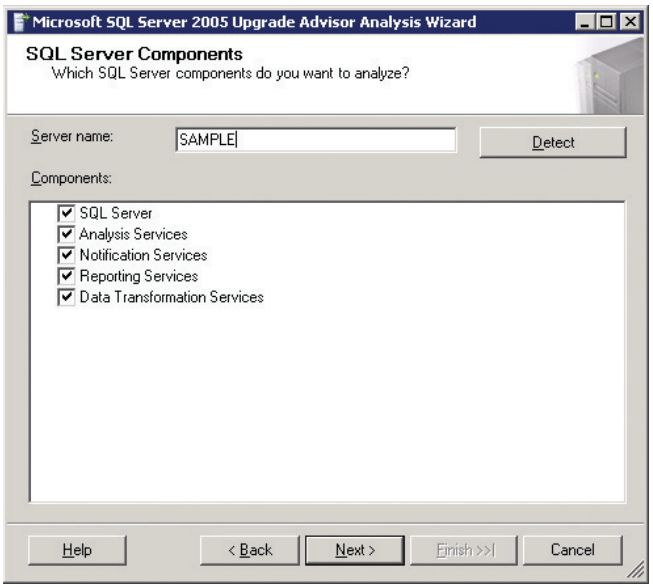


Figure 3. Selecting components for the SQL Server 2005 Upgrade Advisor to analyze

feature when running applications with embedded SQL logic). Administrators can analyze DTS packages that might be stored in files or embedded in the SQL instance they select; they can also select the Notification Services instance at this time.

After the tool completes its analysis, administrators can view a list of issues in the Upgrade Advisor Report Viewer. The Report Viewer provides a summary of issues, noting whether the corrections should be implemented before or after the upgrade. The Report Viewer lets administrators view the details of any server modifications they need to make, which objects (such as scripts or stored procedures) need to be modified, and details about when to make the changes. The Report Viewer also helps administrators manage the modification tasks—letting them check off completed tasks, sort tasks, and create Microsoft Excel spreadsheets of the report details to distribute among members of the project team.

In addition, the Upgrade Advisor lets administrators drill down into the report details, opening a Microsoft Help file that explains how to address specific issues and workarounds. After administrators view the details of a specific issue, they can browse to other rules included in the Help file and see additional areas that the tool evaluates during its analysis. Microsoft includes the Upgrade Advisor with the SQL Server 2005 server installation CD.

Note: Administrators should review the included readme file before installing the Upgrade Advisor; this file contains crucial information about the required prerequisite software and a description of the tool's included rules, known issues, and so forth.

Upgrade process

The Microsoft SQL Server upgrade process can be broken down into four phases: planning and research, testing and process validation, production upgrade, and post-upgrade. This section examines the first three phases. For post-upgrade considerations, see the supplemental online section of this article at www.dell.com/powersolutions.

Planning and research

Developers, database administrators, and application architects should have sufficient resources to start the educational and review process. Their training, experience, and research can drive much of the planning process. Because they intimately know the applications' profiles, they can provide valuable insight into the upgrade details.

The planning phase should move from identifying the databases targeted for the upgrade to determining the changes and processes the upgrade will require. The Upgrade Advisor can help the team determine where to focus its efforts and what to expect. A major decision in this preliminary phase is to decide whether to perform an in-place upgrade or a side-by-side migration. Administrators should base this decision on a combination of factors, including the platform upgrade path available, enhancements to implement during the upgrade, the application architecture, and hardware requirements.

Generally, enterprises should conduct the following planning activities:

- **Learn about the SQL Server 2005 upgrade tools:** Administrators should understand the platform's highlights, examine the functionality, and test the upgrade and migration tools.
- **Assess the application features:** Administrators should evaluate and determine which applications, servers, and databases can benefit most from the upgrade.
- **Select the upgrade path:** Administrators can use the Upgrade Advisor to help determine which upgrade path, in-place upgrade or side-by-side migration, will work best for the environment.
- **Identify the prerequisites for the upgrade process:** Administrators should work with the upgrade team to research compatibility and functionality changes that can help ensure a successful upgrade and can take advantage of the release's enhanced features. The Upgrade Advisor can provide valuable help.
- **Set specific planning and research milestones:** Administrators should determine the upgrade path and steps, set up an initial test plan, and implement a risk mitigation and recovery plan.

Testing and process validation

Nothing can replace testing. Even if administrators plan to upgrade only the SQL Server database engine without changing the application, testing can help identify any backward-compatibility problems and behavioral changes from previous SQL Server releases that the Upgrade Advisor did not detect. Furthermore, testing can help validate data and organize the upgrade process. This phase entails establishing a test environment and composing validation scripts and application functions to confirm a successful upgrade.

The final plan should include a backup of the SQL Server 2000 or SQL Server 7.0 databases and a tested recovery strategy. Also, administrators should identify all application references (such as connection strings, package references, and reports) to the upgraded SQL Server components. For this task, an in-place upgrade offers advantages over a side-by-side migration: When administrators upgrade an earlier SQL Server release in-place through the installation upgrade process, all existing application connections remain the same because the server and the server instance do not change.

Enterprises should conduct the following tasks in the testing and validation phase:

- **Prepare the test environment:** Side-by-side migrations require a separate test SQL Server 2005 installation. In-place upgrades require a test machine running SQL Server 2000 or SQL Server 7.0 and target database copies; hardware comparable to the production setup can allow for production volume testing.
- **Set a pre-upgrade baseline:** This baseline can help administrators evaluate the system post-upgrade and determine any behavioral changes, letting them simulate a typical workload after the upgrade. The baseline can also help administrators confirm functionality and document performance improvements or changes. To set up the baseline, administrators can use familiar tools such as SQL Server Profiler, application load testing tools, Performance Monitor counters, and Showplan statistics.
- **Develop a test plan:** Administrators should set up a generalized testing script or test procedures for the following areas: data validation, data processing, stress and workload, client/server performance, and application functionality
- **Develop a recovery plan:** Administrators should develop upgrade rollback procedures in case of an upgrade interruption. The recovery plan should include running a Database Console Command (DBCC) consistency check on the pre-upgrade databases before backup as well as performing a full restore of the database to validate the backup reliability. After the upgrade, administrators should perform a consistency check and a backup with validation. They also should make sure to test the rollback procedures.
- **Create application-modification procedures:** The test environment should include the full application tier so administrators can confirm that application changes work as expected. These application-modification procedures should include a catalog of affected users. Such procedures also allow for complete documentation of application changes so that they can be applied successfully during the production cutover.
- **Perform an upgrade test run:** A final test run of the upgrade can confirm that the process and procedures work as expected. Administrators can use the Upgrade Advisor after applying the pre-upgrade changes to validate that they have addressed all the problem areas the tool identified.

Production upgrade

The SQL Server 2005 Upgrade Advisor and Setup wizard are designed to help administrators proceed confidently through the planning and testing steps, positioning them for a successful production upgrade. Administrators can use some of the testing steps developed for pre-upgrade use (such as record counts and validation scripts) in validating the upgrade upon completion. Generally, enterprises should perform the following steps for the production upgrade, depending on the SQL Server components being upgraded:

1. **Back up the systems (applications and databases).** Perform a consistency check if applicable, back up the database and related systems, and then validate the backup.
2. **Perform pre-upgrade tasks.** Notify users and then disable the user interface components, pausing all data processing, data entry, and data changes. Make the necessary pre-upgrade changes identified during the testing phase. Re-execute the Upgrade Advisor to validate the pre-upgrade state, and perform an optional secondary backup of the systems before the upgrade.
3. **Perform primary SQL Server back-end platform upgrade tasks.** Run SQL Server 2005 for a side-by-side migration. Install the Microsoft .NET Framework and SQL Native Client. In the Setup wizard, specify the same instance as the legacy installation. Then, specify the same components as the legacy instance (for example, Database Services, Analysis Services, and Reporting Services). Once the setup is complete, perform the tasks required for special upgrade considerations (such as repopulation of full-text indexes, special handling of clusters, or log shipping). Next, make any post-upgrade platform changes, such as scripts or tasks required to support the back-end functionality on the new SQL Server 2005 platform. Finally, run platform data and functionality validation testing scripts to confirm the success of the SQL Server 2005 upgrade.

4. **Make primary application changes.** Make application functionality changes to support the new back-end structures, and make any required database reference changes in application connection strings and other connection references. Test application functionality, including data processing, front-end and report usage, and other application components based on the test procedures created in the planning phase.
5. **Perform post-upgrade steps.** For the database engine, the upgrade automatically sets the compatibility mode to 8.0; however, administrators may wish to switch to compatibility mode 9.0 to take advantage of the features introduced in SQL Server 2005. For side-by-side migrations, stop the former platform services (or set the database to read-only) to prevent unknown data changes. For the relational data, run DBCC consistency checks to validate the data. Back up SQL Server 2005 structures and data with backup validation, and back up application systems and files. Then, re-enable processing and the application user interface, notifying users that the upgrade is complete.

Upgrade considerations for specific SQL Server 2005 components

Given the breadth of the Microsoft SQL Server 2005 platform, upgrade processes vary for different components. This section examines the upgrade considerations for major SQL Server 2005 components.[1]

Upgrading to the SQL Server 2005 database engine

The database engine is the easiest to upgrade of all SQL Server components, and upgrading it can provide an immediate return on investment in the areas of management, performance, and high availability. The two main options for the database engine upgrade are side-by-side migration (in which the SQL Server 2005 engine is installed as a secondary instance on the same server as the SQL Server 2000 or SQL Server 7.0 engine or on a completely separate server) and an in-place upgrade (in which an instance of SQL Server 2000 or SQL Server 7.0 is upgraded through the installation process and databases and other objects are upgraded "in place").

With a side-by-side migration, the most common upgrade path is a simple database detach and re-attach on the SQL Server 2005 instance or a database backup and restore from the older version to the new version. If administrators retain an up-to-date version of the metadata scripts, they also can create the objects on the SQL Server 2005 server and use the bcp utility to export and import the data. The other option is an in-place upgrade, in which administrators upgrade and adapt the databases, settings, and extended features to the SQL Server 2005 engine during the installation process; when running the setup process on a server that has a SQL Server 2000 or SQL Server 7.0 instance, administrators should see an option to upgrade the selected instance to SQL Server 2005.

Note that for the database engine upgrade, all existing Microsoft Data Access Components (MDAC) and ADO.NET applications should continue to function as when they were running on SQL Server 2000 or SQL Server 7.0. In fact, SQL Server 2005 does not include an updated release of MDAC. However, SQL Server 2005 introduces the SQL Native Client, which combines an updated SQL Open Database Connectivity (ODBC) driver and SQL OLE database (OLEDB) provider with network libraries in a single dynamic-link library (DLL). The SQL Native Client lets administrators leverage the SQL Server 2005 client-access features, such as Multiple Active Result Sets (MARS), the XML data type, and user-defined types (UDTs). SQL Server 2005 provides tight integration with the Microsoft .NET Framework 2.0, which includes the latest ADO.NET version.

The in-place server upgrade typically is easier to perform than the side-by-side migration. Although this approach requires a more thorough fallback plan and testing, it also provides seamless connectivity. By performing an in-place upgrade, logins and users remain in sync, database connections remain the same for applications, and SQL Agent jobs and other functionality are concurrently upgraded during the installation. Note that several features, such as log shipping, replication, and cluster environments, have special upgrade considerations.

For the database engine, the upgrade sets the compatibility mode to 8.0. Keeping this setting at 8.0 may be beneficial for certain circumstances, such as for T-SQL references that are no longer supported in SQL Server 2005. The analysis phase of the upgrade process should uncover situations in which using a compatibility setting lower than 8.0 may be preferable. However, best practices recommend fixing any syntax that requires a compatibility level lower than 9.0 (SQL Server 2005) during the upgrade process. By reworking the syntax, developers can have immediate access to the programming enhancements and features in the SQL Server 2005 release. To isolate these type of issues and other syntax that can cause upgrade trouble, administrators can script out the objects and procedures from the previous platform version and attempt to run the scripts within SQL Server 2005. A simple attach or restore might suppress these issues. Also, some SQL logic can be embedded in the application. For data validation, administrators can run the `DBCC checkdb` statement on the attached or restored database to confirm the integrity of the migrated data.

[1] For more information about Microsoft SQL Server 2005 upgrade considerations, see the white papers, Webcasts, and other resources listed at www.microsoft.com/technet/prodtechnol/sql/2005/sqlupgrd.mspx#ECAA.

Note: Microsoft recommends using Information_Schema views to obtain various metadata instead of querying the system tables directly because Microsoft cannot guarantee that the underlying object structure will persist in new platforms. With the release of SQL Server 2005, Microsoft has changed the SQL Server underlying object structure. Also, SQL Server 2005 catalog views and Dynamic Management Views (DMVs) have restricted permissions. PUBLIC users no longer have permissions to view catalog views, and users with GUEST/PUBLIC permissions cannot select from DMVs.

Migrating to SQL Server 2005 Integration Services

Microsoft did not use the name of the SQL Server 2005 Integration Services predecessor, Data Transformation Services, for its SQL Server 2005 extraction, transformation, and loading (ETL) component because SSIS was a complete code rewrite—Microsoft did not use the DTS code to create this component. With industry demands for fast performance and hardware consolidation to handle ever-increasing data complexity and volume, DTS was not positioned as a long-term solution. Although DTS and SSIS are both ETL tools, their architectures diverge greatly. Because of this, moving from DTS to SSIS requires a migration, which involves redesign and solution changes to leverage the SSIS features.

The migration from DTS to SSIS uses wizard-driven output along with manual redesign. Some DTS tasks have a straightforward upgrade path to SSIS and are accommodated by the DTS Migration Wizard. Administrators may be able to use this wizard to upgrade other tasks depending on their use and design, but some tasks may be more difficult to upgrade or not upgradeable. The supplemental online section of this article, available at www.dell.com/powersolutions, describes some issues that administrators may encounter when upgrading DTS packages.

Administrators can incrementally migrate packages to SSIS. When installing SQL Server 2005, they have the option to install the runtime files required for DTS packages to execute on SQL Server 2005—without SQL Server 2000 needing to be installed. This makes the side-by-side migration appealing, especially in an environment where the DTS packages contain many tasks that require manual migration. SSIS also contains an Execute DTS package object when the runtime files or SQL Server 2000 has been installed on the SSIS server. A side-by-side implementation of SQL Server 2005 SSIS and SQL Server 2000 DTS can provide flexibility as administrators approach package migration.

Upgrading Analysis Services

Dimensions, partitions, storage modes, aggregates, and measures—the strengths of Analysis Services 2000—have been preserved in the Analysis Services 2005 release. However, SQL Server 2005 also brings many notable enhancements. The Unified Dimensional Model (UDM), for example, extends beyond traditional online analytical processing (OLAP) sources to allow expanded relational and aggregate data in a unified view. Dimensions are another area with valuable changes from previous SQL Server versions. A shift from a hierarchy-based model to an attribute-based model, with related optimizations on the storage and aggregation side, allows Analysis Services 2005 to scale for enterprise performance and volume.

From an upgrade perspective, Microsoft provides a direct in-place upgrade from Analysis Services 2000 to Analysis Services 2005—preserving cubes, partitions, dimension hierarchies, measures, calculations, and sets. Because Analysis Services objects are built on top of a Data Source View (DSV) referencing database engines, best practices recommend creating the DSV on the base tables that the Analysis Services 2000 objects are built on rather than on views referencing underlying tables. The Migration Wizard generates DSVs that are complete with relationships and attributes from source tables. This can allow developers to add attributes to the cube even though they were not present in Analysis Services 2000.

Note that the Migration Wizard does not optimize the Analysis Services objects; it simply moves the objects in place to the new Analysis Services server. The goal of the wizard is to migrate the cube structures and architecture objects so that client applications relying on the Analysis Services 2000 structures do not fail after administrators have migrated the cube to Analysis Services 2005. Thus, the migrated cube design may not take advantage of SQL Server 2005 enhancements. However, the cubes should have the immediate performance and scalability benefits of the Analysis Services 2005 architecture. When the Migration Wizard finishes its processes, administrators can then reprocess the cube and test the data and reports.

For Analysis Services 2005, the major upgrade considerations revolve around the client-access methods and structure impact to reports. Analysis Services 2005 takes advantage of the Web service protocol for OLAP—XML for Analysis (XML/A)—that Microsoft helped write. (Support for XML/A was available for Analysis Services 2000 as a Web release, letting an Analysis Services 2000 server listen and respond to XML/A requests.) With native support for XML/A in Analysis Services 2005, administrators should update existing client components of OLEDB for OLAP (Pivot Table Services, or PTS) to access SQL Server 2005. That means users require the latest version of PTS that is included with SQL Server 2005. The new driver should be installed side-by-side with the earlier PTS version, letting users access both SQL Server 2005 and SQL Server 2000 Analysis Services.

The second client-access consideration is the OLAP structure and related Multidimensional Expression (MDX) compatibility after the upgrade. MDX is not gracious to members and structures that

have changed. Although the Upgrade Wizard sufficiently recreates the OLAP structure, with the dimension architecture change from hierarchy based to attribute based, administrators may find small structural and data anomalies that appear after the upgrade. Therefore, report and data testing are more critical here than on the database engine side. Administrators may need to recreate reports and underlying MDX for the structures in Analysis Services 2005.

Upgrading Reporting Services

Because Microsoft initially released Reporting Services 2000 in early 2004, the Reporting Services 2005 platform does not introduce major architectural changes, but it does offer features such as multi-select parameters, built-in MDX support, and dynamic report generation. Microsoft provides a direct, in-place upgrade path for moving from Reporting Services 2000 to Reporting Services 2005. Furthermore, Reporting Services 2005 runs Report Definition Language (RDL) report definitions created in Reporting Services 2000 without requiring administrators to upgrade the definitions. However, when developers open a report in Business Intelligence (BI) Development Studio, they are prompted to convert the RDL definitions to the Reporting Services 2005 standards.

A tool for successful upgrades

Managing the upgrade to Microsoft SQL Server 2005 requires significant planning and testing. With appropriate forethought and preparation—and use of the SQL Server 2005 Upgrade Advisor and Setup wizard—administrators can avoid problems and identify the areas where they need to concentrate their efforts. After performing a smooth upgrade, administrators can be ready to fully leverage the power and functionality that SQL Server 2005 is designed to provide.

Erik Veerman is an associate mentor for Solid Quality Learning and has designed several SQL Server–based business intelligence solutions across a broad business spectrum. Erik—an expert in OLAP design, ETL processing, and dimensional modeling—is a frequent presenter for his local Professional Association for SQL Server (PASS) chapter and speaks at the national PASS and *SQL Server Magazine* Connections conferences.

Dynamic Deployment Methodologies for Oracle RAC Databases

Dynamic provisioning allocates resources at the time they are needed. To improve provisioning efficiency, enterprise IT departments can organize resources into components and create an automated process for bare-metal deployments. This approach can prove especially effective for Oracle® Real Application Clusters (RAC) deployments because hardware resources can be dynamically allocated and reprovisioned based on various business priorities.

BY UJJWAL RAJBHANDARI AND DAVID MAR

Related Categories:

Application servers

Cluster management

Clustering

Database

Oracle

Resource allocation

System deployment

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

Creating a dynamically provisioned cluster involves booting the pre-OS environment, installing the OS, installing the node communication application layer, and finally, provisioning for updates. Before building a dynamic provisioning infrastructure, however, enterprises should ascertain the need for dynamic deployments and examine their potential benefits.

Dynamic deployments can help reduce costs resulting from a large number of idle servers. This may occur in IT environments in which applications exhibit large spikes and variations in their loads. For instance, an application that allows students to check their grades might experience an extremely heavy load only at the end of a semester. Another example scenario is at the end of a business quarter, when an enterprise's data center could see extremely high spikes in input and data manipulation. Such systems could be configured to allow reprovisioning to other applications during times of low usage.

Dynamic deployment can also be advisable when an enterprise needs a new cluster of servers and the administrator foresees that the enterprise will require additional servers with similar software configurations in the future. Designing systems that can be dynamically provisioned during the initial deployment may take additional time, but it facilitates adding or scaling out a cluster in the future.

In addition, an enterprise's proactive approach to hardware management can justify dynamic provisioning. Hardware has a predictable life cycle. Administrators usually desire to minimize mean time between failure and consequent down time by periodically replacing older systems with new hardware.

This article explores steps to take when planning and implementing dynamic deployments. It also examines specific steps for dynamically deploying an Oracle Real Application Clusters (RAC) environment.

Planning component phases for deployment

Administrators should divide deployments into components: pre-OS installation, OS installation, and application-layer installation. Organizing dynamic deployment in component phases is essential. Although an imaging solution can be used to configure clusters once they have been deployed, this approach could lead to more work and less flexibility in the future than using a component-based deployment.

Dividing a dynamic deployment into components can provide several benefits. A component-based approach lets IT administrators mix and match software and OS components and specify which applications and services are deployed on each system. Conversely, a monolithic-image approach could leave an administrator with only one software stack and one hardware set on which to deploy it. For example, if an underlying hardware driver changes, an administrator must update the OS with the latest driver. If the administrator had imaged the systems monolithically, an entire new image snapshot of the system would need to be created, introducing additional risk to deployments. Conversely, creating an independent OS module means that only that particular module would need to be modified, leaving the post-OS, application installation, and patch-set modules intact.

Designing systems that can be dynamically provisioned during the initial deployment facilitates adding or scaling out a cluster in the future.

Automating deployment with a component-based approach goes far beyond simply choosing installation packages. When designed correctly, a component-based methodology lets IT administrators use industry-standard components. This allows deployments to work properly on a myriad of hardware platforms—such as Linux® and Microsoft® Windows® operating systems—and permits flexibility in choosing application-layer components. The level of granularity varies based on the deployment. For example, with Linux and Windows, an administrator may choose an OS layer that deploys only the base OS installation. Creating another layer could permit inclusion of advanced components such as a professional or developer's edition of the OS deployment. The best practices presented in this article address the rudimentary components of a dynamic deployment.

Defining the building blocks

A deployment that is broken down into several logical building blocks, as described in the following sections, allows administrators to modify those blocks without changing others.

Pre-OS environment

The pre-OS environment is the condition of the hard drive before any software resides on it—the bare-metal state. A pre-OS environment allows a system to boot up in the absence of an OS. This bootable image lets administrators install an OS or execute setup operations before the OS is deployed.

Various preinstallation operating systems are available, such as Windows Preinstallation Environment (WinPE) and Embedded Linux. The pre-OS environment is ideal for identifying the hardware configuration, choosing the right drivers to read and write to the system, and setting up the network so that each node can access a remote server.

Typically, the pre-OS environment requires the following technologies:

- **Wake-on-LAN:** If a BIOS is enabled for Wake-on-LAN and the system is turned off, it will turn itself on if a wake-up data packet is sent to the node. The packet must contain six bytes of a synchronization stream (FFh) followed by 16 copies of the node's Ethernet address. This feature is particularly useful when no one is present at the data center and a need arises to reprovision or install an OS. Dell™ PowerEdge™ servers have built-in Wake-on-LAN functionality, which can easily be enabled to manage the servers remotely.
- **PXE:** Preboot Execution Environment (PXE) is used for dynamically booting or provisioning a server. PXE allows a system to boot over the network. Using PXE requires a BIOS and network interface card (NIC) that support the PXE framework. The network cards and BIOS in Dell PowerEdge servers have integrated PXE capabilities.
- **DHCP:** The task of a Dynamic Host Configuration Protocol (DHCP) server is to assign IP addresses to all nodes on a network so that each computer may communicate. Before a computer begins talking on the network, it must obtain an address that uniquely identifies it. It is analogous to a telephone, which cannot make or receive calls until it has received a number. Enabling the DHCP server for each node is the first step in dynamically adding a node to the infrastructure.
- **TFTP:** Trivial FTP (TFTP) is generally used for booting thin clients from a network host or transferring very small files. It uses User Datagram Protocol (UDP) on port 69 as its transport protocol and—unlike its more robust cousin, FTP—lacks features such as encryption, authentication, or the ability to list directory contents. However, like FTP it has the ability to read and write files from a remote server and can transfer both binary files and ASCII files. Each of the nodes uses TFTP to obtain a pre-OS environment as the first step toward deployment.

OS installation

Deploying an OS through scripted installation allows the installation to not be bound to a particular image, enabling flexibility for OS provisioning. For example, with a scripted OS installation, the same installation scripts can be used on multiple types of machines.

Two predominant methods are used to script an installation: Red Hat® kickstart installation using the ks.cfg file and Windows unattended installation using the unattend.txt file. Both allow scripted customization of an OS installation.

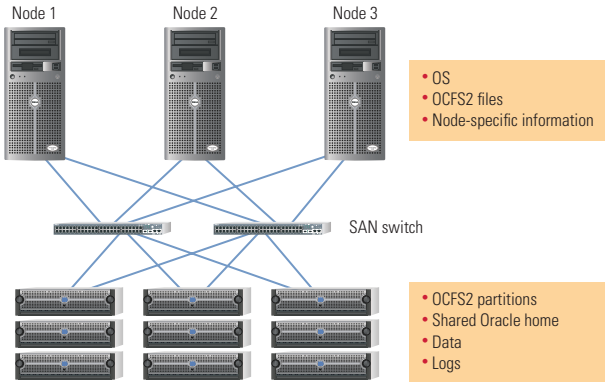


Figure 1. Typical Oracle RAC implementation with shared Oracle home

Application-layer installation

The application-layer phase establishes a means of connecting individual servers to a cluster. During OS installation, each server operates individually without knowledge of its peers. This phase involves reestablishing network connectivity and establishing shared storage among the server nodes. Many technologies exist that allow an enterprise's cluster to share storage.

Deploying Oracle RAC using the shared-home method and OCFS2

After the OS is deployed, Oracle RAC can be deployed using the shared home method. This method installs all Oracle binaries in a common location and configures the RAC nodes to contain only node-specific information to supplement the common Oracle home. Each node can boot using the local disk that contains the OS, and at the same time, all Oracle binaries are located on a common storage area network (SAN) disk.

Oracle Cluster File System 2 (OCFS2) provides shared-home support for Oracle RAC environments. This file system enables Oracle RAC database files to be stored on shared storage accessible by all cluster nodes. Using OCFS2, all cluster nodes can access files used for database storage as well as the binary files for running the database. Consequently, application-layer installation is simplified because the Oracle database needs to be installed only once.

Because OCFS2 for a given node is specific to that RAC node, the OCFS2 parameter files must reside on the local disk. However, after OCFS2 is installed and the common attached storage disks are formatted for OCFS2, those partitions can be accessed by all nodes of the RAC cluster. Figure 1 shows a typical OCFS2-based shared-home implementation. The RAC nodes contain the OS and node-specific parameters such as host name, IP addresses, and so forth. They also contain the OCFS2 parameter files that are tied to the node. The common Oracle binaries are located in the common storage disk partition in the SAN.

When a need arises to expand RAC functionality, administrators simply deploy the OS on an additional node and set up the OCFS2 modules on that node. The node can then be added to the common storage subsystem where it can access the shared-home Oracle binaries.

The shared-home approach enables administrators to easily manage the Oracle RAC environment. Administrators can spend minimal time managing each RAC node individually, thereby enhancing administrator productivity. In addition, with storage backup functionality available on all back-end storage, an image of the shared home can be easily created, which can help provide effective disaster recovery.

Oracle Universal Installer

After the OS has been installed on systems that are part of the Oracle RAC environment, the Oracle binaries must be deployed to all the nodes of the cluster. This can be done using the Oracle Universal Installer (OUI). The OUI helps administrators install RAC and perform Oracle prerequisite checks. It also enables automated installation of Oracle RAC using response files.

Provisioning servers dynamically

The building blocks for dynamic provisioning include booting the pre-OS environment, installing the OS, and setting up node communication at the application layer. Dividing the deployment process into these components can give administrators flexibility in planning and managing an enterprise's growing IT infrastructure.

Ujjwal Rajbhandari is a systems engineer in the Database and Application Engineering Department of the Dell Product Group. He is responsible for validating Oracle RAC solutions on Dell PowerEdge servers and Dell/EMC Fibre Channel storage. Ujjwal has a B.E. in Electrical Engineering from the Indian Institute of Technology, Roorkee, and an M.S. in Electrical Engineering from Texas A&M University.

David Mar is a senior software engineer in the Dell Database and Application Engineering Department of the Dell Product Group. His principal focus is developing deployment strategies for Oracle-based solutions on Dell PowerEdge servers, Dell PowerVault™ storage, and Dell/EMC storage. David has a B.S. in Computer Engineering and an M.S. in Computer Science from Texas A&M University.

FOR MORE INFORMATION

Oracle Real Application Clusters:
www.oracle.com/technology/products/database/clustering/index.html

IT Executive Learning Series

By IT Leaders, for IT Leaders

Engage in an IT-to-IT Discussion

Dell IT ELS events offer a number of informative sessions hosted by Dell's top IT executives. These sessions give a behind-the-scenes look at Dell's IT organization and cover a range of topics including supply chain management, security, disaster recovery, data management, and much more*. These complimentary events provide an outstanding opportunity for you to connect with your peers and gain insight into several Dell IT best practices. The IT Executive Learning Series is an IT-to-IT event specifically designed and presented "By IT Leaders, For IT Leaders."

*Topics vary by event

What Your Peers Have Said About Dell IT Events

- "I found the series to be very informative, and more importantly, thought provoking."
- "Excellent, candid presentations…"
- "Substantially more relevant than going to a Gartner or META conference."
- "I appreciated the IT-to-IT discussion as opposed to a sales event."
- "Time well spent, and I look forward to coming back again."

Visit **www.dell.com/it** for dates, topics, and registration information

DELL™

Information Technology

Testing Oracle 10*g* RAC Scalability

on Dell PowerEdge Servers and Dell/EMC Storage

Oracle® 10*g* Real Application Clusters (RAC) software running on standards-based Dell™ PowerEdge™ servers and Dell/EMC storage can provide a flexible, reliable platform for a database grid. Administrators can scale the database easily and reliably simply by adding nodes to the cluster. A team of engineers from Dell and Quest Software ran benchmark tests against a Dell-based Oracle 10*g*® Release 1 RAC cluster to demonstrate the scalability of this platform.

BY ZAFAR MAHMOOD; ANTHONY FERNANDEZ; BERT SCALZO, PH.D.; AND MURALI VALLATH

Related Categories:

Characterization

Oracle

Benchmark Factory

Quest Software

Scalable enterprise

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

When businesses start small and have little capital to invest, they tend to keep their computer systems simple, typically using a single database server and a storage array to support a database. However, as a business grows, this simple database configuration often cannot handle the increased workload. At this point, businesses typically upgrade their hardware by adding more CPUs and other required resources or they add servers to create a cluster. Increasing resources, or *vertical scaling*, is like placing a temporary bandage on the system—it solves the current situation of increased workload and resource demand but does not address the possibility of future workload increase. Instead of adding more memory or CPUs to the existing configuration, organizations can add servers and configure them to function as a cluster to provide load balancing, workload distribution, and availability—a process known as *horizontal scaling*.

The growth potential for a vertically scalable system is limited because it reaches a point at which the addition of resources does not provide proportionally improved results. In contrast, horizontal scalability enabled by clustered hardware can distribute user workload among multiple servers, or *nodes*. These nodes may be relatively small and can use inexpensive standards-based hardware, offering economical upgradeability options that can enhance a single large system. In addition, clusters offer both horizontal and vertical scalability, providing further investment protection.

Understanding Oracle 10*g* Real Application Clusters

Oracle 10*g* Real Application Clusters (RAC) is designed to provide true clustering in a shared database environment. A RAC configuration comprises two or more nodes supporting two or more database instances clustered together through Oracle clusterware. Using Oracle's cache fusion technology, the RAC cluster can share resources and balance workloads, providing optimal scalability for today's high-end computing environments. A typical RAC configuration consists of the following:

- A database instance running on each node
- All database instances sharing a single physical database
- Each database instance having common data and control files
- Each database instance containing individual log files and undo segments
- All database instances simultaneously executing transactions against the single physical database
- Cache synchronization between user requests across various database instances using the cluster interconnect

Figure 1 shows the components of a typical RAC cluster.

Oracle clusterware

Oracle clusterware comprises three daemon processes: Oracle Cluster Synchronization Services (CSS), Oracle Event Manager (EVM), and Oracle Cluster Ready Services (CRS). This clusterware is designed to provide a unified, integrated solution that enables scalability of the RAC environment.

Cluster interconnect

An interconnect is a dedicated private network between the various nodes in a cluster. The RAC architecture uses the cluster interconnect for instance-to-instance block transfers by providing cache coherency. Ideally, interconnects are Gigabit Ethernet adapters configured to transfer packets of the maximum size supported by the OS. Depending on the OS, the suggested protocols may vary; on clusters running the Linux® OS, the recommended protocol is UDP.

Virtual IP

Traditionally, users and applications have connected to the RAC cluster and database using a public network interface. The network protocol used for this connection has typically been TCP/IP. When a node or instance fails in a RAC environment, the application is unaware of failed attempts to make a connection because TCP/IP can take more than 10 minutes to acknowledge such a failure, causing end users to experience unresponsive application behavior.

Virtual IP (VIP) is a virtual connection over the public interface. If a node fails when an application or user makes a connection using VIP, the Oracle clusterware—based on an event received from EVM—will transfer the VIP address to another surviving instance. Then, when the application attempts a new connection, two possible scenarios could ensue, depending on the Oracle 10*g* database features that have been implemented:

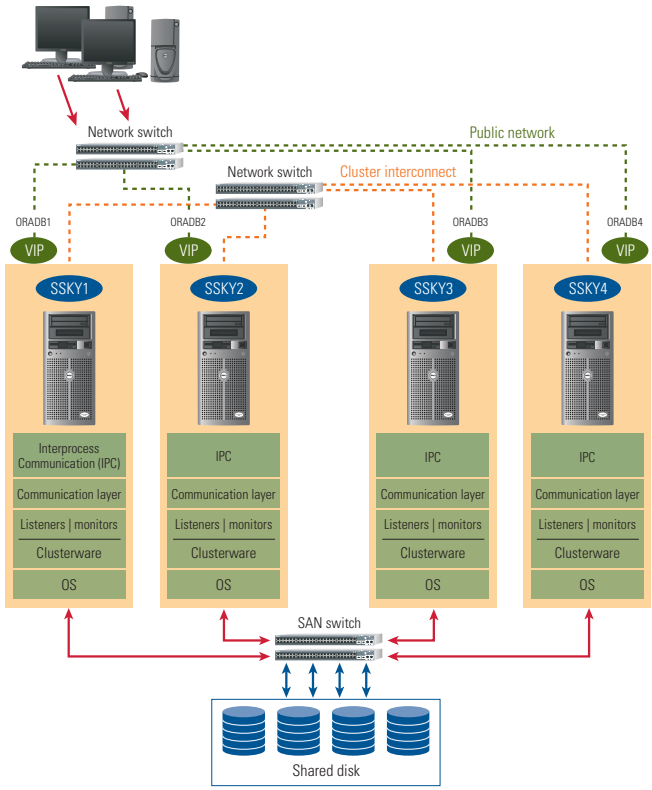
- If the application uses Fast Application Notification (FAN) calls, Oracle Notification Services (ONS) will inform ONS running on the client systems that a node has failed, and the application—using an Oracle-provided application programming interface (API)—can receive this notification and connect to one of the other instances in the cluster. Such proactive notification mechanisms can help prevent connections to a failed node.
- If the application attempts to connect using the VIP address of the failed node, the connection will be refused because of a mismatch in the hardware address and the application is immediately notified of the failure.

Shared storage

Another important component of a RAC cluster is its shared storage, which is accessed by all participating instances in the cluster. The shared storage contains the data files, control files, redo logs, and undo files. Oracle Database 10*g* supports three methods for storing files on shared storage: raw devices, Oracle Cluster File System (OCFS), and Oracle Automatic Storage Management (ASM).

Raw devices. A raw device partition is a contiguous region of a disk accessed by a UNIX® or Linux character-device interface. This interface provides raw access to the underlying device, arranging for direct I/O between a process and the logical disk. Therefore, when a process issues a write command to the I/O system, the data is moved directly to the device.

Oracle Cluster File System. OCFS is a clustered file system developed by Oracle to provide easy data file management as well



Source: *Oracle 10g RAC Grid, Services & Clustering* by Murali Vallath, 2005.

Figure 1. Components within a typical Oracle 10*g* RAC cluster

	Hardware	Software
Oracle 10*g* RAC cluster nodes (10)	Dell PowerEdge 1850 servers, each with: • Two Intel® Xeon® processors at 3.8 GHz • 4 GB of RAM • 1 Gbps* Intel NIC for the LAN • Two 1 Gbps LAN on Motherboards (LOMs) teamed for the private interconnect • Two QLogic QLA2342 HBAs • Dell Remote Access Controller • Two internal RAID-1 disks (73 GB 10,000 rpm) for the OS and Oracle Home	• Red Hat Enterprise Linux AS 4 QU1 • EMC PowerPath 4.4 • EMC Navisphere® agent • Oracle 10*g* R1 10.1.0.4 • Oracle ASM 10.1.0.4 • Oracle CRS 10.1.0.4 • Linux bonding driver for the private interconnect • Dell OpenManage
Benchmark Factory for Databases servers (2)	Dell PowerEdge 6650 servers, each with: • Four Intel Xeon processors • 8 GB of RAM	• Microsoft Windows Server™ 2003 • Benchmark Factory application and agents • Spotlight on RAC
Storage	• Dell/EMC CX700 storage array • Dell/EMC Disk Array Enclosure with 30 disks (73 GB 15,000 rpm) • RAID Group 1: 16 disks with four 50 GB RAID-10 logical units (LUNs) for data and backup • RAID Group 2: 10 disks with two 20 GB LUNs for the redo logs • RAID Group 3: 4 disks with one 5 GB LUN for the voting disk, Oracle Cluster Repository (OCR), and spfiles • Two 16-port Brocade SilkWorm 3800 Fibre Channel switches • Eight paths configured to each logical volume	• EMC FLARE™ Code Release 16
Network	• 24-port Dell PowerConnect™ 5224 Gigabit Ethernet switch for the private interconnect • 24-port Dell PowerConnect 5224 Gigabit Ethernet switch for the public LAN	• Linux binding driver used to team dual on-board NICs for the private interconnect

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

Figure 2. Hardware and software configuration for the test environment

Database	Oracle Database 10*g* R1 10.1.04 Enterprise Edition
ASM disk groups	SYSTEMDG: 50 GB DATADG: 50 GB INDEXDG: 50 GB REDO01DG: 20 GB REDO02 DG: 20 GB All disk groups were created using the external redundancy option of ASM.
Tablespaces	Quest_data in the DATADG disk group (40 GB) using the OMF feature Quest_index in the INDEXDG disk group (10 GB) using the OMF feature All other database tablespaces were created in the SYSTEMDG disk group. Redo log files were created in the REDO01DG and REDO02DG disk groups

Figure 3. Database configuration for the test environment

as performance levels similar to raw devices. OCFS 1.0 supports only database files to be stored on devices formatted using OCFS, while OCFS 2.0 supports both Oracle and non-Oracle files. OCFS supports both Linux and Microsoft® Windows® operating systems.

Oracle Automatic Storage Management. ASM is a storage management feature introduced in Oracle Database 10*g*. ASM is designed to integrate the file system and volume manager. Using Oracle Managed Files (OMF) architecture, ASM distributes the I/O

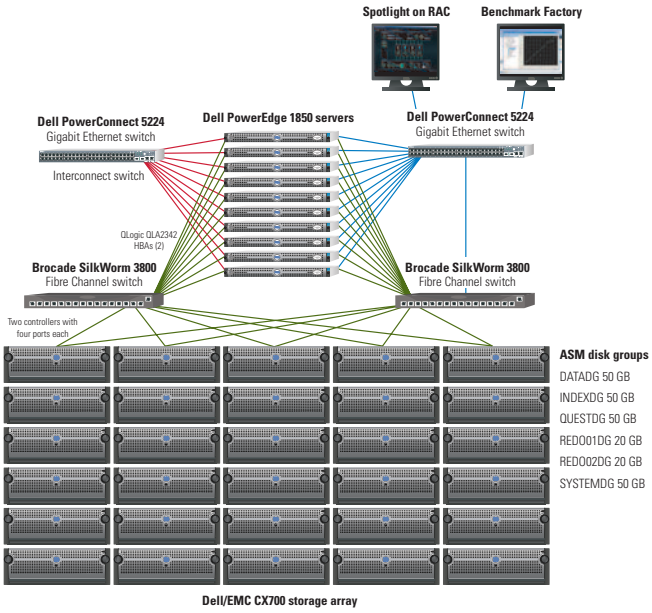


Figure 4. Ten-node cluster architecture for the test environment

load across all available resources to help optimize performance and throughput.

Testing Oracle RAC environments for scalability

The primary advantages of Oracle RAC systems, apart from improved performance, are availability and scalability. Availability is enhanced with RAC because, if one of the nodes or the instances in the cluster fails, the remainder of the instances would continue to provide access to the physical database. Scalability is possible because, when the user workload increases, users can access the database from any of the available instances that have resources available. Database administrators (DBAs) also can add nodes to the RAC environment when the user base increases.

When organizations migrate to a RAC environment, best practices recommend conducting independent performance tests to determine the capacity of the cluster configured. Such tests can help determine when a cluster will require additional instances to accommodate a higher workload. To illustrate this, in August and September 2005 engineers from the Dell Database and Applications team and Quest Software conducted benchmark tests on Dell PowerEdge servers and Dell/EMC storage supporting an Oracle 10*g* RAC database cluster. The results of these tests demonstrate the scalability of Dell PowerEdge servers running Oracle RAC and ASM.

Figure 2 lists the hardware and software used in the test environment, while Figure 3 describes the database configuration, including the disk groups and tablespaces. Figure 4 shows the layout of the cluster architecture.

Benchmark Factory for Databases

Benchmark Factory® for Databases from Quest Software provides a simple yet robust graphical user interface (GUI), shown in Figure 5, for creating, managing, and scheduling industry-standard database benchmarks and real-world workload simulation. It helps determine accurate production database hardware and software configurations for optimal effectiveness, efficiency, and scalability. Using Benchmark Factory, DBAs can address two challenging tasks: selecting a hardware architecture and platform for deployment and determining the appropriate performance-related service-level agreements.

While Benchmark Factory offers numerous industry-standard benchmarks, the test team selected benchmarks similar to the TPC-C benchmark from the Transaction Processing Performance Council (TPC). This benchmark measures online transaction processing (OLTP) workloads, combining read-only and update-intensive transactions that simulate the activities found in complex OLTP enterprise environments.

The benchmark tests simulated loads from 100 to 5,000 concurrent users in increments of 100, and these tests used a 10 GB database created by Benchmark Factory. The goal was to ascertain two critical data points: how many concurrent users each RAC node could sustain, and whether the RAC cluster could scale both predictably and linearly as additional nodes and users were added.

Spotlight on RAC

Spotlight® on RAC from Quest Software is a database monitoring and diagnostic tool that extends the proven architecture and intuitive GUI of the Spotlight on Oracle tool to RAC environments. Spotlight on RAC is designed to provide a comprehensive yet comprehendible overview of numerous internal Oracle RAC settings and metrics represented by a dashboard-like display (see Figure 6) that enables easy monitoring and diagnostics of a database cluster. With this tool, DBAs can easily monitor their clusters to detect, diagnose, and correct potential problems or hotspots. Spotlight on RAC also provides alarms with automatic prioritization and weighted escalation rankings to help less-experienced RAC DBAs focus their attention on the most critical or problematic issues. Spotlight on RAC requires only a Windows OS on the client system to monitor all the nodes on the cluster. It does not require server-side agents or a data repository.

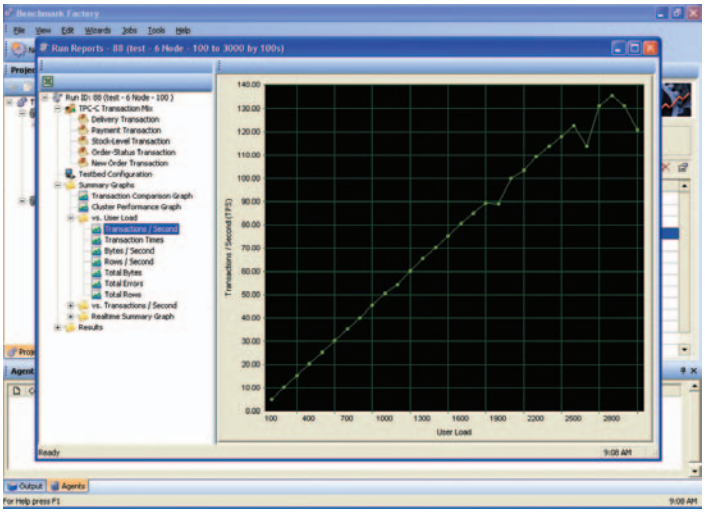


Figure 5. Benchmark Factory for Databases GUI

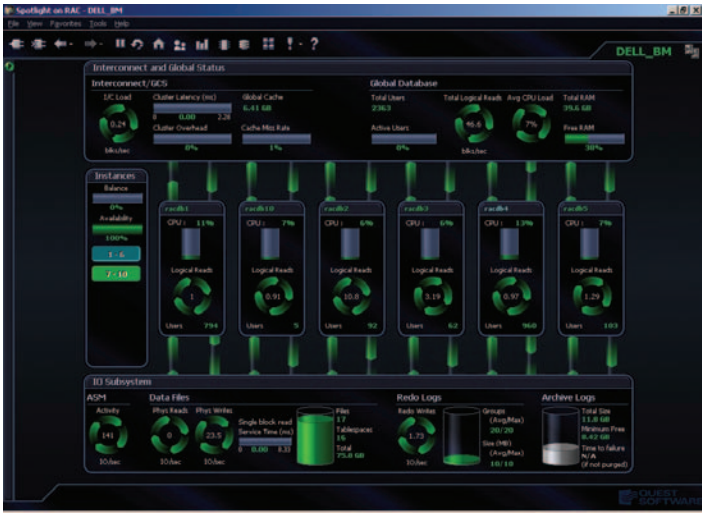


Figure 6. Spotlight on RAC GUI

Defining the testing methodology

Methodology is critical for any reliable benchmarking exercise, especially for complex and repetitive benchmark tests. A methodology allows for comparison of current activity with previous activities, while recording any changes to the baseline criteria. Oracle RAC testing is no different—a methodology to identify performance candidates, tune parameters or settings, run the tests, and then record the results is critical. And because of its highly complex multi-node architecture, RAC benchmarking should follow an iterative testing process, as described in this section.

For the first node and instance, administrators should take the following steps:

1. **Establish a fundamental baseline.** Install the OS and Oracle database (keeping all normal installation defaults); create and populate the test database schema; shut down and restart the database; and run a simple benchmark (such as TPC-C for 200 users) to establish a baseline for default OS and database settings.

2. **Optimize the basic OS.** Manually optimize the typical OS settings; shut down and restart the database; run a simple benchmark (such as TPC-C for 200 users) to establish a new baseline for basic OS improvements; and repeat the prior three steps until a performance balance results.
3. **Optimize the basic non-RAC database.** Manually optimize the typical database spfile parameters; shut down and restart the database; run a simple benchmark (such as TPC-C for 200 users) to establish a new baseline for basic Oracle database improvements; and repeat the prior three steps until a performance balance results.
4. **Ascertain the reasonable per-node load.** Manually optimize the scalability database spfile parameters; shut down and restart the database; run an increasing user load benchmark (such as TPC-C for 100 to 800 users, with user load increasing by increments of 100) to determine how many concurrent users a node can reasonably support, a measurement referred to as the "sweet spot"; monitor the benchmark test via the vmstat command, looking for the points at which excessive paging and swapping begins and the CPU idle time consistently approaches zero; record the "sweet spot" number of concurrent users, which represents an upper limit; and reduce the "sweet spot" number of concurrent users by some reasonable percentage to account for RAC architecture and inter- and intra-node overheads (for example, reduce it by 10 percent).
5. **Establish the baseline RAC benchmark.** Shut down and restart the database; create an increasing user load benchmark based on the node count and the "sweet spot" number (such as TPC-C for 100 to node count multiplied by the "sweet spot" number of users, with user load increasing by increments of 100); and run the baseline RAC benchmark.

For the second through n^{th} nodes and instances (where n is the number of nodes in the cluster), administrators should take the following steps:

1. **Duplicate the environment.** Install the OS and duplicate all of the base node's OS settings.
2. **Add the node to the cluster.** Perform node registration tasks; propagate the Oracle software to the new node; update the database spfile parameters for the new node; and alter the database to add node-specific items (such as redo logs).
3. **Run the baseline RAC benchmark.** Update the baseline benchmark criteria to include user load scenarios from the prior run's maximum up to the new maximum based on node count multiplied by the "sweet spot" number of concurrent users (and relying upon Benchmark Factory's automatic ability to balance the new user load); shut down and restart the database, adding the new instance; run the baseline RAC benchmark; and plot the transactions-per-second graph showing this run versus all the prior baseline benchmark runs—the results should show a predictable and reliable scalability factor.

As with any complex testing endeavor, the initial benchmarking setup and sub-optimization procedure can be time-consuming. In fact, nearly two-thirds of the overall effort is expended in setting up the first node and instance correctly and properly defining the baseline benchmark. However, once that initial work is completed, the remaining steps of adding nodes and retesting progresses rather quickly. In addition, if the DBA duplicates all the nodes and instances using the first node and instance, then the additional node benchmarking can be run with little or no DBA interaction (that is, steps 1 and 2 for setting up the second through n^{th} nodes and instances can be eliminated). This also provides flexibility to test various scenarios and in any order that the DBA prefers (for example, testing 10 nodes down to 1 node).

Testing the Oracle 10*g* RAC environment

In the benchmarking test case described in this article, the first three steps for setting up the first node and instance (establish a fundamental baseline, optimize the basic OS, and optimize the non-RAC database) are straightforward. The test team installed Red Hat Enterprise Linux AS 4 Update 1, the device drivers necessary for the hardware, Oracle 10*g* Release 1, and the Oracle 10.1.0.4 patch. Then, the team modified the Linux kernel parameters to best support Oracle by adding the following entries to /etc/sysctl.conf:

- kernel.shmmax = 2147483648
- kernel.sem = 250 32000 100 128
- fs.file-max = 65536
- fs.aio-max-nr = 1048576
- net.ipv4.ip_local_port_range = 1024 65000
- net.core.rmem_default = 262144
- net.core.rmem_max = 262144
- net.core.wmem_default = 262144
- net.core.wmem_max = 262144

Next, the test team performed the following steps to help ensure that asynchronous I/O feature was compiled into the Oracle binaries and is currently being used:

1. Go to the Oracle Home directory and rebuild the Oracle binaries:

```
cd $ORACLE_HOME/rdbms/lib
make -f ins_rdbms.mk async_on
make -f ins_rdbms.mk ioracle
```


2. Set the necessary spfile parameter settings:

```
disk_asynch_io = true
filesystemio_options = setall
```

The default value of `disk_asynch_io` is "true." The "setall" value for `filesystemio_options` enables both asynchronous and direct I/O.

Note: In Oracle 10*g* Release 2, asynchronous I/O is compiled in by default.

The test team then created the RAC database and initial instance using Oracle Database Configuration Assistant (DBCA), selecting parameter settings suited for the proposed maximum scalability (10 nodes). Finally, the team manually made the following spfile adjustments:

- cluster_database = true
- cluster_database_instances = 10
- db_block_size = 8192
- processes = 16000
- sga_max_size = 1500m
- sga_target = 1500m
- pga_aggregate_target = 700m
- db_writer_processes = 2
- open_cursors = 00
- optimizer_index_caching = 80
- optimizer_index_cost_adj = 40

The primary goal was to consume as much System Global Area (SGA) memory as possible within the 32-bit OS limit (about 1.7 GB). Because the cluster servers had only 4 GB of RAM each, allocating half of the memory to Oracle was sufficient—the remaining memory was shared by the OS and the thousands of dedicated Oracle server processes that the benchmark created as its user load.

Finding the "sweet spot"

The next step was to ascertain the reasonable per-node load that the cluster servers could accommodate. This is arguably the most critical aspect of the entire benchmark testing process—especially for RAC environments with more than just a few nodes. The test team initially ran the benchmark on the single node without monitoring the test via the `vmstat` command. Thus, simply looking at the transactions-per-second graph in the Benchmark Factory GUI yielded a deceiving conclusion that the "sweet spot" was 700 users per node. Although the transactions per second continued to increase up to 700 users, the OS was overstressed and exhibited minimal thrashing characteristics at about 600 users. Moreover, the test team did not temper that value by reducing for RAC overhead.

The end result was that the first attempt at running a series of benchmarks for 700 users per node did not scale reliably or predictably beyond four servers. Because each server was pushed to a near-thrashing threshold by the high per-node user load, the nodes did not have sufficient resources to communicate in a timely fashion for inter- and intra-node messaging. Thus, the Oracle database assumed that the nodes were either down or non-respondent. Furthermore, the Oracle client and server-side load balancing feature allocates connections based on which nodes are responding, so the user load per node became skewed in this first test and then exceeded the per-node "sweet spot" value. For example, when the team tested 7,000 users for 10 nodes, some nodes appeared down to the Oracle database and thus the load balancer simply directed all the sessions across whichever nodes were responding. As a result, some of the nodes tried to handle far more than 700 users—and this made the thrashing increase.

Note: This problem should not occur in Oracle Database 10*g* Release 2. With the runtime connection load-balancing feature and FAN technology, the Oracle client will be proactively notified regarding the resource availability on each node, and the client can place connections on instances that have more resources. Load balancing can be performed based on either connections or response time.

With a valuable lesson learned by the first test attempt, the test team made two major improvements. First, they reevaluated the "sweet spot" number by carefully monitoring the single-node test (in which user load increased from 100 to 800) for the onset of excessive paging, swapping, or a consistent CPU idle time near zero. The team determined that the "sweet spot" number was actually 600 users, not 700. They then reduced that number to 500 users to accommodate overhead for the RAC architecture, which would require approximately 15 percent of the system resources. This amount is not necessarily a recommendation for all RAC implementations; the test team used this amount to help yield a positive scalability experience for the next set of benchmarking tests. A less conservative "sweet spot" number could have been used if the team were able to keep repeating the tests until a definitive reduction percentage could be factually derived. Instead, the test team chose a "sweet spot" value that they expected would work well yet would not overcompensate. In addition, the team used the load-balancing feature of Benchmark Factory—which allocates one-n^{th} of the jobs to each node (where n is the number of nodes in the cluster)—to help ensure that the number of users running on any given node never exceeds the "sweet spot" value.

Increasing the user load to determine scalability

With the "sweet spot" user load identified and guaranteed through load balancing, the test team then ran the benchmark on the cluster nodes as follows:

- One node: 100 to 500 users
- Two nodes: 100 to 1,000 users

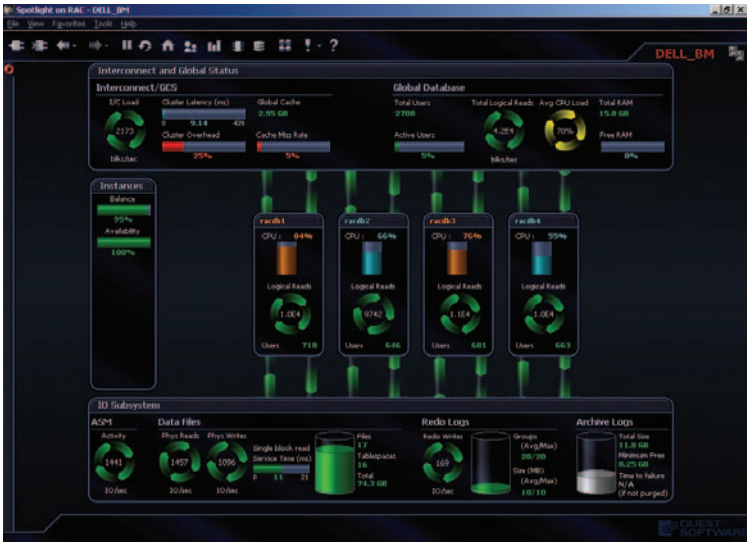


Figure 7. Spotlight on RAC GUI showing high CPU usage on nodes racdb1 and racdb3 during the four-node test

- Four nodes: 100 to 2,000 users
- Six nodes: 100 to 3,000 users
- Eight nodes: 100 to 4,000 users
- Ten nodes: 100 to 5,000 users

For each scenario, the workload was increased in increments of 100 users. The Benchmark Factory default TPC-C-like test iteration requires about four minutes for a given user load. Therefore, for the single node with five user loads, the overall benchmark test run required 20 minutes.

During the entire testing process, the load was monitored using Spotlight on RAC to identify any problems. As shown in Figure 7, when the four-node tests were conducted, Spotlight on RAC identified that CPUs on node racdb1 and racdb3 reached 84 percent and 76 percent, respectively. This high CPU utilization probably was caused by a temporary overload of users on these servers and the ASM response time. To address this problem, the test team increased the SHARED_POOL and LARGE_POOL parameters on the ASM instance from their default values of 32 MB and 12 MB, respectively, to 67 MB each. They then ran the four-node test again, and none of the nodes experienced excessive CPU utilization. This was the only parameter change the team made to the ASM instance.

Figure 8 shows the cluster-level latency graphs from Spotlight on RAC during the eight-node test. These graphs indicated that the interconnect latency was well within expectations and in line with typical network latency numbers.

Figure 9 shows the final Benchmark Factory results for all the node and user load scenarios tested. These results show that RAC scaled predictably as nodes and users were added. The scalability was near linear because the cluster interconnect generated a small amount of overhead during block transfers between instances. However, the interconnect performed well. The network interface card (NIC) bonding feature of Linux was implemented to provide load balancing across the redundant interconnects, which also helped provide availability in the case of interconnect failure.

The Dell/EMC storage subsystem that consisted of six ASM disk groups for the various data files types provided high throughput as well as high scalability. EMC PowerPath® software provided I/O load balancing and redundancy utilizing dual Fibre Channel host bus adapters (HBAs) on each server.

Note: Figure 9 shows a few noticeable troughs for the 8- and 10-node clusters. These performance declines were likely caused by undo tablespace management issues on one of the nodes in the cluster; these issues were later resolved by increasing the size of the tablespace.

The test team also monitored the storage subsystem using Spotlight on RAC. As shown in Figure 10, the Spotlight on RAC performance graphs indicated that ASM performed well at the peak of the scalability testing—10 nodes with more than 5,000 users. ASM achieved fast service times, with performance reaching more than 2,500 I/Os per second.

As the Spotlight on RAC results show, Oracle RAC and ASM performance were predictable and reliable when the cluster was

Figure 8. Spotlight on RAC GUI showing cluster-level latency graphs during the eight-node test

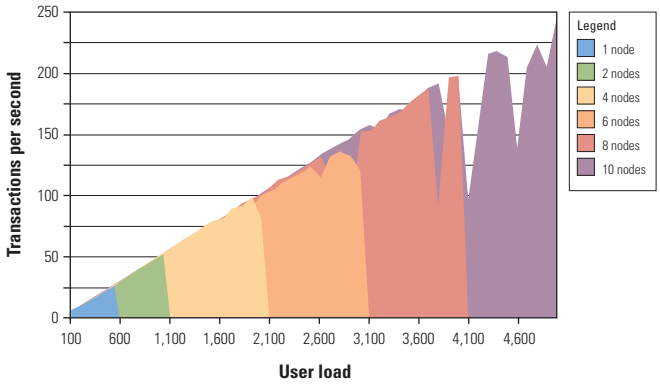


Figure 9. Benchmark Factory for Databases results for 1 to 10 RAC nodes

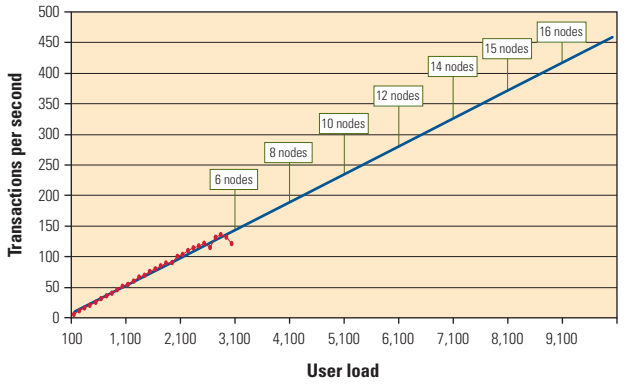


Figure 11. Projected RAC scalability for up to 17 nodes and 10,000 users

scaled horizontally. Each successive node provided near-linear scalability. Figure 11 shows projected scalability for up to 17 nodes and approximately 10,000 concurrent users based on the results of the six node scenarios that were tested. In this projection, the cluster is capable of achieving nearly 500 transactions per second.

Optimizing Oracle 10*g* RAC environments on Dell hardware

As demonstrated in the test results presented in this article, an Oracle 10*g* RAC cluster can provide excellent near-linear scalability. Oracle 10*g* RAC software running on standards-based Dell PowerEdge servers and Dell/EMC storage can provide a flexible, reliable platform for a database cluster. In addition, Oracle 10*g* RAC databases on Dell hardware can easily be scaled out to provide the redundancy or additional capacity that database environments require.

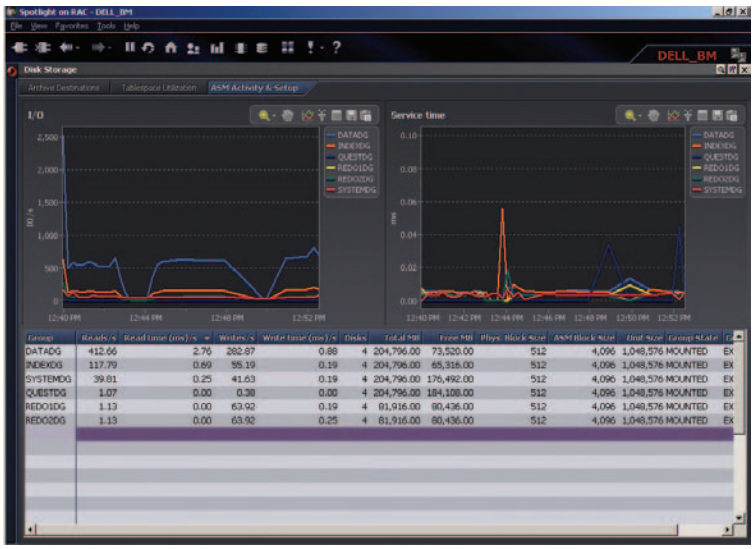


Figure 10. Spotlight on RAC GUI showing ASM performance for 10 RAC nodes

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

Zafar Mahmood is a senior consultant in the Dell Database and Applications Team of Enterprise Solutions Engineering, Dell Product Group. Zafar has an M.S. and a B.S. in Electrical Engineering, with specialization in Computer Communications, from the City University of New York.

Bert Scalzo, Ph.D., is a product architect for Quest Software and a member of the Toad® development team. He has been an Oracle DBA and has worked for both Oracle Education and Consulting. Bert has also written articles for the Oracle Technology Network, *Oracle Magazine, Oracle Informant, PC Week, Linux Journal,* and Linux.com as well as three books. His key areas of DBA interest are Linux and data warehousing. Bert has a B.S., an M.S., and a Ph.D. in Computer Science as well as an M.B.A., and he holds several Oracle Masters certifications.

Murali Vallath has more than 17 years of IT experience designing and developing databases, including more than 13 years of working with Oracle products. He has successfully completed more than 60 small, medium, and terabyte-sized Oracle9*i*™ and Oracle 10*g* RAC implementations for well-known corporations. Murali also is the author of the book *Oracle Real Application Clusters* and the upcoming book *Oracle 10*g *RAC Grid, Services & Clustering*. He is a regular speaker at industry conferences and user groups—including Oracle Open World, the UK Oracle User Group, and the Independent Oracle Users Group—on RAC and Oracle relational database management system performance-tuning topics. In addition, Murali is the president of the Oracle RAC SIG (www.oracleracsig.org) and the Charlotte Oracle Users Group (www.cltoug.org).

Best Practices:

Enterprise Testing Fundamentals

Testing is a critical aspect of making sound enterprise deployment decisions. Given a large field of possible test scenarios, administrators must decide when enough testing has occurred based on indicators of readiness for the production environment. This article, addressing fundamentals of the test process, explores best practices for enterprise testing—including recommendations for a phased approach to component, feature, and system level testing.

BY CYNTHIA LOVIN AND TONY YAPTANGCO

Related Categories:

Application development

Enterprise management

Enterprise testing

Performance

Planning

Project management

Scalable enterprise

System deployment

Testing

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

Testing enterprise solutions can be challenging, particularly when determining the right mix of approaches to apply within tight cost and schedule constraints. Effective testing can enable sound decisions about production readiness—and ineffective testing can be costly, either because of excessive testing or undetected defects that can lead to downtime and support issues. This article focuses on testing fundamentals that are key to effective enterprise testing.

The benefits of testing

To yield high-quality products, enterprises could test every feature of a product on every possible configuration using every approach or sequence that a customer might use. Of course, such an exhaustive process would take so long that the product might be obsolete before it ever reaches customers.

The key is to create a scientific, structured approach to testing—an approach that will find and fix the highest number of defects in the shortest possible time at the most reasonable cost. In short, the goal is to optimize test activities for a fast time to market while delivering a product that meets quality expectations.

An introduction to test processes

In product development, the initial view of a product is a broad overview of what a customer needs, usually in the form of product requirements generated by an enterprise's marketing team. Product developers then begin an iterative process of defining the required design. These preliminary designs are refined into detailed plans. Product modules or components are then defined and coded by the respective hardware or software engineers.

The test process can be viewed as reverse engineering of the product-development process. It is usually a three-step process consisting of component or module verification, feature verification, and usage validation.

Best practices for unit testing

The process of component or module verification is referred to as unit or structural testing. Unit testing requires knowledge of the code or circuitry—this type of testing is known as "glass box" or "white box" testing. Typically, it is performed by members of the development team.

Unit testing software and firmware

Unit testing can occur as soon as a respective component has been developed. Because this is a test of the

component only, it is not necessary for the rest of the product to be available. Therefore, unit tests can occur in parallel with the design and coding of other components. As new, untested routines are added to the component and defects are found, developers can isolate the defects and attribute them to the new routine or to its interactions with what has already been tested.[1] Unit test cases should be written in advance of, or alongside, the code modules as they are developed.

Unit testing may be used to verify that requirements and design have been implemented as well as to test data flow, checking all logic paths and error handling.[2] Unit testing is also used to augment code and design reviews on critical portions of their products.

Unit testing hardware

Unit testing for hardware is designed to find low-level flaws in chips, subassemblies, and interfaces. For example, the chip specifications may be compared to readings taken with oscilloscopes or voltage meters.[3] Hardware unit tests are performed on single components before related system components are ready to be added and verified.

Best practices for product testing

As components complete unit testing, the product reaches readiness for feature verification. This testing phase is referred to as product or integration testing. Product testing is an in-depth verification of the integrated product that includes all options of its respective features. Test cases determine whether the product's designed and implemented features and functionality meet requirements and specifications. Product testing is performed on hardware, firmware, and software and may use "glass box" or "black box" testing or both. Unlike "glass box" testing, "black box" testing does not require knowledge of the code structure or circuitry.

Product testing should be performed by an independent test team (not members of the development team) and can begin when enough interrelated, unit-tested components are available to allow test progress in one or more major areas of functionality. Product testing introduces additional complexity—for example, when customer requirements call for a software application to be supported on various server, storage, and OS platforms. Everything that a customer can do should be tested during this phase. Test engineers also should anticipate how the product might be used improperly to help ensure the product handles these types of error conditions gracefully.

		Type of test and when to use it		
Technique	**Description**	**Unit**	**Product**	**System**
Alpha testing	Allows carefully selected customers to test the partially verified product in real-world environments; provides early design feedback	Yes	Yes, early	No
Boundary testing	Attempts to overflow data structures, field lengths, and other boundary conditions	Yes, internal boundaries	Yes, all boundaries at user interface	Yes, final verification
Negative testing	Tests the product in unsupported circumstances or environments to help ensure that the user is given clear directions and that the product exits gracefully	Yes	Yes	Yes
Automated testing	Does not require human execution—a test harness or script runs unattended (useful for regression and boundary testing)	Yes	Yes	Yes
Stress and load testing	Verifies correct product operation under stress and load conditions; verifies robustness of design	No	Yes, feature level	Yes, system level
Performance testing	Verifies that product performance meets requirements, such as system boot time	No	Yes	Yes
Usability testing	Provides structured testing for an intuitive, clear user interface; may be performed with internal testers or a sample of external users similar to potential customers; should be performed as early as possible because it may drive design changes	No	Yes	Yes
Beta testing	Allows carefully selected customers to test the functionally verified product in real-world environments; validates a successful design in an almost-finished product	No	No	Yes
Reliability testing	Tests acceptable product operation over time without failure (mean time between failures) or to minimize downtime (mean time to repair); can be particularly useful for hardware testing	No	No	Yes
Acceptance testing	Uses testing performed by customers to determine product's readiness to deploy in the production environment	No	No	Yes, before release to customer
Out-of-box audits	Performs sample testing on the boxed product from the factory; may uncover packaging, manufacturing, or transportation handling flaws	No	No	Yes, during or after release to customer

Figure 1. Test types, descriptions, and phases

Ideally, unit testing and reviews in the earlier test phases will have uncovered the majority of component defects, and product testing will find defects primarily in the interfaces among these components. Each component should work well individually as demonstrated by the unit tests, but when put together, they may fail. If unit test resources or automated unit tests are available, they may continue to be used in parallel with product testing.

Best practices for system testing

System testing, the third level of testing, is a validation of the product's required usage after installation. During this phase, the test team attempts to replicate the users' operating environments.

[1] Source: *Code Complete* by Steve McConnell, published by Microsoft Press, 1993.

[2] For a "logic coverage" approach using minimal test cases, see "Structured Basis Testing" in *Code Complete* by Steve McConnell, published by Microsoft Press, 1993.

[3] Source: *Managing the Testing Process* by Rex Black, published by Wiley Publishing, 2002.

System testing occurs when all of the product features and functionality have been fully implemented. System test execution requires all components to be available in relatively stable, customer-ready revisions.

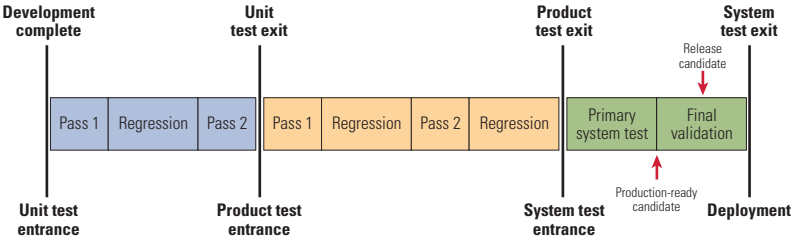


Figure 2. Phases of a typical test model

This phase often includes operating the product in a variety of environments to uncover defects. Variables that may be introduced include different communication networks and other products that will interoperate with the product under test. Products are validated under stress and load conditions. Reliability and other long-duration testing or test cases that have historically uncovered design flaws should be started as soon as possible in the system test phase (see Figure 1). This allows enough time for development teams and suppliers to respond with fixes and for the test team to verify the fixes, while preserving the desired project schedule.

System testing should help determine production readiness. To the extent possible, customer acceptance test techniques must be integrated into system testing. This can facilitate the acceptance process at the customers' sites. If unit and product test resources or automated regression tests are available, they should continue to be run in parallel with system testing.

Repetition and regression testing

A primary purpose of any testing is to uncover defects in the product and to characterize the circumstances that allow these defects to surface. As defects are discovered and reported, development engineers analyze the defects, develop fixes for problems, and provide those fixes back to the test team for verification.

Test engineering strategy must incorporate some repetition of core functionality testing during defect fix verification. These fixes should resolve the reported problems and not cause any other issues in adjacent functionality. This process is referred to as regression testing. Regression testing usually occurs at the end of the execution of a test phase. Depending on the number of fixes that have been incorporated during the test phase and the extent of any design changes required to fix defects, significant portions of the testing may need to be run again in a new test pass.

A typical test model, such as the one shown in Figure 2, incorporates the elements of each test phase and multiple test passes. Many specific test methods or techniques exist within the broader categories of unit, product, and system testing. Figure 1 describes some of these techniques and gives guidelines for applying them.

When developing their test processes, enterprises should consider the following factors:

- Who will perform the testing
- Combinations of possible configurations to be tested or bypassed
- A strategy for developing test cases
- Time allotments for each test phase
- The sequence of test phases (serial or parallel)
- When to start and complete testing
- Appropriate measurements of product quality
- Appropriate measurements of test efficiency and adequacy
- Tools to help manage the test process and improve its efficiency
- A defect management system and strategy

The optimal test process

The test process consists of iterative verification for each level of product integration and complexity. A certain amount of repetitive testing should be built into the test model to help ensure that previously tested features and functionality are not adversely affected by product and code changes during the test cycle. Knowing when to apply various testing techniques and methods in the test process is fundamental to the testing success. In addition, testing should enhance product quality while minimizing resource costs and time to market.

Cynthia Lovin is a senior consultant test engineer in the Dell Product Group Global Test Department, with 16 years of test- and quality-engineering experience in companies ranging from startups to Fortune 500 enterprises. She has Six-Sigma Green Belt certification and a B.A. in Business Administration from The University of Texas at Austin.

Tony Yaptangco is the director of the system test group within the Dell Product Group Global Test Department. He has 25 years of experience in various software and test environments, including 16 years in engineering management positions. Tony has a B.S. in Computer Science from San Diego State University and an M.S. in Engineering Management from National Technological University.

FOR MORE INFORMATION

IEEE Software Engineering Standards Zone:
www.standards.ieee.org/software

VeriTest Testers' Network:
www.veritest.com/tn

Software Quality Engineering:
www.sqe.com

American Society for Quality:
ww.asq.org

Building Clustered Enterprise Applications

with JBoss Application Server on the Dell PowerEdge 1855 Blade Server

JBoss Application Server (JBoss AS) is a standards-based Java platform for scalable enterprise applications. The Dell™ PowerEdge™ 1855 blade server can provide a cost-effective system for hosting scaled-out applications on the JBoss AS platform. To demonstrate the ease of migrating applications to this platform, a team of engineers from Dell and JBoss ported a Web application to JBoss AS on a PowerEdge 1855.

BY TODD MUIRHEAD; DAVE JAFFE, PH.D.; NORMAN RICHARDS; AND SHAUN CONNOLLY

Related Categories:

Clustering

Dell PowerEdge blade servers

JBoss

Linux

Novell SUSE

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

JBoss Application Server (JBoss AS) has emerged as a popular Java 2 Platform, Enterprise Edition (J2EE) application server for enterprise data centers. The cornerstone of JBoss Enterprise Middleware Suite (JEMS), JBoss AS is designed to help enterprise IT departments build and deploy revenue-generating applications. JBoss AS provides a standards-based platform for scalable enterprise applications—with a zero-cost software license—enabling organizations to scale out applications without incurring prohibitive per-CPU licensing costs.

The cost-effective, easy-to-manage Dell PowerEdge 1855 blade server is an excellent platform for JBoss AS. The PowerEdge 1855, now a certified platform for JBoss AS,[1] supports up to 10 server blades in a 7U chassis. Each blade can be configured with up to two Intel® Xeon® processors with Intel Extended Memory 64 Technology (EM64T) designed to support the latest 64-bit operating systems.

The certification of JBoss AS on the Dell PowerEdge 1855 exemplifies Dell support for open source software. The combination of JBoss AS and MySQL database from MySQL AB, both of which are included with Novell® SUSE LINUX Enterprise Server 9, can provide a cost-effective, Dell-supported platform for enterprise applications. The Dell and JBoss relationship further extends the Professional Open Source model, which combines the cost-savings benefits of open source with the development methodologies, support, and accountability expected from leading hardware and software vendors.

To demonstrate the ease of building applications on the JBoss AS platform, in July 2005 a team of engineers from Dell and JBoss ported a Web application—which uses JavaServer Pages (JSP) to implement the front end of an online DVD store—to JBoss AS running on two PowerEdge 1855

[1] JBoss Inc. certified JBoss AS software to run on Dell PowerEdge servers, including the Dell PowerEdge 1855.

blade servers in less than a week. In the process, the team added features to the Web application, including secure sign-on, shopping-cart persistence, clustering, failover, and internationalization.[2] The MySQL database back end required no modifications to be ported to the JBoss AS platform. This article provides an overview of the features provided by JBoss AS—with a focus on the features added to the DVD store application when ported to JBoss AS—as well as details of the JBoss AS DVD store implementation.

Understanding JBoss AS

JBoss AS is a standards-based, J2EE 1.4 certified application server providing the foundation upon which enterprises, software vendors, integrators, and solution providers can build applications and Web services. Application platforms like JBoss AS are typically referred to as middleware and are designed to simplify the development of business applications by providing a layer of software infrastructure that abstracts the low-level operating systems, communication protocols, and hardware details.

Companies rely on JBoss AS for deploying scalable and secure Web applications that are composed of complex business logic and serve thousands of concurrent requests. Because JBoss AS is founded on a service-oriented microkernel architecture, it is designed to provide services in a plug-and-play fashion, including all J2EE 1.4 services such as Enterprise JavaBeans (EJB), JSP, servlets, and Web services, along with extended enterprise services for clustering, caching, failover, persistence, and distributed deployment. This service orientation allows small, medium, and large enterprises to standardize on JBoss AS because it is designed to scale from single, low-end server configurations to clustered environments that contain hundreds of high-end servers.

JBoss AS is designed to provide a complete platform for building a Web application's user interface, business logic, and data access logic. This article and the DVD store demonstration application are focused on the following features:

- Ease of development
- Data persistence
- Application security
- Clustering and failover
- Internationalization

Ease of development

JBoss AS enables developers to leverage enterprise features without undue complexity—as reflected in the design of features such as clustering, which requires no application changes or design-time code modifications. Moreover, JBoss AS 4.0 continues to simplify development by including support for two important technologies: JavaServer Faces (JSF) and ESB 3.0.

JSF provides a standards-based framework for handling a Web application's presentation layer, and EJB 3.0 provides a simple programming model for the business-logic layer and data-access layer. Both JSF and EJB 3.0 are critical components of the Java Platform, Enterprise Edition 5 (Java EE 5) standard, which is designed to directly address the needs of developing enterprise-class applications (including user interface, business logic, data access, and persistence) in a dramatically simplified manner. By enabling developers to isolate their business logic from the user interface and data-access logic, JBoss AS helps simplify development, improve application maintenance, and enable IT organizations to deliver high value in a relatively short period of time.

Application platforms like JBoss AS are typically referred to as middleware and are designed to simplify the development of business applications by providing a layer of software infrastructure that abstracts the low-level operating systems, communication protocols, and hardware details.

Data persistence

Because JBoss AS supports the EJB 3.0 specification, it inherently provides a highly flexible and productive mechanism for storing Java objects and EJB components in relational database tables. JBoss AS not only supplies high-performance access to data, it also enables the application to support any relational database for its back-end data. This support for transparent data persistence means that although the DVD store application uses MySQL as its database, other relational databases such as an Oracle® or Microsoft® SQL Server™ database can be used without any changes to the application.

Application security

Because application servers are platforms for enterprise applications, they are expected to provide a wide range of mission-critical services and security features. JBoss AS supports the Java Authentication and Authorization Service (JAAS) standard application programming interface (API), which is designed to provide a seamless security architecture across J2EE applications. Using simple

[2] For more information about the original JSP-based application, see the "Three Approaches to MySQL Applications" presentation by Dave Jaffe and Todd Muirhead, available at conferences.oreillynet.com/cs/mysqluc2005/view/e_sess/6226.

declarative security statements, developers can restrict J2EE application access to authenticated users. Moreover, security can be externally configurable so there are no explicit security checks within an application.

Clustering and failover

JBoss AS achieves scalability and fault tolerance through its clustering technology, which makes it suitable for deployment across large numbers of servers. The clustering technology is designed to be transparent to the application, so cluster nodes automatically discover one another on boot-up—with no additional configuration. Any application can be made to run on a JBoss AS cluster, and clustering can be activated by changing a JBoss AS configuration setting. Doing so is enough to enable load balancing, state replication, and failover for an application's Java beans.

Internationalization

Organizations creating globally deployed applications require a platform that is designed for internationalization, such as JBoss AS. By using JSF, developers can create applications that target a global audience. JSF isolates the user interface's text to help simplify the localization of the application to a specific language.

Implementing the DVD store application

The Dell and JBoss team implemented the DVD store application using the standard Java EE 5 technologies available in JBoss AS 4.0.3. This section describes the implementation and configuration of the application running on JBoss AS.

A multi-tiered architecture

The DVD store application uses a standard multi-tiered architecture consisting of a database tier, an EJB 3.0 tier, and a Web tier. Although the Web and EJB 3.0 tiers are logically separate layers, they are colocated to maximize performance. This colocation is the recommended deployment strategy for applications built on JBoss AS.

Figure 1 shows the overall system architecture used in the demonstration implementation. JBoss AS 4.0.3 was deployed on two Dell PowerEdge1855 blade servers. The DVD store application was deployed on each JBoss AS instance, with JBoss Cache providing the clustering link between them. Additional PowerEdge 1855 nodes could have been added to the cluster with no additional JBoss AS configuration required. Both JBoss AS instances communicated with a MySQL database running on a Dell PowerEdge 2800 server.

A front-end load balancer is required to provide HTTP failover between nodes. A hardware load balancer is generally preferred; however, JBoss AS also supports the use of the Apache HTTP server

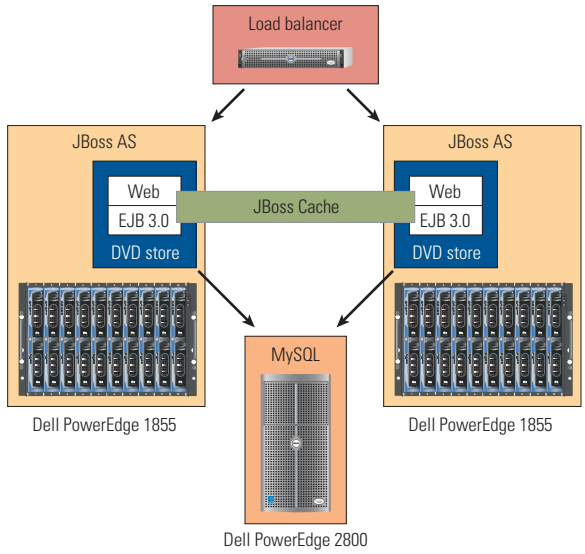


Figure 1. System architecture for the DVD store Web application

with the mod_jk connector to provide software load balancing.

Database tier. MySQL provided the database tier in the implementation described in this article (see the "The DVD store MySQL database" sidebar in this article for database details). The connection between JBoss AS and MySQL is managed in JBoss AS using a standard J2EE Connector Architecture (JCA) connection pool. The connection pool allows for fast database access and removes the connection management code from the application.

EJB 3.0 tier. The EJB 3.0 tier provides the data management and core business logic features of the application. It does so through EJB 3.0 entity beans, which provide the mapping between Java objects and data in the database, and stateless session beans, which provide the business logic that the application will use.

Compared to earlier EJB versions, EJB 3.0 entity beans provide a greatly simplified method of mapping relational data to the database. The entity beans are implemented as lightweight plain old Java objects (POJOs) with minimal use of annotations to provide persistence details. Unlike EJBs under earlier J2EE versions, no special interfaces, magic methods, or XML configurations are needed.

The Dell and JBoss team created one EJB 3.0 entity bean for each table in the database.[3] The application provides table and column mapping details but is otherwise relieved of the burden of managing the object-relational mapping.

EJB 3.0 persistence eliminates the need for an application to manage SQL queries. Figure 2 lists the code required to load the orders for a customer from the database.

Although native SQL queries can be used to optimize performance, the application as written is portable to any database that

[3] For code of the EJB 3.0 entity bean that corresponds to the PRODUCTS table, see the supplemental online section of this article at www.dell.com/powersolutions.


```
public List<Order> getOrders(Customer customer) {
    return
        em.createQuery(
                "from Order o where o.customer = :customer")
          .setParameter ("customer", customer)
          .getResultList();
}
```

Figure 2. Java code for loading orders for a customer from the database

JBoss AS supports. No application code must be rewritten to switch to a different database.

The application provides a single stateless session bean that provides the core business logic of the application. The DvdStoreBean provides the logic for data lookup, such as the getOrders() method, inventory management, and purchase processing. Figure 3 shows the external interface to the DVD store session bean.

Like EJB 3.0 entity beans, session beans are similarly lightweight and require no heavy interfaces or external XML configuration files to take advantage of enterprise aspects such as security and transactions. For example, the purchase method of the DvdStoreBean processes the order and creates all of the objects that must be written to the database.[4] It also handles the application-specific requirements when a transaction cannot be completed because of insufficient quantities of a DVD.

Web tier. Although the business logic is provided by the EJB 3.0 tier, the Web tier assembles the logic and presents a coherent application to the user. The DVD store uses a combination of JSF and JSP to manage the user experience.

Using a JSF component–based architecture has many advantages. The model-view-controller (MVC) architecture allows for a clean separation of the Web application responsibilities. The JSF backing beans provide a rich model for the interface, and they map user requests onto the EJB 3.0 business logic. The JSP pages provide a simple view with no embedded business logic and minimal HTML and HTTP details exposed. JSF also provides powerful internationalization features that externalize text for easy localization of the application in a specific language.

Externalized security

Security is externalized in the DVD store application—it has no explicit security checks. The Dell and JBoss team defined a JAAS domain that could understand the usernames and passwords in the relational database. Using simple declarative security statements, developers can restrict application access to authenticated users. JAAS enables developers to replace the security domain with one that consults a Lightweight Directory Access Protocol (LDAP) server or makes use of another authentication technology without changing any application code.

Clustering support

The Dell and JBoss team added clustering support to both the Web tier and the EJB 3.0 tier of the DVD store application. The application uses HTTP session replication and clustered single sign-on in the Web tier. If any JBoss AS node were to fail, a front-end HTTP load balancer would fail the request over to another JBoss AS server. This application server would be in the right state to process the request, and the user would continue using the application with no noticeable interruption of service.

Additionally, a second-level clustered entity cache in the DellJBoss implementation maintained primarily read-only data such as product categories. Once this data was loaded into the cache, the EJB 3.0 tier no longer needed to consult the database to load that data. The DVD store application did not need to be changed in any way to support the cache. All clustering options in JBoss AS make use of the JBoss Cache/JGroups stack and require multicasting communication between the nodes.

Porting applications to the JBoss AS platform

The DVD store application was ported in less than a week, with no changes required to the database layer. Through the use of

```
public interface DvdStore
{
    public Customer getCustomer(String user);
    public List<Order> getOrders(Customer customer);
    public List<Product> getRecentHistory(
           Customer customer, int howmany);
    public List<Product> getProducts();
    public List<Product> searchProducts(String title,
           String actor, Category category, int howmany);
    public Order purchase(Customer customer,
           List<OrderLine> lines)
       throws InsufficientQuantityException;
    public List<Category> getCategories();
}
```

Figure 3. Java code for the external interface to the DVD store session bean

[4] For code of the purchase method of the DvdStoreBean session bean, see the supplemental online section of this article at www.dell.com/powersolutions.

THE DVD STORE MYSQL DATABASE

The back-end database employed in the Dell-JBoss implementation was a large MySQL database (100 GB total size), representing an online DVD store with 1 million DVD titles, 200 million customers, and 120 million orders. Advanced database features such as transactions and referential integrity constraints were employed.

The MySQL DVD store database comprised seven main tables and one small table (see Figure A).

The CUSTOMERS table was prepopulated with 200 million customers: 100 million U.S. customers and 100 million customers from the rest of the world. The ORDERS table was prepopulated with 10 million orders per month for a full year. The ORDERLINES table was prepopulated with an average of five items per order. The PRODUCTS table contained 1 million DVD titles, each with a principal actor listed for search purposes. For realism, titles and actor names were generated by taking combinations of actual movie titles and actor names. Additionally, the CATEGORIES table contained the 16 DVD categories.

The schema is fully documented in the database build script, which can be found in the supplemental online section of this article at www.dell.com/powersolutions. The complete MySQL DVD store code (as well as implementations for Oracle and Microsoft SQL Server) is available under the open source GNU General Public License (GPL) from linux.dell.com/dvdstore.

Table	Columns	Number of rows
CUSTOMERS	CUSTOMERID, FIRSTNAME, LASTNAME, ADDRESS1, ADDRESS2, CITY, STATE, ZIP, COUNTRY, REGION, EMAIL, PHONE, CREDITCARDTYPE, CREDITCARD, CREDITCARDEXPIRATION, USERNAME, PASSWORD, AGE, INCOME, GENDER, PROD_ID_IDX, PROD_ID1, PROD_ID2 … PROD_ID10	200 million
ORDERS	ORDERID, ORDERDATE, CUSTOMERID, NETAMOUNT, TAX, TOTALAMOUNT	120 million
ORDERLINES	ORDERLINEID, ORDERID, PROD_ID, QUANTITY, ORDERDATE	600 million
CUST_HIST	CUSTOMERID, ORDERID, PROD_ID	600 million
PRODUCTS	PROD_ID, CATEGORY, TITLE, ACTOR, PRICE, SPECIAL, COMMON_PROD_ID1, COMMON_RATING1, COMMON_PROD_ID2, COMMON_RATING2, COMMON_PROD_ID3, COMMON_RATING3	1 million
INVENTORY	PROD_ID, QUAN_IN_STOCK, SALES	1 million
REORDER	PROD_ID, DATE_LOW, QUAN_LOW, DATE_REORDERED, QUAN_REORDERED, DATE_EXPECTED	Variable
CATEGORIES	CATEGORY, CATEGORYNAME	16

Figure A. The DVD store database schema

EJB 3.0 entity beans to map the relational database data to application objects, no changes to the application would be needed for it to work with any database supported by JBoss AS, including Microsoft SQL Server and Oracle. The use of stateless session beans simplified the implementation of the application's business logic. The Web front end, implemented with JSF, can be modified in appearance or language. Built-in JBoss AS features such as JAAS and JBoss Cache provide secure sign-on and clustering. All of these features were built and successfully tested on two PowerEdge 1855 blade servers, with the MySQL database running on a PowerEdge 2800 server.

Todd Muirhead is a senior engineering consultant on the Dell Scalable Enterprise Technology Center team. He specializes in storage area networks, virtualization, and database systems. Todd has a B.A. in Computer Science from the University of North Texas and is Microsoft Certified Systems Engineer + Internet (MCSE+I) certified.

Dave Jaffe, Ph.D., is a senior consultant on the Dell Scalable Enterprise Technology Center 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 and the Dell Technology Showcase. He has a B.S. in Chemistry from Yale University and a Ph.D. in Chemistry from the University of California, San Diego.

Norman Richards is a developer at JBoss, Inc. He is the co-author of *JBoss: A Developer's Notebook* and *XDoclet in Action.*

Shaun Connolly is the vice president of product management at JBoss, Inc. Shaun is responsible for the products that make up JEMS. Prior to joining JBoss, Shaun served in vice president– and director-level product management and product development positions at Princeton Softech, HP Middleware, Bluestone Software, and Primavera Systems. Shaun has a B.S. in Electrical Engineering from Drexel University and has been a panelist, speaker, and contributor of articles on a wide range of IT topics.

FOR MORE INFORMATION

Dell and JBoss:
www.dell.com/jboss
www.jboss.com/dell

VMware ESX Server Performance

on Dell PowerEdge 2850 and PowerEdge 6850 Servers

VMware® ESX Server™ software on Dell™ PowerEdge™ servers enables a virtualization platform that can support a wide variety of applications and operating systems. Dell engineers tested this software on Dell PowerEdge 2850 and PowerEdge 6850 servers to show how ESX Server can manage multiple workloads.

BY TODD MUIRHEAD; DAVE JAFFE, PH.D.; AND SCOTT STANFORD

Related Categories:

Dell PowerEdge servers

Dell/EMC storage

Linux

Microsoft Exchange

Microsoft Windows Server 2003

MySQL

Novell SUSE Linux

Performance

Virtualization

VMware

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

VMware ESX Server software enables many different applications to run on the same hardware at the same time. ESX Server provides the virtualization layer that resides just above the hardware to create virtual machine (VM) containers. Each VM runs its own OS, which in turn has its own set of applications and services. Because each VM is isolated by ESX Server from the other VMs in the same way that physical systems are isolated from each other, ESX Server offers flexibility for concurrently running multiple types of applications and operating systems.

Each individual VM can be rebooted or powered down without affecting other VMs running on the same physical server. Because each application runs in its own VM, an application can be patched or upgraded and the VM rebooted without bringing down all other applications running on that physical server.

In addition, VMware ESX Server supports VMotion™ technology, which allows for a running VM to be moved from one physical server to another physical server without any downtime for the VM. In a pool of servers running ESX Server software and using VMotion, VMs can be moved around for load balancing or maintenance without affecting end user applications, because no downtime occurs during the movement of VMs. To enable this feature, administrators must connect all servers involved in the move to the same shared storage in the storage area network (SAN). During a VMotion move, only the RAM contents of the VM or VMs being moved are copied across the network because the VM's disk is on the shared storage and is already accessible by all servers.

In a typical ESX Server farm, many different operating systems and applications run concurrently on the same physical server. To measure the performance and scalability of ESX Server on eighth-generation Dell PowerEdge servers, Dell engineers modeled and tested a variety of workloads. These tests were conducted in July 2005.

Setting up the test systems

VMware ESX Server 2.5.1 software was installed on both Dell PowerEdge servers used in testing, the PowerEdge 2850 and PowerEdge 6850, instead of a traditional OS such as the Microsoft® Windows Server™ 2003 or Red Hat® Enterprise Linux® 4 OS. The fastest processors available for each system at the time of testing in July 2005 were used. The PowerEdge 2850 used two Intel® Xeon® processors DP at 3.6 GHz with 2 MB level 2 (L2) cache. The Dell team tested two different processor configurations for the PowerEdge 6850: four Intel Xeon processors MP at 3.66 GHz with 1 MB L2 cache and four Intel Xeon processors MP at 3.33 GHz with 8 MB level 3 (L3) cache.

Each server was connected to the SAN with two QLogic SANblade QLA2340 host bus adapters (HBAs). ESX Server provides failover across the multiple paths to logical units (LUNs) through both HBAs. SAN storage was provided by

a Dell/EMC CX700 array. The VMs used in the testing were spread across forty-eight 73 GB 10,000 rpm Fibre Channel disks located in the CX700 array. Eight five-disk RAID-5 LUNs were created. Four of these LUNs were assigned as VMs for the Microsoft SQL Server™ database server; the LAMP (Linux, Apache, MySQL, and PHP) Web platform; the NetBench benchmarking tool; and Microsoft Exchange Server. Additionally, each Exchange VM used one of the eight LUNs for the mail store data drives. The final eight drives were configured as four two-disk RAID-1 transaction log volumes for the Exchange VMs.

Understanding the four VM-testing workloads

To simulate typical application usage in VMs on ESX Server, Dell engineers tested the PowerEdge servers by increasing the number of VMs for four different workloads until CPU utilization for the entire server reached 85 percent. A utilization level of 85 percent was chosen as a reasonably high level of server usage that might be reached in production but is well below the maximum 100 percent utilization that is used in many industry-standard benchmarks and is hopefully not reached in production.

The four workloads used to test virtualization on the servers were Microsoft SQL Server 2000 with an online transaction processing (OLTP) database, SUSE Linux with a LAMP stack, Microsoft Windows Server 2003 with NetBench 7.03, and Microsoft Exchange Server 2003 with LoadSim 2003. Initially, each workload was tested individually. To accomplish this, Dell engineers set up a workload in multiple VMs, keeping all settings on the VMs and driver systems the same, and then observed how many VMs could run concurrently under the same load—thus measuring the number of VMs the server could support. Figure 1 shows the configuration of the VMs used in all four workloads.

Microsoft SQL Server 2000

The SQL Server 2000 VMs used in testing were the same VMs used in two previous Dell studies (the test team used VMotion to move the VMs to the servers for these tests).[1] These VMs had an OLTP database simulating the database back end of a simple Web-based DVD store front. The database size of about 1 GB is small and representative of a database used for development or testing.

To simulate a load against the VM, Dell engineers created a C program to make connections to the database and to simulate the actions of a user, but with a very small think time between actions. This small think time enables a single driver system to use a small number of connections to simulate the workload of a large amount of users. The stress and connectivity patterns found in the combination of small think times and the single driver system is similar to how the database would be used if a middle-tier application server was used to pool the connections to the database. For this test, each SQL Server 2000 VM was driven by a single thread of the driver application with a 20 millisecond (ms) delay.

SUSE LAMP

To add variation to the testing and to underscore that VMware ESX Server can support Linux VMs as well as Microsoft Windows® VMs, Dell engineers built the DVD store database and Web-based application on the LAMP stack.[2] All components of this LAMP stack[3] are available through open source. Although in this test all stack layers were contained in a single VM, the Apache/PHP Web layer can easily be moved to one or more VMs, and the MySQL database can run on a master VM and one or more slave VMs using MySQL replication.

The DVD store used database features such as foreign keys, transactions, and full-text searching that were not used in the SQL Server VM workload. Using these features caused the overall orders per minute to decrease in relation to the additional work performed by the database for each order, because of the additional CPU overhead needed for these features.

The driver for the LAMP stack differs from the driver used for the SQL Server testing in that it emits HTTP requests and receives HTML code returned from the Apache/PHP layer, whereas the SQL Server driver communicates directly with the database. However, both drivers measure the same parameters: the total orders per minute handled by the application and the average response time experienced by the simulated customers. For this test, each SUSE LAMP VM was driven by a single thread of the driver program with a 30 ms delay.

NetBench 7.03

NetBench 7.03 is a benchmark tool designed to simulate a file server load. A set of files is created and accessed according to predefined

Workload	RAM	Disk	Virtual NIC	Virtual CPU
Microsoft SQL Server 2000	512 MB	10 GB	Vmxnet	1
SUSE LAMP	1,024 MB	10 GB	Vlance	1
NetBench 7.03	512 MB	10 GB	Vmxnet	1
Microsoft Exchange Server 2003	512 MB	10 GB for the Windows OS 130 GB for data 10 GB for logs	Vmxnet	1

Figure 1. Configuration of the VMs used in testing

[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, www.dell.com/downloads/global/power/1q04-jav.pdf; and "Evaluating Price/Performance of VMware ESX Server on Dell PowerEdge Servers" by Todd Muirhead; Dave Jaffe, Ph.D.; and Felipe Payet in *Dell Power Solutions,* June 2004, www.dell.com/downloads/global/power/ps2q04-006.pdf.

[2] The complete DVD store application code, including a SQL Server and a LAMP version, is freely available for public use under the GNU General Public License (GPL) at linux.dell.com/dvdstore.

[3] The LAMP stack has been fully documented in the Dell Enterprise Product Group white paper "MySQL Network and the Dell PowerEdge 2800: Capacity Sizing and Performance Tuning Guide for Transactional Applications" by Todd Muirhead, Dave Jaffe, and Nicolas Pujol, www.dell.com/downloads/global/solutions/mysql_network_2800.pdf.

Virtualization software	VMware ESX Server 2.5.1
CPU	Two Intel Xeon processors DP at 3.6 GHz with 2 MB L3 cache
Memory	8 GB
Internal disk	Two 73 GB drives
NIC	Two 10/100/1000 Mbps internal NICs One Intel PRO/1000 XT NIC
Disk controller	Dell PowerEdge Expandable RAID Controller 4, embedded internal (PERC 4/ei)
Fibre Channel HBA	QLogic SANblade QLA2340
Form factor	2U (3.5 inches)

Figure 2. Configuration of the PowerEdge 2850 server used in testing

scripts. Typically, NetBench is run with an increasing number of client engines making requests against a single server to measure the throughput (in megabytes per second) that can be achieved with a given number of connections.

To test the number of VMs on an ESX server, Dell engineers increased the number of VMs and the number of client engines at the same rate, until the CPU utilization for ESX Server reached 85 percent. NetBench 7.03 with the included standard DiskMix script was used with an 0.6 second think time to connect two client engines to each VM. This setup simulated multiple file servers hosted on the same ESX Server, similar to a file-server consolidation scenario. The driver systems on which the client engines ran mapped drives to all VMs in the test. In NetBench, the test-directories path file was modified so that, as successive client engines were added, they would use the next drive letter, which corresponded to the next VM.

Microsoft Exchange Server 2003

Similar to the other tests, the objective of the Exchange Server 2003 test was to scale the number of simulated Microsoft Outlook® mail user connections that an ESX Server host could support by scaling the number of concurrently running Exchange Server 2003 VMs.

LoadSim 2003 is a very effective tool for analyzing how a back-end Exchange Server 2003 system will perform under typical workday messaging loads. The program simulates Outlook Messaging Application Programming Interface (MAPI)–based user profiles and actions that are found in typical corporate messaging environments.[4] Because the I/O generated by LoadSim 2003 mimics real-world Outlook user activities, the Exchange Server 2003 VM CPU, memory, network, and disk I/O subsystems were all subjected to moderately heavy random workloads. These workloads thereby stressed and utilized all of the ESX Server host-based virtualized hardware components. The scaling results are a good indicator of how a messaging-server consolidation strategy scales on an ESX Server–virtualized PowerEdge server configuration.

For the other three workloads, a large number of small VMs were used; but for the Exchange Server 2003 workload, a small number of large VMs was modeled. Given the typical memory and storage requirements for back-end Exchange servers, it seemed unlikely that a large number of small Exchange VMs would be used in the same way as large numbers of file servers, small SQL Server databases, or LAMP stack applications. Using large VMs in one test also provided another dimension to the testing by including VMs with greater disk, RAM, and processor requirements per VM.

Testing ESX Server on PowerEdge 2850 servers

The Dell PowerEdge 2850 is a 2U dual-processor server with an 800 MHz frontside bus that can use up to 12 GB of RAM. In the Dell tests described in this article, the PowerEdge 2850 was configured with 8 GB of RAM. The PowerEdge 2850 can support six internal disks, offers three PCI slots, and includes dual on-board Gigabit Ethernet[5] network interface cards (NICs). Two of the PCI slots were used in the test environment for QLogic SANblade QLA2340 HBAs to provide Fibre Channel connectivity to the SAN. An Intel Gigabit Ethernet NIC was used in the final slot in addition to the two on-board Gigabit Ethernet NICs, making the total number of NICs in the system three. This configuration allowed one NIC to be used for the ESX Server service console, one for the VMs, and one for VMotion. Figure 2 summarizes this configuration.

Results for the four workloads

Figure 3 displays the results for all four workloads running on the PowerEdge 2850. The SQL Server, SUSE LAMP, and NetBench test results all show that approximately the same number of VMs could be supported at approximately the same ESX Server CPU utilization. Even though each of these three workloads was quite different, the number of VMs that could be run was about the same.

In the case of the SUSE LAMP VMs, ESX Server was overcommitted for RAM. Each of the 10 SUSE LAMP VMs was assigned 1 GB of RAM, but the PowerEdge 2850 was configured with only 8 GB of RAM. ESX Server has a swap file available if the memory overcommitment at the VM level causes the server to actually run

Workload	Number of VMs on PowerEdge 2850	ESX Server CPU utilization
Microsoft SQL Server 2000	11	86%
SUSE LAMP	10	87%
NetBench 7.03	11	85%
Microsoft Exchange Server 2003	3	63%

Figure 3. Results of workload tests on the PowerEdge 2850 server

[4] For more information about LoadSim 2003, see the Dell Enterprise Product Group white paper "MMB3 Comparative Analysis" by Scott Sanford, www.dell.com/downloads/global/solutions/MMB2_Comparison_with_MMB3.doc.

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

Virtualization software	VMware ESX Server 2.5.1
CPU	Four Intel Xeon processors MP at 3.66 GHz with 1 MB L2 cache or Four Intel Xeon processors MP at 3.33 GHz with 8 MB L3 cache
Memory	16 GB
Internal disk	Two 73 GB drives
NIC	Two 10/100/1000 Mbps internal NICs Two Intel PRO/1000 XT NICs
Disk controller	Dell PERC 4, Dual Channel (PERC 4/DC)
Fibre Channel HBA	QLogic SANblade QLA2340
Form factor	4U (7 inches)

Figure 4. Configuration of the PowerEdge 6850 server used in testing

out of memory. However, because of the way that ESX Server shares common memory pages between VMs and manages active memory, no swapping occurred during the testing.

The Exchange Server test was quite different from the other three workload tests, in that each VM supported 1,000 LoadSim 2003 users and used 2 GB of RAM. To support the I/O requirements of these users, Dell engineers added another data LUN, whereas the other workloads were able to support their I/O requirements with the same disk used for the OS. The large size of the Exchange Server VMs also did not provide a level of fine granularity that could have helped reach the desired 85 percent ESX Server CPU utilization level. With four Exchange VMs, the ESX Server CPU utilization was at 100 percent on the PowerEdge 2850.

The test results show that the PowerEdge 2850 was able to handle different types of workloads equally well in a virtualized environment. Additionally, the high performance of the dual-processor PowerEdge 2850 makes it well suited for scaling out VMware ESX Server farms. Using the PowerEdge 2850 as the building block of an ESX Server farm can be cost-effective for adding capacity because of its low acquisition cost. The excellent performance results of the PowerEdge 2850 demonstrate its suitability in virtualization environments that do not require more than three PCI slots for connectivity (NICs and Fibre Channel) and do not have large RAM requirements (greater than 12 GB per system).

Testing ESX Server on PowerEdge 6850 servers

The Dell PowerEdge 6850 is a 4U server that supports up to four Intel Xeon processors MP and 32 GB of RAM. Processor cache sizes available on the PowerEdge 6850 are 1 MB L2 and 8 MB L3. The 8 MB L3 cache enables better performance than the 1 MB L2 cache; however, processors with a 1 MB cache typically are less expensive than those with an 8 MB cache.

The PowerEdge 6850 has seven I/O slots and an optional QLogic SANblade QLA2362 Fibre Channel daughtercard. Four of the slots are PCI Express, and three are PCI Extended (PCI-X). The large number of slots provides room for additional NICs, which are often needed for ESX Server environments with high network bandwidth or a large number of isolated networks. If the QLogic daughtercard is used for Fibre Channel connectivity, all slots are left free for NICs or other connectivity options. Hot-plug PCI Express and memory RAID features on the PowerEdge 6850 are designed to increase the server's overall availability. Administrators can replace or add PCI cards without having to power down the server, and memory RAID allows the server to continue operating despite a memory failure.

In the Dell tests, the PowerEdge 6850 was configured with 16 GB of RAM and two PCI-X–based Intel Gigabit Ethernet NICs in addition to the system's two internal Gigabit Ethernet NICs. Two PCI-X–based QLogic QLA2340 adapters also were used for Fibre Channel connectivity. Figure 4 summarizes this configuration.

Results for the four workloads

The PowerEdge 6850 provided excellent performance in a single system and high capacity in terms of I/O and RAM. ESX Server running with all four workloads performed better on the PowerEdge 6850 using 3.33 GHz processors with 8 MB L3 cache, but the PowerEdge 6850 using the less-expensive 3.66 GHz processors with 1 MB L2 cache was able to host at least 75 percent of the number of VMs as the server using 8 MB cache processors. Figure 5 shows the complete results of the PowerEdge 6850 testing under different workloads.

During the Exchange Server testing on the PowerEdge 6850 with the 3.33 GHz, 8 MB cache processors, Dell engineers again encountered challenges in achieving the 85 percent CPU utilization level; however, these challenges differed from those encountered with the less resource-intensive VM stacks in the PowerEdge 2850 Exchange Server test. In an effort to increase CPU utilization, the test team upgraded the data disk from a 5-disk RAID-5 configuration to a 10-disk RAID-10 configuration, doubled the number of LoadSim 2003 drivers, and increased the number of LoadSim 2003 users from 4,000 to 4,400. None of these actions significantly raised the CPU utilization, although LoadSim 2003 latency time was reduced after moving to RAID-10 and increasing the number of driver systems. The additional CPU headroom that is available in such an

	PowerEdge 6850 with 3.66 GHz, 1 MB cache processors		PowerEdge 6850 with 3.33 GHz, 8 MB cache processors	
Workload	**Number of VMs**	**ESX Server CPU utilization**	**Number of VMs**	**ESX Server CPU utilization**
Microsoft SQL Server 2000	12	85%	16	87%
SUSE LAMP	14	85%	16	85%
NetBench 7.03	14	89%	17	86%
Microsoft Exchange Server 2003	4	81%	4	66%

Figure 5. Results of workload tests on the PowerEdge 6850 servers

environment can be used to support additional VMs and is not wasted as it would be on a traditional nonvirtualized server.

The PowerEdge 6850 provides an excellent platform for virtualization, with support for up to 32 GB of RAM, seven I/O slots, and several high-availability features. The large amount of RAM lends itself to supporting a large number of VMs that have significant memory requirements. The seven I/O slots allow for a large number of NICs to be installed to support VMs that have significant network-bandwidth requirements. Moreover, the QLogic QLA2362 daughtercard can provide dual-port Fibre Channel connectivity without occupying a PCI slot. The high-availability features of hot-plug PCI and memory RAID are designed to increase hardware availability, as explained earlier in this article.

To determine which processor to use in the PowerEdge 6850, IT departments must weigh their performance requirements against cost savings. In the Dell tests, the Intel Xeon processors with the larger 8 MB cache demonstrated higher performance than the Intel Xeon processors with the 1 MB cache; thus, the 8 MB cache processors would be better suited for performance-sensitive environments than the 1 MB cache processors. In contrast, a PowerEdge 6850 with 32 GB of RAM and the 1 MB cache processors would be well suited for hosting a large number of VMs with performance requirements that do not dictate 8 MB cache processors.

Testing efficiency in a mixed-workload scenario

To test the efficiency of VMware ESX Server running multiple workloads at the same time—a scenario very similar to how VMs are run on ESX Server in a server consolidation environment—Dell engineers ran all four workloads at the same time. Approximately one-fourth of each workload was run simultaneously on both the PowerEdge 2850 server and the PowerEdge 6850 server (using the 3.33 GHz, 8 MB cache processors). To measure efficiency, Dell engineers calculated the total CPU utilization by adding the individual CPU utilization for each workload test and then compared that calculated total with the actual total CPU utilization measured during the mixed-workload test.

Results of the mixed-workload test (see Figure 6) show that ESX Server—on both the PowerEdge 2850 and PowerEdge 6850—was able to efficiently handle multiple workloads running concurrently. The actual utilization of the PowerEdge 2850 was only 5 percent higher than the calculated value, while the actual utilization of the PowerEdge 6850 was slightly lower than the calculated total. The four-processor PowerEdge 6850 was better at handling the mixed workload than the two-processor PowerEdge 2850, but both were within 5 percent of the calculated value—showing that running multiple workloads simultaneously on the same server does not incur a significant penalty. Because of the high efficiency demonstrated by ESX Server, all the VMs on a physical server do not need to be of the same workload type in an ESX Server farm.

Workload	PowerEdge 2850		PowerEdge 6850 with 3.33 GHz, 8 MB cache processors	
	Number of VMs	ESX Server CPU utilization	Number of VMs	ESX Server CPU utilization
Microsoft SQL Server 2000	3	23.40%	4	21.75%
SUSE LAMP	2	17.40%	4	21.50%
NetBench 7.03	3	23.18%	4	21.25%
Microsoft Exchange Server 2003	1*	15.75%	1**	16.50%
Calculated mixed CPU utilization		80%		81%
Actual mixed CPU utilization		85%		80%

*In this test environment, 750 users were simulated.
**In this test environment, 1,000 users were simulated.

Figure 6. Results of the mixed-workload test on the PowerEdge servers

Providing excellent platforms for virtualization

The Dell PowerEdge 2850 and PowerEdge 6850 servers can provide excellent platforms for running VMware ESX Server virtualization software. By testing multiple workloads and a mixed-workload environment, Dell engineers demonstrated that ESX Server can consistently manage a large number of VMs with a range of I/O and memory-usage patterns. All four test workloads achieved similar results on each server, showing that any of these workloads can be suitable for overall VM-capacity sizing. Testing with a mixed workload running concurrently demonstrated that ESX Server can efficiently manage different I/O and memory-usage patterns.

When selecting an appropriate platform for an ESX Server environment, IT departments must consider performance and capacity requirements. The dual-processor PowerEdge 2850 offers a cost-effective, high-performance option that is well suited to building out a farm of ESX Server systems. The PowerEdge 6850 provides high performance and high capacity with support for up to 32 GB of RAM, seven I/O slots, and four processors. By determining the most critical factor for their virtualization environments, IT departments can select the most suitable Dell PowerEdge server.

Todd Muirhead is a senior engineering consultant on the Dell Scalable Enterprise Technology Center team. Todd has a B.A. in Computer Science from the University of North Texas and is Microsoft Certified Systems Engineer + Internet (MCSE+I) certified.

Dave Jaffe, Ph.D., is a senior consultant on the Dell Scalable Enterprise Technology Center team who specializes in cross-platform solutions. He has a B.S. in Chemistry from Yale University and a Ph.D. in Chemistry from the University of California, San Diego.

Scott Stanford is a systems engineer on the Scalable Enterprise Computing team within the Dell Enterprise Solutions Engineering Group. He has a B.S. from Texas A&M University and an M.S. in Community and Regional Planning from The University of Texas at Austin, and he is pursuing an M.S. in Computer Information Systems at St. Edward's University.

VIRTUAL addition

The New Math: Less = More

Addition
Rapidly Add New Server Resources

Subtraction
Subtract Server Maintenance Downtime

Multiplication
Multiply Your Business Continuity

Division
Divide and Conquer

Virtual Addition = The New Math

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

= unparalleled flexibility, fast time to productivity and economic benefits

Get the Power of VMware Virtualization

GET MORE OUT OF NOW

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

Enabling VMware ESX Server VLAN Network Configurations

for the Dell PowerEdge 1855 Blade Server

When used in conjunction with virtual LAN (VLAN) technology, server virtualization software can help build virtual infrastructures to support the scalable enterprise. In particular, VMware® ESX Server™ software, modular Dell™ PowerEdge™ 1855 blade servers, and Dell PowerConnect™ 5316M switches can be used along with VLAN configurations to create complex network infrastructures in virtualized data centers.

BY BALASUBRAMANIAN CHANDRASEKARAN, KYON HOLMAN, CUONG T. NGUYEN, AND SCOTT STANFORD

Related Categories:

Blade servers

Data networking

Dell PowerEdge blade servers

Internetworking

Network fabric

Virtualization

VMware

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

Virtualization allows for the creation of multiple virtual machines (VMs) that can run simultaneously on a single physical server. These VMs can communicate among each other and with other physical systems through virtual switches. Virtual switches are software entities that provide the functionality of a physical Ethernet switch. Coupled with virtual LAN (VLAN) technology, virtualization can be effectively used to set up complex network infrastructures for test, development, and production environments. For example, VLANs can provide traffic isolation channels and enhanced network security models for physical server hosts and the VMs running on them.

To show how VLANs can improve security and traffic isolation, engineers from the Dell PowerConnect Networking and Scalable Enterprise Development teams have developed best-practices methodologies for deploying VLANs and designed four network deployment models for VLAN architecture based on Dell PowerEdge 1855 servers, Dell PowerConnect 5316M switches, and VMware ESX Server 2.5 software. IT administrators and systems engineers can use these best practices and models to implement and support a secure, scalable virtualized network infrastructure.

Configuring the Dell PowerEdge 1855 for VMware ESX Server

VMware ESX Server software includes the VMotion™ feature, which allows for live migration of a VM from one physical server to another. VMotion copies the memory state of a VM from the source physical server to the destination server through the network fabric. The virtual disk is stored in a Fibre Channel storage device, which is shared between the two physical servers.

The Dell PowerEdge 1855 blade server can hold up to 10 server blades within its chassis, which is called the Dell Modular Server Enclosure. To enable VMotion in a blade environment using Fibre Channel storage area network (SAN) connectivity, a Fibre Channel daughtercard must be used in the server blade. Because the PCI daughtercard slot is used for Fibre Channel connectivity, this leaves room for only two network interface cards (NICs) per server blade. The two NICs are exposed to a server blade through the midplane interconnect. In an optimal ESX Server environment, four NICs are required: one for management, one for VMotion, and two teamed together for redundant connectivity for VMs. Management is enabled through the VMware service console, which is

a Linux®-based environment running an Apache Web server. The configurations described in this article assume only two NICs are used for the VMware ESX Server environment.

The following sections describe the advantages of using VLAN configurations in combination with the Dell PowerConnect 5316M switch—which is the network I/O module in the Dell Modular Server Enclosure—to support an optimal VMware ESX Server environment that enables highly available and secure network traffic. Four configurations that leverage the Dell PowerEdge 1855 and PowerConnect 5316M switch infrastructure are provided as well instructions on how to set up the ESX Server environment and the advantages of each configuration.

Exploring key concepts and advantages of VLANs

Current-generation blade servers provide a unique challenge because of the limited network ports available when compared to non-blade servers. Each server blade in the PowerEdge 1855 comes with two embedded Gigabit Ethernet[1] NICs, which are referred to as NIC 0 and NIC 1. For VMware ESX Server 2.5.1 software, the NICs must be either dedicated to the VMware service console, dedicated to the VMs, or shared between the service console and the VMs. In a default installation, NIC 0 is dedicated to the service console and NIC 1 is dedicated to the VMs.

To better utilize the network bandwidth and to provide redundancy, the NICs can be configured in one of the four configurations described in Figure 1: default, segregated traffic, dedicated VMotion network, and redundant. When added to an ESX Server environment, a VLAN can help improve traffic isolation and thereby the security of each of these configurations.[2] However, adding a VLAN may introduce some complexity to the initial setup stage. In addition, PowerEdge 1855 and ESX Server infrastructure maintenance tasks must be modified to include the additional VLAN layers. Still, these potential disadvantages are significantly offset by the benefits of increased security and traffic isolation, especially when VMotion or the service console shares a physical NIC with production VM traffic.

Additional important VLAN concepts include the following:

- Ethernet traffic sent on one VLAN will not be forwarded to another VLAN in a Layer 2 Ethernet switch. In addition, broadcast traffic will be sent only within the same VLAN.
- VLAN membership can be defined in the switch by associating specific network and virtual switch ports to a given VLAN. This means that only the ports associated with the VLAN may communicate with each other. Also, broadcast messages sent to any of these ports for a given VLAN can only be sent to other ports belonging to the same VLAN.
- A port may belong to different VLANs at the same time. In this case, the packet sent to the port must identify the VLAN to which it belongs. If the packet sent does not identify its VLAN, the switch will automatically associate the port to a default VLAN as configured in the port.
- VLAN configuration allows traffic to be groomed in the inbound direction and for nonconforming traffic to be automatically dropped.
- A VLAN does not necessarily guarantee quality of service (QoS). For example, if two VLANs share the same port, they will not split the network bandwidth between them. Thus, in heavily oversubscribed switch configurations, network traffic on one VLAN can potentially cause congestion of network traffic in the other VLAN.
- QoS parameters may be configured on the switch to select specific traffic types and provide priority to the traffic type (such as VM traffic versus VMotion traffic). The QoS setup is beyond the scope of this article.

If VLAN configurations are used, then once a particular configuration is selected, this same configuration choice must be applied to all the server blades in a chassis. As shown in Figure 1, for example, if the default configuration is selected (in which the service console uses NIC 0 and NIC 1 is reserved for the VMs and VMotion), then this configuration must also be used for all the other server blades. This restriction will simplify VLAN configuration in Ethernet switches. It is possible to circumvent this restriction, but the complexity of the configuration in that case is beyond the scope of this article.

Key points to consider for all four configurations shown in Figure 1 include:

- VMware management traffic from the service console is encrypted by default.
- VMotion traffic is not encrypted. To help ensure effective and secure VMotion events, best practices recommend configuring the VMotion NIC on a separate VLAN or using a separate physical NIC for VMotion.
- The service console does not generate only VMware management traffic. Systems management software suites, such as Dell OpenManage™ software, are also installed on the service console. In addition, baseboard management console (BMC) traffic, Simple Network Management Protocol (SNMP) traffic, and backup traffic use the same NIC as the service console.
- The three non-default configurations require sharing NIC 0 between the service console and the VMs, which requires administrators to perform additional steps during installation.

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

[2] For more information about the use of VLANs, see the *Dell PowerConnect 5316M Ethernet Switch Module User's Guide* at support.dell.com/support/edocs/network/PC5316M/en/UG/index.htm.

Configuration type	Use of NIC 0	Use of NIC 1	Fault tolerance for VMs	Installation type	Traffic isolation	VM performance	Security (VLAN)
Default	Service console	VMs and VMotion	No	Standard	Moderate	Acceptable if VMotion events are infrequent	**Moderate:** VMotion runs on a production network; adding a VLAN helps improve security and isolation of VMotion traffic
Segregated traffic	Service console and VMotion (shared)	VMs	No	Command-line post-installation steps	Good	Acceptable if management traffic is infrequent	**Good:** The service console runs on a private network; adding a VLAN helps improve isolation of traffic between VMotion and the service console
Dedicated VMotion network	Service console and VMs (shared)	VMotion	No	Command-line post-installation steps	Moderate	Acceptable if management traffic is infrequent and VMotion events are frequent	**Moderate:** The service console runs on a production network; adding a VLAN helps improve security and isolation of the service console traffic
Redundant NIC	NIC 0 shared between the service console and the VMs; NIC 0 and NIC 1 teamed and used by the VMs and VMotion		Yes	Command-line post-installation steps	Moderate and flexible	Acceptable if VMotion events are infrequent	**Poor:** VMotion and the service console run on a production network; adding VLAN helps improve security and isolation of the service console and VMotion traffic

Figure 1. NIC and VLAN configurations for Dell PowerEdge 1855 blade servers

Using the default VLAN configuration

In the default configuration, NIC 0 is dedicated to the service console for management traffic and NIC 1 is used for VMotion and traffic. No special steps are required to enable this configuration—it is part of the default installation of VMware ESX Server. This configuration provides network isolation by separating the management network from the VM network.

A VLAN can help address the problem of securing VMotion traffic by creating a separate virtual network for VMotion. Figure 2 shows the default configuration with a VLAN enabled. ESX Server allows administrators to create virtual switches, which are connected to the VMs for networking and also can be configured for VMotion. These virtual switches support the VLAN.

Enabling the default configuration

The following steps can be used to secure VMotion traffic through VLANs:

1. **Set up virtual switches in the ESX Server software to provide VLAN tags for all VM traffic.** Administrators can define the same VLAN tag for all VMs to share or specify a VLAN tag for each VM or for any group of VMs. The decision for assigning specific VLAN tags to the VMs depends on whether the VMs need to communicate with each other. If only a few VMs need to communicate with each other, administrators should group those VMs together and configure the switch to tag all traffic from those VMs with the same VLAN ID. Allowing the switch to tag the traffic can help improve ESX Server performance by offloading VLAN tag and packet inspection and routing processing tasks to the Dell PowerConnect 5316M application-specific integrated circuit (ASIC).
2. **Set up a VLAN on the PowerConnect 5316M switch connected to NIC 1.** This switch is represented as "Switch 1" in Figure 2. All the internal ports connected to the server blades are configured with the VLAN for VMotion (shown as VLAN ID = A in the figure). The PowerConnect switch connected to the service console NIC ("NIC 0")—represented as "Switch 0" in Figure 2—is left unconfigured.
3. **Assign one external port on Switch 1 to be a member of VLAN A.** Administrators should make sure this external port's permanent VLAN ID (PVID) is set to 4095, which will cause all untagged traffic to be discarded. Alternatively, if the external system does not support VLANs, then untagged traffic could be allowed into this port and the switch could automatically tag the traffic for VMotion. This option is slightly less secure but easier to manage than the first option of discarding all untagged traffic. The external port should be used only for transmitting and receiving VMotion traffic. This configuration means that VMotion traffic coming from an external source to the server blade must go through this port.

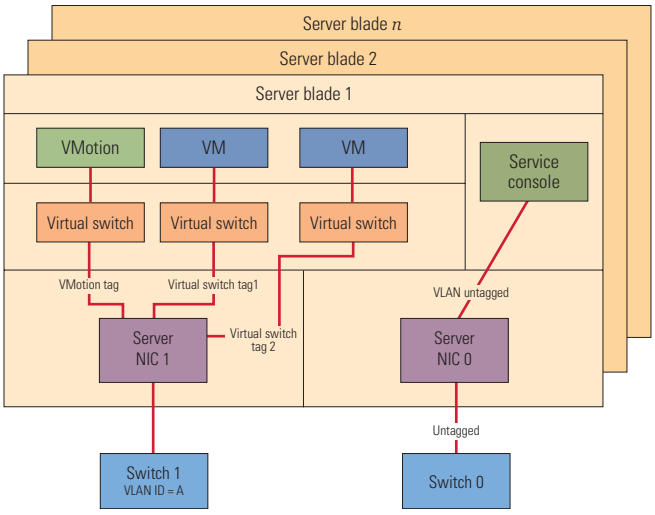


Figure 2. Default VLAN architecture using ESX Server virtual switches and physical Dell PowerConnect 5316M switches

4. **Configure the remaining external ports on Switch 1 to be specific members of the VM VLANs, if desired.** By assigning ports to specific VM VLANs, administrators can help isolate traffic from various VMs.

Note: For VMotion to function properly, the switch names should be identical across all blade servers.

For the three non-default VLAN configurations, the following sections provide a general overview of the advantages provided by each configuration. For detailed setup and configuration steps for these VLAN configurations, see the supplemental online section of this article at www.dell.com/powersolutions.

Implementing a segregated traffic configuration

In the segregated traffic configuration, the service console resides on a private network. Adding VLANs can help improve isolation of VMotion traffic and service console traffic. This configuration can help prevent a loss of management capabilities for the ESX Server host and provide a moderate level of virtual and physical switch port isolation for VMs that reside on VMotion-enabled ESX Server hosts running on a PowerEdge 1855 blade server.

Implementing a dedicated VMotion network configuration

In contrast to the segregated traffic configuration, the service console resides on a production network in the dedicated VMotion network configuration. Adding VLANs can help improve security by isolating service console traffic while also supporting an dedicated VMotion environment. For deployments of ESX Server 2.x software on PowerEdge 1855 blade servers, which require a high degree of management and VMotion security and isolation, this configuration provides systems engineers and VMware administrators with options to enable a granular level of logical and physical distinction between management and VM-related traffic.

Implementing a redundant NIC configuration

Building upon the segregated traffic and dedicated VMotion network configurations, the redundant NIC configuration provides an infrastructure for production environments that require even higher levels of network redundancy that achievable with the other VLAN configurations. Because the VMotion and service console networks reside on a production network in this configuration, adding VLANs can help improve security and isolation of service console and VMotion traffic. This configuration is well suited for environments that require high availability at the VM and physical network infrastructure layers.

Enabling fault tolerance using PowerConnect 5316M switch modules

To enable NIC failover when using PowerConnect 5316M switches in the redundant NIC configuration, administrators can select the advanced configuration in the Options tab of the ESX Server Management User Interface and change the `Net.ZeroSpeedLinkDown` parameter value to 1.

Examining the impact of VMotion events on VM production traffic

In some of the configurations described in this article, particularly the default configuration, VMotion traffic shares a NIC with VM production traffic. To learn more about the impact of VMotion events on VM production traffic when network resources are shared, see the Dell white paper "VMware VMotion Performance on the Dell PowerEdge 1855 Blade Server" in the Dell/VMware Resource Center at www.dell.com/vmware.

Following Dell best practices in virtual infrastructures

The Dell best practices and deployment models introduced in this article are designed to help enterprise IT organizations determine the appropriate network configurations and requirements for VMware ESX Server 2.x environments on Dell PowerEdge 1855 blade servers. The four VLAN configurations presented in this article can be implemented to meet various VM, VMotion, and service console management scenarios. Systems engineers, enterprise architects, and VMware administrators can employ these best practices and deployment models as resources and guides to help them build a scalable, flexible virtual network infrastructure.

Balasubramanian Chandrasekaran is a systems engineer in the Scalable Enterprise Computing Lab at Dell. His research interests include virtualization of data centers, high-speed interconnects, and high-performance computing. Balasubramanian has an M.S. in Computer Science from The Ohio State University.

Kyon Holman is a lead software engineer on the tape storage team in the Dell Enterprise Product Group. He has a B.S. in Computer Science from the University of Michigan at Ann Arbor and an M.S. in Software Engineering from The University of Texas at Austin, and he is pursuing an Executive M.B.A. from The University of Texas at Austin.

Cuong T. Nguyen is a systems engineer on the Dell PowerConnect Ethernet switch team. His expertise is in network management systems, and he has more than 15 years of networking and telecommunications industry experience. Cuong has a B.S. in Computer Information Science from the University of California at Irvine.

Scott Stanford is a systems engineer in the Scalable Enterprise Computing team within the Dell Solutions Engineering Group. His current focus is on performance characterization and sizing for virtualized solutions. He has a B.S. from Texas A&M University and an M.S. in Community and Regional Planning from The University of Texas at Austin, and he is pursuing an M.S. in Computer Information Systems at St. Edward's University.

Architectural Implementation of a Boot-from-SAN Manager

This article describes an architectural approach to booting servers from storage area networks (SANs) that employs an out-of-band boot-from-SAN manager. This approach incorporates both present and future industry standards.

BY MATTHEW BRISSE, AHMAD TAWIL, AND DRUE REEVES

Related Categories:

Boot-from-SAN

Dell/EMC storage

Fibre Channel

iSCSI storage

Standards

Storage

Storage management

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

Few architectural implementations that enable booting servers from storage area networks (SANs) exist today. These range from manual implementations using best-practices procedures to appliances designed to help simplify and automate the boot-from-SAN process. Many of these approaches fail to leverage industry standards, can be complex and costly, and can lead to dependence on a particular vendor.[1] This article describes a step-by-step approach to implementing a boot-from-SAN manager (BSM) that employs current and future standards and can help eliminate the need for proprietary approaches. Because no commercial BSM product currently exists, this architectural approach can be implemented to help scale and simplify the boot-from-SAN process.

Figure 1 depicts an example boot-from-SAN architecture using industry standards and out-of-band management of various SAN devices over an Ethernet management network. Out-of-band management provides a remote, common management approach for various interconnect technologies such as Fibre Channel, Internet SCSI (iSCSI), and Serial Attached SCSI (SAS).

Advantages of booting from a SAN

A common management architecture can establish a single point of management for all the components in a SAN—servers, host bus adapters (HBAs), switches, and storage subsystems. This architecture works on a standard PC-based server platform, including the BSM, which runs on any standard management station. By using industry-standard components, this architecture can be cost-effective and can easily scale from small to large SANs.

Low-level application programming interfaces (APIs) are required within each device to configure and manage the SAN devices to automate the boot-from-SAN operation. The BSM interfaces with the servers, fabric, and storage subsystems as an entire system (see Figure 2).

Common configuration of devices

The key to the architecture described in this article is the common configuration—based on current and future industry standards—of the underlying devices in the SAN. This configuration enables a BSM to manage each device without specific knowledge of the proprietary low-level APIs.

[1] For more information about the systems management benefits and complexities of boot-from-SAN implementations, see "Streamlining Server Management with Boot-from-SAN Implementations" by Matthew Brisse and Ahmad Tawil in *Dell Power Solutions,* August 2005, www.dell.com/downloads/global/power/ps3q05-20050101-Brisse.pdf.

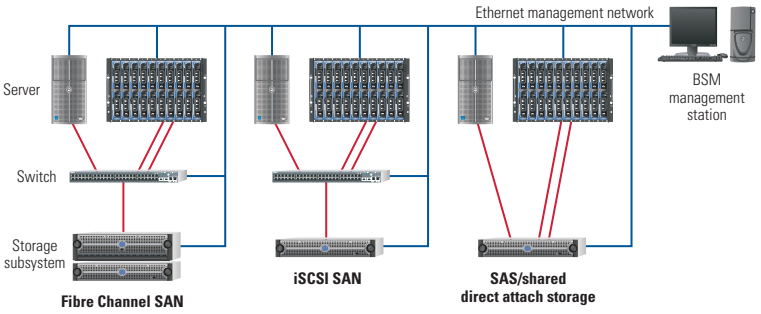


Figure 1. Boot-from-SAN architecture incorporating an out-of-band boot-from-SAN manager

In this architecture, the BSM could perform the following steps to set up the boot-from-SAN implementation:

1. Discover all the devices in the SAN.
2. Configure the storage subsystem to allocate storage for the boot logical units (LUNs).
3. Configure the fabric to assign the server and the storage subsystem to the same zone.
4. Assign the boot LUN to the HBA in the host, along with the associated data LUNs if desired.

The BSM can use several methods to discover the devices in the SAN. A standards-based approach would be to use the Service Location Protocol (SLP) in conjunction with the Storage Management Initiative Specification (SMI-S), which was developed by the Storage Networking Industry Association (SNIA). Other approaches involve proprietary agents or configurations that use Simple Network Management Protocol (SNMP), Domain Name System (DNS), or manual intervention.

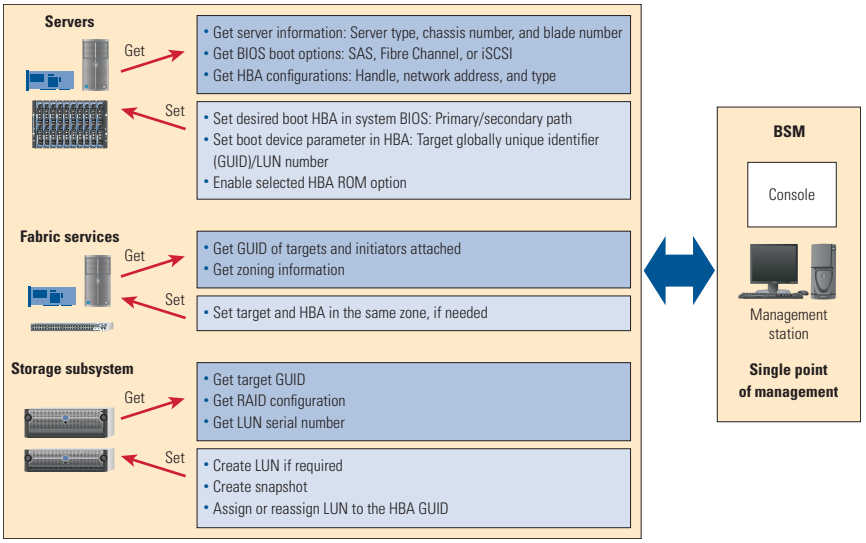


Figure 2. Examples of boot-from-SAN APIs

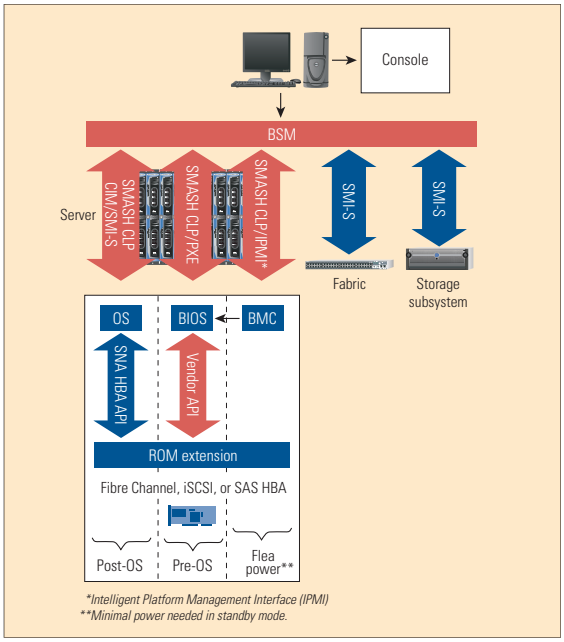


Figure 3. Interfaces between the boot-from-SAN manager and PXE

HBA configuration in the host, storage subsystems, and fabric should also use existing industry standards such as SMI-S. Using SMI-S, the BSM can perform boot LUN creation and assignment in the storage subsystem and fabric zoning in the SAN.

SMASH CLP

This architecture described in this article requires the BSM to communicate with the server to configure the server's BIOS and HBA (see Figure 3). In a pre-OS environment, the interface between the server and the BSM is based on the *Systems Management Architecture for Server Hardware Command-Line Protocol* (SMASH CLP), developed by the Server Management Working Group (SMWG) in the Distributed Management Task Force (DMTF). Until they become standardized, the low-level HBA APIs—required to program boot parameters in the HBA associated with the host—are based on a vendor-unique approach.

The BSM instructs the host to enter a Preboot Execution Environment (PXE) boot process to load and execute a boot agent on the host (see Figure 4). The purpose of the boot agent is to load a SMASH CLP environment on the host. The BSM executes remote SMASH CLP commands on the host to configure the boot parameters in the system BIOS and the HBA on the host side.

In a post-OS environment, the host requires an SMI-S or SMASH CLP environment, or both. This environment allows the BSM to configure the host's system BIOS and HBA boot parameters using SMI-S and SMASH CLP (see Figure 5).

UEFI

In an ideal solution, PXE and the vendor-unique HBA APIs could be eliminated with a Unified Extensible Firmware Interface (UEFI). UEFI is an emerging server system BIOS standard that provides pre-OS components—including a network stack—and the abilities to run pre-OS applications and load various drivers.

One of these pre-OS applications is the SMASH CLP environment that allows the BSM to evoke commands that configure the host's system BIOS and HBA boot parameters. In addition, UEFI allows the inclusion of a common set of UEFI APIs for the low-level configuration of a HBA boot device connected to Fibre Channel, iSCSI, or SAS storage.

Future development of industry standards

With further development of the UEFI, SMASH CLP, and SMI-S standards, the architecture described in this article could become the industry standard for configuring and managing the server, HBA, fabric, and storage components within a boot-from-SAN environment. This cost-effective, flexible, and scalable architecture provides a common approach for Fibre Channel, iSCSI, and SAS storage.

Today, the SMI-S and UEFI specifications do not contain all the necessary functionality to perform boot-from-SAN in the architecture described in this article. However, the Host Management Working Group within the SNIA Storage Management Initiative is working toward this goal by describing the boot capabilities of the host through SMI-S profiles using Common Information Model (CIM) classes. Also, the UEFI working group is building boot capabilities into the server system BIOS.

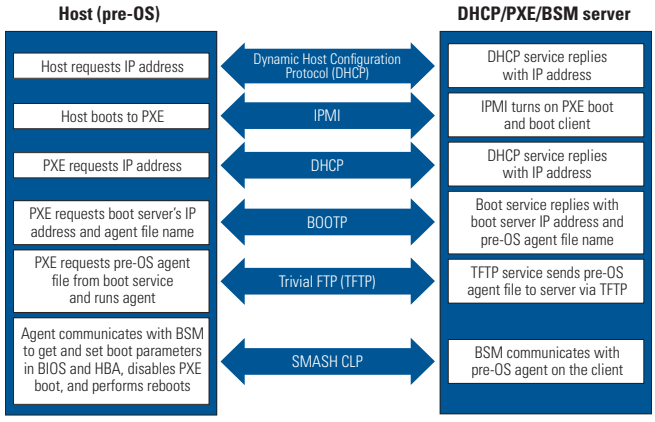


Figure 4. PXE-based boot-from-SAN server configuration sequence

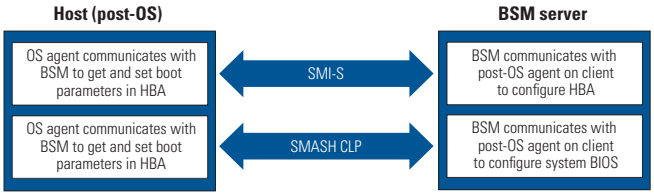


Figure 5. Post-OS server configuration interfaces

Enhanced boot-from-SAN capabilities

Independent software vendors can use the architecture proposed in this article to create a boot-from-SAN manager that is based on open standards and supports a heterogeneous SAN environment across different I/O interfaces. For enterprise IT departments, standardization can enhance product choice and flexibility. As industry standards evolve, automated boot-from-SAN will likely become possible—enabling further systems management capabilities such as policy-based management, server repurposing, image management, component self-discovery, pathway invisibility, and common authentication and authorization.

Matthew Brisse is a technology strategist in the office of the Dell CTO and is vice chair of the SNIA board of directors. At Dell, Matthew chairs the Standards Review Board and is a member of the Systems Management Architecture team.

Ahmad Tawil is a technology strategist in the Storage Architecture and Technology Group in the office of the Dell CTO, where he focuses on future networked I/O technologies.

Drue Reeves is a storage technology strategist in the office of the Dell CTO. He is co-chair of the SNIA Host Technical Working Group, a member of the Technical Steering Group, and the Dell representative for SMI-S–related activities. At Dell, Drue focuses on storage management architectures. He also holds a patent in management and security technologies.

FOR MORE INFORMATION

SNIA SMI:
www.snia.org/smi

SMASH:
www.dmtf.org/standards/smash

UEFI:
www.uefi.org

Background Patrol Read

for Dell PowerEdge RAID Controllers

Background Patrol Read, a new feature in Dell™ PowerEdge™ RAID Controllers (PERCs), is designed to help prevent data loss in a redundant array. This article describes how Background Patrol Read works and how it interoperates with Consistency Check and SMART alerts within the PERC Fault Management Suite.

BY DREW HABAS AND JOHN SIEBER

Related Categories:

Dell PowerEdge RAID Controller (PERC)

Dell PowerEdge servers

Fault tolerance

RAID

Storage

Systems management

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

Over time, sectors on a hard drive can become damaged and unusable. To help address this problem, hard drives keep track of unusable sectors in a grown-defect list. Once a sector is added to this list, future attempted accesses to it are remapped to a good sector—a back-up sector in the hard drive designated for this purpose.

Typically, bad sectors are discovered through normal read or write accesses to the drive. Using a redundant RAID level such as RAID-1 or RAID-5, a drive array can survive this data integrity threat by reconstructing the data from the good sectors of the other drives and remapping the data to an unused backup sector on the affected drive. Dell PowerEdge RAID Controller (PERC) products are designed to manage this process seamlessly and independently from the rest of the server without administrative intervention. As long as the array remains redundant, the server can remain online and running with data intact.

However, when a drive fails and there is an unknown bad sector on one of the nonfailed drives, problems may occur. As an example, consider a RAID-5 array that has experienced a drive failure and—unknown to the RAID controller—one of the nonfailed drives contains a hard-drive media defect. The administrator replaces the failed drive with a new one, and the array starts to rebuild the data onto the new drive based on the parity and peer data sectors from the remaining good drives. While data is being rebuilt on the new drive, the drive with the bad sector is read, but because of the media defect the data cannot be read. Without all of the available peer data and the parity, the new drive's sector cannot be regenerated. At this point, the array has lost data.

In the past when drive capacities were smaller, this type of problem was less likely to occur because small arrays typically contain fewer media defects than large arrays. Large hard drives are more prone to media defects. Media defects are specified as *x* number of defects per *y* number of bits. Therefore, larger drives are prone to more defects because they contain a greater number of bits. Today, hard-drive capacities have increased remarkably, and the likelihood has grown that one or more media defects will occur over the lifespan of the drive. In addition, large arrays take longer to rebuild than small arrays, thus increasing the amount of time the array is not redundant. Today's RAID systems need a proactive tool—such as the Background Patrol Read feature of Dell PERCs—to help avert such data problems by fixing the bad sectors when all of the drive array members are online and redundant.

Understanding how Background Patrol Read works

Background Patrol Read is designed to proactively detect hard-drive media defects while the array is online and redundant, and then proceeds to recover data. This tool provides three functions: data protection, variable run modes, and dynamic performance management.

Data protection

This function concerns data reconstruction and remapping. Background Patrol Read issues commands to each drive in the array to test all sectors. When a bad sector is found, the PERC instructs the hard drive to reassign the bad sector, and then reconstructs the data using the other drives. The affected hard drive then writes data to the newly assigned good sector. These operations continue so that all sectors of each configured drive are checked, including hot spares. As a result, bad sectors can be remapped before data loss occurs.[1]

Variable run modes

Background Patrol Read includes two run modes to help enhance flexibility and data protection:

- **Auto mode:** In this mode, the tool is always on. Once Background Patrol Read has checked all sectors of the array, it repeats this check indefinitely. This mode is recommended so that the PERC can maintain optimal array health.
- **Manual mode:** This mode is used to perform a single, quick check of the array. After Background Patrol Read has checked all sectors of the array, it stops and will not start again until started manually by the administrator. In Manual mode, Background Patrol Read commands are given a higher priority than in Auto mode so that the check completes significantly faster.

Manual mode is recommended during periods of low drive activity or during system maintenance. Background Patrol Read can be controlled by running the MegaPR utility within a Microsoft Windows® or Linux® OS.

Note: As its name indicates, Background Patrol Read is always a background or secondary process. Data I/O remains the highest priority for the RAID subsystem, whereas Background Patrol Read uses spare bandwidth.

Dynamic performance management

Background Patrol Read is designed to provide a balance between data protection and high performance. As a result, it uses an intelligent algorithm to adjust how much bandwidth it consumes based on the current data workload. During periods of light workload, the volume of Background Patrol Read commands increase. Conversely, during periods of heavy workload, the volume of Background Patrol Read commands decrease.

To gauge the intensity of the present workload, Background Patrol Read senses the volume of outstanding I/O operations pending for each hard drive. With this information, Background Patrol Read then adjusts both the frequency and size of the commands it sends to each drive.

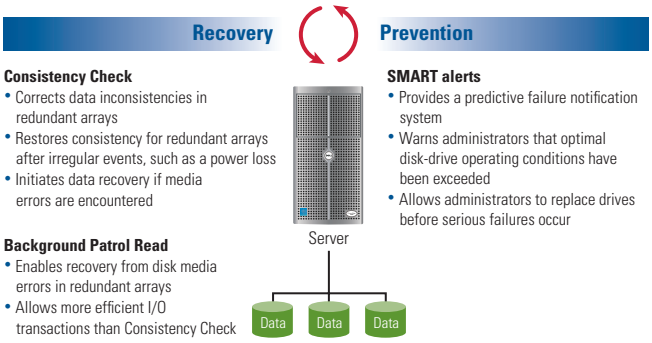


Figure 1. PERC Fault Management Suite

Exploring other components of the PERC Fault Management Suite

Background Patrol Read is just one of several PERC features that constitute the PERC Fault Management Suite. These features work in concert to help provide robust data protection. Other components of this suite are Consistency Check and SMART alerts. Figure 1 summarizes the components of the PERC Fault Management Suite and their functions.

Consistency Check

Consistency Check is designed to correct data inconsistencies in redundant arrays. A RAID-5 array is inconsistent when the data and parity do not match. Likewise, a RAID-1 array is inconsistent when the data and mirror do not match. Data inconsistencies can arise when all writes to an array are not completed because of catastrophic events such as a power loss.

Performance problems can arise when an array is inconsistent. RAID-5 arrays employ consistency checks to help improve write performance. For example, if the RAID controller must write new data to only a portion of a data stripe, it does not need to access every drive in the array if the array is consistent. When the array is consistent, the controller can read the new data from the host, read the old data from the affected drive or drives, and read the parity information. The RAID controller can then calculate the new parity and write the information to only the affected drives.[2]

An added bonus of running Consistency Check is that drive media are checked for errors and data recovery is initiated, similar to Background Patrol Read. However, Background Patrol Read is more efficient than Consistency Check in addressing this concern because Consistency Check is a data-level check and requires more controller resources to read and compare data. Also, because of the additional resources required, Consistency Check is not designed to

[1] Hard drives that are not part of a RAID array or are not assigned as hot spares are not scanned for media defects because these drives do not yet contain data.

[2] For more information about this read-modify-write operation, see "Understanding RAID-5 and I/O Processors" by Paul Luse in *Dell Power Solutions,* May 2003, www1.us.dell.com/content/topics/global.aspx/power/en/ps2q03_luse?c=us&l=en&s=corp.

run continuously. Rather, it should be scheduled to run at a regular interval, preferably during periods of low drive activity.

SMART alerts

Self-Monitoring Analysis and Reporting Technology (SMART) gauges hard-drive health. The PERC and hard drives work together to monitor various aspects of drive performance. They determine whether the drives are behaving normally and provide status information. Administrators can then choose to replace a drive before a failure occurs. SMART focuses on predictive errors, which are errors that occur over a long period and provide early warning signs that the situation is deteriorating.[3]

SMART alerts work in concert with Background Patrol Read and Consistency Check. During the life of a hard drive, Background Patrol Read and Consistency Check help maintain drive health and protect data. As the drive ages, more sectors are added to the grown-defect list and early warning conditions arise—thus triggering SMART alerts. SMART alerts allow administrators to assess a drive's health and consider replacement before failures occur.

Providing robust data protection

The PERC Fault Management Suite emphasizes proactive error detection, data recovery, and error prevention to help keep data safe and minimize downtime. In particular, Background Patrol Read blends proactive data protection with dynamic performance management and run-mode flexibility. This tool is designed to both protect data and provide optimal end-user experience by automatically maintaining performance and allowing administrators to make runtime adjustments to fit their data environments.

Drew Habas is a RAID product marketing manager at Dell. He is responsible for planning and managing Dell's RAID products. He has a B.S. in electrical engineering from the University of Illinois at Urbana–Champaign and an M.S. in engineering management from The University of Texas at Austin.

John Sieber is the lead engineer for Dell's SCSI RAID development group. He has extensive experience with RAID, storage area network, and network attached storage technologies. John has a B.S. in computer engineering from Texas A&M University.

[3] For more information about this technology, see www.pcguide.com/ref/hdd/perf/qual/featuresSMART-c.html.

Dell PowerVault ML6000 Tape Library Raises Storage IQ

Intelligent management and Ultrium 3 Linear Tape-Open (LTO-3) technology are built into the first truly modular tape library from Dell for intelligent backup and affordable scalability.

Dell PowerVault ML6020 CM

The balancing act between unrelenting data growth and sufficient storage management is a repeat performance for many organizations, with constricted backup windows, inadequate disaster recovery, and complex–and costly–storage area network (SAN) environments. The Dell™ PowerVault™ ML6000 family of modular tape libraries was designed to stop the teetering and raise the intelligence quotient for entry-level and enterprise SAN backup solutions and disaster recovery preparedness.

Modular forethought and exceptional scalability

Well versed in expandability and versatility, the modular architecture of the PowerVault ML6000 helps manage growth and protect long-term investments with a flexible 5U control module; 9U expansion modules; and nondisruptive, on-demand scalability. The control module can be placed anywhere in the expansion stack for easy customization of the library, and expansion modules can be located above or below the control module to easily expand the library as data grows.

Thinking ahead is made simple with nondisruptive, capacity-on-demand scalability. The PowerVault ML6000 family provides flexible configurations for preinstalled storage capacity–to activate additional library capacity in 46 slot increments through license keys without disrupting library operations–and scales from 14.4 TB to 51 TB with future growth to 161 TB of native storage capacity, using Ultrium 3 Linear Tape-Open (LTO-3) SCSI or Fibre Channel tape drives.

Intelligent backup

With intelligent diagnostics, intuitive wizards, and proactive triage, the PowerVault ML6000 can prepare the storage environment for quick disaster recovery and optimum availability. The self-awareness and built-in intelligence features of the PowerVault ML6000 help predict and isolate failures, speed resolution time, and limit unplanned downtime.

Easy-to-use setup wizards can guide administrators through installation step-by-step. Proactive monitoring supervises all systems, and components are designed to be operational and initiate preventative maintenance. Predictive self-diagnostics and service wizards generate automated service tickets, e-mail and pager notifications, and specific troubleshooting instructions. Library partitioning can enhance flexibility and backup consolidation while helping to simplify management and reduce rack-space requirements.

Proven reliability and unsurpassed performance

With high-capacity LTO-3 technology, WORM (write once, read many) functionality, continuous robotics, and up to 80 MB/sec transfer rates, the PowerVault ML6000 is designed to deliver reliability and performance that can scale with the library while extending and protecting stored data–and significantly helping to reduce backup windows and cycle times.

Clenching a fistful of benefits is not difficult with the PowerVault ML6000, engineered to protect investments by supplying compatibility with a myriad of storage software, Dell PowerEdge™ servers, and Dell storage systems. For the enterprise with shrinking backup windows, rampant data growth, and compliance requirements, the Dell PowerVault ML6000 offers a smart choice for scaling in tempo with growth.

For more information:

Dell PowerVault: **www.dell.com/storage**

DELL LTO-3

Moving backup *forward.*

Dell PowerVault ML6010 CM

Dell PowerVault 110T

Dell PowerVault 114T

Dell PowerVault 124T

Dell PowerVault 132T

Dell PowerVault ML6020 CM

Dell | Enterprise

Put your tape libraries in motion. Evolve with the Dell™ PowerVault™ series of Ultrium® 3 Linear Tape-Open™ (LTO-3) products,[1] designed to add flexibility, speed, and scalability to your storage requirements. Third-generation LTO® technology advances beyond the design of earlier LTO models, with greater capacity and performance, and high levels of reliability in desktop and automation-class products.

Bigger

More capacity can mean more savings. LTO-3 doubles the native cartridge capacity to 400 GB[2] to help reduce tape quantities required for backup, simplify cartridge management, and better utilize automation cartridge slots.

Faster

Unsurpassed performance meets the need for speed. LTO-3 races ahead with native transfer rates of up to 80 MB/sec[3]—and continues with Digital Speed Matching and super-fast load/unload cycle times for maximum performance.

Better

Reliability and functionality enhancements abound. LTO-3 tape drives feature field-proven components, firmware maturity, low power consumption, WORM (write once, read many) capabilities, and continued backward compatibility.[4]

LTO-3 drives are certified with Dell PowerEdge™ servers and available in the Dell PowerVault 110T, PowerVault 114T, PowerVault 124T, PowerVault 132T, and exclusively in the Dell PowerVault ML6000 family.

ULTRIUM LTO 3

GET MORE OUT OF NOW.

Visit www.dell.com/storage for more information.

[1]Dell PowerVault LTO-3 products utilize IBM LTO-3 tape drives. [2]LTO-2 drive models have a native capacity of 200 GB. [3]LTO-2 drive models vary, with the highest throughput rate being up to 35 MB/sec. [4]LTO-3 reads and writes LTO-2, and reads LTO-1.

Configuring a Highly Available Linux Cluster for SAP Services

Clusters of Dell™ PowerEdge™ servers using Oracle9*i*™ Real Application Clusters (RAC) can provide SAP® software environments with a flexible, scalable, and highly available database platform. The database will continue to run if one of the Oracle9*i* RAC database nodes fails; however, vital SAP functionality such as the message server and the enqueue server can still be single points of failure. To help protect these services from failure and thus unwanted downtime or even data loss, IT administrators can run them on a Red Hat® Enterprise Linux® OS–based cluster to complement the Oracle9*i* RAC database cluster and maintain service in a highly available manner.

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

Related Categories:

Clustering

Database

Dell/EMC storage

High availability (HA)

Linux

Oracle

Red Hat Enterprise Linux

SAP

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

Setting up a highly available SAP system on Linux requires eliminating any possible single point of failure for the database as well as for the various SAP components of the overall system. While the database is made highly available by means of Oracle9*i* Real Application Clusters (RAC) technology, SAP applications can be made highly available by protecting the SAP central instance—which includes the message server and the enqueue server—from failure. In addition, SAP management tools require a common shared $ORACLE_HOME directory, which requires the Highly Available Network File System (HA NFS) service exporting the Oracle® executables, the SAP executables, and SAP shared files such as profiles and the sapglobal directory.

On Linux, the node membership for Oracle9*i* RAC database nodes is managed by the Oracle cluster manager (oracm), which is designed specifically to manage RAC nodes. Therefore, administrators should implement a second, independent cluster to make the $ORACLE_HOME directory and the SAP central instance services highly available. This must be performed on a second set of hosts, because each node can be a member of only one cluster. Membership in two independent clusters with potential conflicts on current node status would render the cluster nodes unusable for each of the respective clusters. This second cluster uses the Red Hat Cluster Suite.

Setting up the Red Hat cluster

To set up the Red Hat cluster for the SAP software, administrators should first determine whether the Red Hat Package Manager (RPM™) packages for the Red Hat Cluster Suite are installed (see Figure 1). Depending on


```
[root@ls3220 root]# rpm -q clumanager
clumanager-1.2.16-1

[root@ls3220 root]# rpm -q redhat-config-cluster
redhat-config-cluster-1.0.2-2.0
```

Figure 1. Checking for Red Hat Cluster Suite RPM packages

the availability of updates, the version numbers may differ. Administrators should install the most recent version of these packages. They should then prepare the shared storage and the network connections. Throughout the example scenario used in this article, the server names ls3219 and ls3220 are used for the first and second cluster nodes, respectively.

Configuring the network

Both nodes must have two available Ethernet interfaces. One interface is used for cluster communication between the two nodes and should be on a private network. The other is the publicly visible network interface. The private network interfaces in this example are named dell3219 and dell3220, respectively. Depending on the specific requirements of the environment, administrators may want to set up four interfaces—two for each node—using the Linux kernel bonding mechanism. This provides a highly available network connection on each channel and secures the cluster against failure of one single component (network interface card, network cabling, or switch) on the respective communication channel.

Administrators should reserve one public IP address for each node. In this example scenario, these addresses are 10.17.64.25 for node ls3219 and 10.17.64.26 for node ls3220. Administrators should also reserve one private IP address for each node. For dell3219, the private address is 172.16.42.34; for dell3220, the private address is 172.16.42.35.

Additionally, administrators should reserve three IP addresses for the cluster services to be used as virtual IP addresses. They should configure the interfaces (or virtual interfaces) with these addresses, either by using the redhat-config-network program or editing the respective interface setup files in /etc/sysconfig/network-scripts. Figure 2 shows what the public interface on ls3219 should look like.

Administrators should set up all the interfaces on the nodes according to the host names and IP addresses. This is the same setup principle that is used in the Oracle9*i* RAC cluster: one public and one private IP address per node.[1]

Configuring the shared storage

After testing the network connections, administrators can set up the shared storage. The cluster software needs two small partitions as quorum devices, which should be configured on separate logical units (LUNs) to maximize independence and minimize possible contention. The partitions must have a minimum size of 10 MB each. However, the usual minimum size for a LUN on a Dell/EMC storage array is 100 MB. The quorum LUNs will be bound later as raw devices.

Additionally, administrators should create one or more LUNs to hold the file systems for the data to be exported via the HA NFS server. They should follow Oracle recommendations regarding the size for $ORACLE_HOME and SAP recommendations for the executables (approximately 300 to 400 MB depending on the kernel version), and they should take into account the data that will be stored in the LUNs as well. Once the LUNs have been created on the storage system, administrators can make them available to the nodes.

The next step is to create partitions on the quorum and NFS storage LUNs. For the quorum LUN, one partition is enough. Because administrators will bind these partitions as raw devices, they can set the partition type to "da" (non-file-system data) with the `t` option of `fdisk`. Figure 3 shows what the quorum partitions would look like on the host in the example scenario. Administrators should create and format partitions on the LUNs for the NFS directories.

Next, administrators should create persistent symbolic names for the partitions with devlabel. This program makes the partition device names resilient against device name reordering (for example, when the SCSI scan order is different). In the example scenario, the persistent symbolic name /dev/homedir is created for the partition used for the NFS export.

Because the quorum disks are raw devices, they must be bound so as to be available to the kernel. When the special symbolic name /dev/raw/raw*n* is used with devlabel, the link is created and the partition is bound as a raw device. Note that the identifier changes

```
[root@ls3219 root]# cat /etc/sysconfig/
   network-scripts/ifcfg-eth0
DEVICE=eth0
BOOTPROTO=static
IPADDR=10.17.64.25
NETMASK=255.255.252.0
ONBOOT=yes
TYPE=Ethernet
GATEWAY=10.17.64.1
```

Figure 2. Example public interface setup file for node ls3219

[1] For more information, refer to the "Checking the network connections" section in "Creating Flexible, Highly Available SAP Solutions Leveraging Oracle9*i* and Linux on Dell Servers and Dell/EMC Storage" by David Detweiler, Achim Lernhard, Florenz Kley, Thorsten Staerk, and Wolfgang Trenkle in *Dell Power Solutions*, November 2005; www.dell.com/downloads/global/power/ps4q05-20050174-SAP.pdf.

```
Disk /dev/sdb: 314 MB, 314572800 bytes
64 heads, 32 sectors/track, 300 cylinders
Units = cylinders of 2048 * 512 = 1048576 bytes

Device Boot  Start  End  Blocks  Id  System
/dev/sdb1      1    300  307184  da  Non-FS data

Disk /dev/sdc: 314 MB, 314572800 bytes
64 heads, 32 sectors/track, 300 cylinders
Units = cylinders of 2048 * 512 = 1048576 bytes

Device Boot  Start  End  Blocks  Id  System
/dev/sdc1      1    300  307184  da  Non-FS data
```

Figure 3. Example quorum partitions on host

to "RAW." Administrators can check the result with the `devlabel status` command (see Figure 4).

Administrators can check whether the raw devices are bound. As shown in Figure 5, the `raw` command displays the major and minor numbers of the bound devices. Administrators can check these numbers against the currently assigned block devices (from `devlabel status`).

Once the devlabel settings have been finalized, administrators can copy the /etc/sysconfig/devlabel file to the same directory on the other host. Then, they can log in to that host and issue the `devlabel restart` command. Administrators should not try to add raw devices and symbolic names themselves; they should allow devlabel sort out the unique IDs collected on the other node to ensure that the same physical device is bound under the same symbolic name.

Configuring a clustered NFS service

Once the network and devices are configured, administrators can activate the Red Hat cluster. Logged in as the root user, administrators can check whether the cluster services are running (see Figure 6). If the output does not show that the cluster services have stopped, administrators should stop them by issuing the `stop` argument to the init-script. Then, they can start the redhat-config-cluster program, preferably in a Virtual Network Computing (VNC) session.

Administrators should begin by setting up the raw devices for the cluster quorum. In the Cluster Configuration Tool, administrators should go to Cluster > Shared State to display the Shared State dialog box. In this box, administrators should enter the names of the two raw devices: /dev/raw/raw1 and /dev/raw/raw2.

They should then add the two nodes as members of the cluster by clicking the Members tab and going to File > New. In the dialog box, administrators should enter the name of the host (ls3219 in this example scenario). They should then repeat this step for the other cluster member (ls3220). They can leave "Enable SW Watchdog" checked—this enables the software watchdog timer, which allows a cluster member to reboot itself if the system hangs.

```
[root@ls3220 root]# devlabel status
brw-rw----   root   disk   /dev/raw/raw1
   --[RAW]--> /dev/sdb1
brw-rw----   root   disk   /dev/raw/raw2
   --[RAW]--> /dev/sdc1
brw-rw----   root   disk   /dev/homedir ->
   /dev/sdd1
```

Figure 4. Checking the status of partitions

Next, administrators can name the NFS service by clicking the Services tab and then the New button to display the Service dialog box. They should provide a name without spaces or special characters (this simplifies querying the status from the command line). In the example scenario, the HA NFS service name is RAC_NFS_directories. Neither a failover domain nor a user script should be assigned. Administrators can click the OK button and go to File > Save to save the changes to the cluster configuration.

In the next step, administrators can add a device to mount, a service IP address, and NFS clients to the service. In the example scenario, 10.17.64.27 is used as the IP address for the HA NFS service. To specify a service IP address, administrators should click the Services tab of the Cluster Configuration Tool. Then they can select the service and click the Add Child button. Next, administrators select "Add Service IP Address" in the popup window and specify an IP address with netmask and broadcast in the next window. Although the netmask and broadcast addresses are optional, best practices recommend setting them.

```
[root@ls3220 root]# raw -qa
/dev/raw/raw1:  bound to major 8, minor 17
/dev/raw/raw2:  bound to major 8, minor 33
```

Figure 5. Checking whether raw devices are bound

```
[root@ls3220 root]# /etc/init.d/clumanager status
clumembd is stopped
cluquorumd is stopped
clulockd is stopped
clusvcmgrd is stopped
```

Figure 6. Checking whether cluster services are running

EXPERT ADVICE. PEER SUPPORT. LAME JOKES.

How many experts does it take to solve a custom development problem? At sdn.sap.com, you'll find 250,000 developers, system managers and other insiders to help with your toughest applications challenges and coding snafus. Not to mention free sample downloads, advice from SAP staff and maybe even a few new punch lines.

// JOIN IN AT SDN.SAP.COM

THE BEST-RUN BUSINESSES RUN SAP™

ORACLE CLUSTER FILE SYSTEM

A cluster file system allows all nodes in a cluster to concurrently access a device via the standard file system interface—enabling easy management of applications that need to run across a cluster. Oracle Cluster File System (OCFS) Release 1 was introduced in December 2002 to enable Oracle RAC users to run the clustered database without having to use raw devices. This file system was designed to store database-related files such as data files, control files, redo logs, and archive logs.

OCFS2—the next generation of the Oracle Cluster File System—was introduced in August 2005 for Red Hat Enterprise Linux 4 and Novell® SUSE™ Linux Enterprise Server 9 platforms. This high-performance, general-purpose cluster file system can store not only database-related files on a shared disk, but also Oracle binaries and configuration files (shared Oracle home)—helping to make the management of Oracle RAC easier using OCFS2 than OCFS1. Also, any non-Oracle binaries or non-Oracle configuration files, such as shared SAP directories, can be stored on shared disks.

In addition, OCFS2 provides metadata caching; metadata journaling; cross-node file data consistency; easy administration, including operation as a shared root file system; support for multiple block sizes; support for up to 254 cluster nodes; context-dependent symbolic link (CDSL) support for node-specific local files; asynchronous and direct I/O support for database files for improved database performance; and full integration with Linux kernel 2.6 and later. With the release of OCFS2 on Linux, enterprises can implement Oracle RAC with these capabilities while enhancing the overall environment by not having to use HA NFS Linux volumes for the required shared SAP directories.

To add the device to be mounted by the service, administrators can click the Services tab of the Cluster Configuration Tool, select the service, and click the Add Child button. They can then select "Add Device" and click the OK button. Then they can specify a device special file (in this example, /dev/homedir) and a mount point (/sapmnt/clu_export). Each device must have a unique device special file and a unique mount point within the cluster and across service boundaries. Administrators can specify a directory from which to mount the device in the Mount Point field. This directory cannot be listed in /etc/fstab because it is automatically mounted by the Red Hat Cluster Manager when the service is started. Administrators should choose a file system type from the FS Type list (ext3 is used in the example scenario).

Administrators can specify options for the device. If the Options field is left blank, the default mount options (rw, suid, dev, exec, auto, nouser, and async) are used.[2] Administrators can check "Force Unmount" to force any application that has the specified file system mounted to be shut down prior to disabling or relocating the service (when the application is running on the same cluster member that is running the disabled or relocated service). When finished, administrators can click the OK button and go to File > Save to save the changes to the /etc/cluster.xml configuration file, and go to File > Quit to exit the Cluster Configuration Tool.

Testing the cluster

Once the cluster is configured, administrators can begin the first cluster test. First, administrators should restart the redhat-config-cluster program and go to Cluster > Start local cluster daemons. Once the status display shows that the host has changed from "Unknown" to "Active," administrators can enable the service by selecting it in the Services window and clicking "Enable." The service status should change from "Disabled" (red) to "Running" (black). On this node, administrators should now see the mounted device under the configured mount point. If administrators do not see the device, they should check the system log for cluster service error messages.

After a successful test on one node, administrators can copy the /etc/cluster.xml file into the same directory on the other node. Then, they can start the cluster services there, either with the init script or with the redhat-config-cluster graphical user interface (GUI). Administrators also should test switching the service between the two cluster hosts.

Adding clients to the clustered NFS service

After testing that the cluster runs properly, administrators should extend the configuration of the NFS service to export one or more directories to the clients. Administrators should check that the NFS daemon and the portmapper run on both hosts and are configured to start automatically. They should execute the following commands on both hosts:

```
/sbin/chkconfig --level 345 nfs on
/sbin/chkconfig --level 345 portmap on
```

This enables automatic starting in the runlevels 3, 4, and 5. Administrators can check the result by entering the following command:

```
/sbin/chkconfig -list service
```

[2] For a description of the available options, administrators should refer to the mount man page.

The output should look similar to the following:

```
[root@ls3220 root]# /sbin/chkconfig --list nfs
nfs    0:off    1:off    2:on    3:on    4:on
    5:on    6:off
```

In addition, the output should look similar for services that will be set up later. Administrators should perform these steps on both hosts.

Next, administrators should return to the cluster configuration GUI and click the Services tab. They should select the NFS service and click the collapse/expand indicator, or *twistie,* on the left to display the contents. Administrators should see the service IP address and the service device. They can then select the device and click the Add Child button. A popup window asks for the export directory name. In this example scenario, everything below /sapmnt/clu_export is exported, and the following directories are exported with different access permissions:

- /sapmnt/clu_export/readonly (ro, async)
- /sapmnt/clu_export/read_write (rw, sync)
- /sapmnt/clu_export/read_write_root (rw, sync, no_root_squash)

Even when it is not immediately necessary to create an export that is root-writeable and preserves the user ID (without reassigning nfsnobody to root), best practices recommend configuring the export with these settings—to enable backups and quick file distribution among the hosts.

The allowed client and permissions are attributes of the NFS Export Client object, which is a child of the NFS Export object. Administrators can add the clients again by selecting "NFS Export" and clicking the Add Child button. When finished, the NFS service structure should resemble the structure of the XML file /etc/cluster.xml (see Figure 7).

Note: The NFS export is under the control of the Red Hat cluster, and the directories exported there must not appear in the /etc/exports file used by the non-clustered NFS daemon.

Client-side mount options

On the client side, the directories are mounted under /sapmnt/homedir/readonly following a schema where /sapmnt/*hostname*/*directory* mounts directories exported by *hostname*. All clients mount the exported directories there. The /etc/fstab entries for the example scenario are shown in Figure 8.

```
Service RAC_NFS_directories

service_ipaddress
ipaddress="172.16.42.60"
netmask="255.255.255.0"
broadcast="172.16.42.255"

device name="/dev/homedir"
mountpoint="/sapmnt/clu_export"
fstype="ext3"
forceunmount="yes"

nfsexport name="/sapmnt/clu_export/readonly"
client name="172.16.42.0/24"
options="ro,async"

nfsexport name="/sapmnt/clu_export/readwrite"
client name="172.16.42.0/24"
options="rw,sync"

nfsexport name="/sapmnt/clu_export/readwrite_root"
client name="172.16.42.0/24"
options="rw,sync,no_root_squash"
```

Figure 7. NFS service directory structure

Adapting the SAP directory structure

Locally, symbolic links point to the NFS-mounted directories. For example, the SAP instance DVBGS00 would expect the directory structure shown in Figure 9 on its server. The directories are located on NFS and can be found in the same location (that is, with the identical pathname) on every host. The /usr/sap/RAC/SYS directory links to /sapmnt/RAC (see Figure 10), and the /sapmnt/RAC directory links to the NFS directories (see Figure 11).

In the example scenario, the NFS directories are organized by system ID (SID) to support more than one SAP system (see Figure 12). The readonly/ and readwrite/ incarnations of the RAC_sapsystem directory show that the directories used by an SAP system are divided by these attributes, as shown in Figure 13.

```
# HA NFS exports
homedir:/sapmnt/clu_export/readonly         /sapmnt/homedir/readonly        nfs\ hard,intr,nolock,ro,bg  0 0
homedir:/sapmnt/clu_export/readwrite        /sapmnt/homedir/readwrite       nfs\ hard,intr,sync,bg      0 0
homedir:/sapmnt/clu_export/readwrite_root   /sapmnt/homedir/readwrite_root  nfs\ hard,intr,sync,bg      0 0
```

Figure 8. Example /etc/fstab entries


```
ls3220:racadm-DVBGS00 > find /usr/sap/RAC/
   DVBGS00/* -prune
/usr/sap/RAC/DVBGS00/data
/usr/sap/RAC/DVBGS00/log
/usr/sap/RAC/DVBGS00/sec
/usr/sap/RAC/DVBGS00/work
```

Figure 9. Directory structure for DVBGS00 SAP instance

Configuring SAP central instance services for the cluster

Access to data in the underlying database of an SAP system is synchronized with a special lock system called the SAP enqueue mechanism. This mechanism serializes access and prevents access from being changed for more than one requesting party.

The enqueue server usually runs as a service of the SAP central instance. If clients run in the same SAP instance, they can contact the enqueue server via the UNIX® Interprocess Communication (IPC) mechanism; if they are not part of the central instance, clients contact the enqueue server via the SAP message server. As opposed to all other components of the SAP system on the application layer, the enqueue server holds a state—an in-memory table of granted locks—that cannot be recovered gracefully if the service fails. The message server, which consequently plays an important role in contacting the enqueue server, holds no state; it receives only incoming connection requests and transfers them to the addressee. The message server can be restarted after failure, with no impact other than delayed communications. The enqueue server is a potential single point of failure in an SAP system, isolated from the failover provided at the database layer.

The SAP solution to the enqueue challenge is the stand-alone enqueue server and the enqueue replication mechanism. With these components, the enqueue server runs as a stand-alone program and can be contacted directly by enqueue clients. Additionally, a second enqueue server—called the enqueue replication server—is started; its only task is to maintain a second copy of the enqueue state table (lock table). Communicating regularly with the enqueue server, the enqueue replication server keeps its copy of the enqueue table current. If the enqueue server fails, it can be restarted on the host where the enqueue replication server runs. When the enqueue replication server recognizes that the enqueue server has started up, it will transfer its current lock table before exiting. The newly started enqueue server can now continue without losing valuable enqueue state information. Additionally, OS-level high-availability software makes the enqueue server available via a virtual, clustered IP address, masking the restart from the clients so that they always connect to the same IP address.

Splitting the central instance

To secure the SAP system's services in a high-availability cluster, administrators must split the traditional central instance into dedicated instances because a large "service block" can be difficult to monitor. Furthermore, this large block makes restarting services difficult, because administrators must also restart parts of the central instance that have not failed.

To run the enqueue server as a master/slave service, the enqueue service and the enqueue replication service should always reside on different hosts. The message server is not bound to a particular host. Because these are the two services that constitute a central instance, the cluster can run only those services,

```
ls3220:racadm-DVBGS00 > ls -l /usr/sap/RAC/SYS/
total 4
drwxr-xr-x    2 racadm   sapsys  4096 Nov 11 12:44 exe
lrwxrwxrwx    1 racadm   sapsys    18 Jan 24 14:48 global -> /sapmnt/RAC/global
lrwxrwxrwx    1 racadm   sapsys    19 Jan 24 14:49 profile -> /sapmnt/RAC/profile
```

Figure 10. Links to /sapmnt/RAC

```
ls3220:racadm-DVBGS00 > ls -l /sapmnt/RAC/
total 0
lrwxrwxrwx    1 racadm   sapsys    49 Jan 24 14:49 exe -> /sapmnt/homedir/readonly/RAC_sapkernel/exe-640-21
lrwxrwxrwx    1 racadm   sapsys    46 Jan 24 14:49 global -> /sapmnt/homedir/readwrite/RAC_sapsystem/global
lrwxrwxrwx    1 racadm   sapsys    46 Jan 24 14:49 profile -> /sapmnt/homedir/readonly/RAC_sapsystem/profile
```

Figure 11. Links to the NFS directories


```
[root@ls3220 root]# ls -l /sapmnt/clu_export/*
/sapmnt/clu_export/readonly:
total 12
drwxr-xr-x    3 root     root         4096 Jan 21 19:05 RAC_oracle
drwxr-xr-x    3 racadm   sapsys       4096 Jan 21 12:19 RAC_sapkernel
drwxr-xr-x    4 racadm   sapsys       4096 Jan 21 17:47 RAC_sapsystem

/sapmnt/clu_export/readwrite:
total 8
drwxr-xr-x    2 orarac   dba          4096 Jan 21 03:03 RAC_oracle
drwxr-xr-x    8 racadm   sapsys       4096 Jan 27 13:50 RAC_sapsystem

/sapmnt/clu_export/readwrite_root:
total 0
```

Figure 12. Directories organized by system ID

and all application servers must be outside the cluster. However, for systems management purposes, the message server can run together with a dialog service, and an application server can reside in the cluster, or close to it. In the example scenario, the enqueue server; the enqueue replication server; the message server; and an application instance with dialog, update, batch, and spool work processes all run as services in the cluster.

In the example scenario, the traditional central instance DVEBMGS is split into multiple instances as follows (the two numerals at the end of each instance name represent the system number):

- **DVBGS00:** Dialog, update, batch, gateway, and spool work processes
- **DM01:** Dialog service (for local administration) and message server
- **E02:** Enqueue server
- **R02:** Enqueue replication server

Note that the enqueue server and the enqueue replication server must have the same system number (02) separate from the rest of the instances; otherwise, the takeover of the enqueue table will fail.

For each instance, administrators must define an instance profile and a start profile, according to SAP documentation. However, to put these instances under the control of the high-availability cluster, administrators must provide scripts for the cluster that conform to the UNIX System V init conventions—that is, Bourne shell (bash) scripts that offer a start, stop, and status function.

For example instance profiles and start profiles, see the supplemental online section of this article at www.dell.com/powersolutions.

Switching between different user environments

To start the SAP services, the programs use the environment of the SAP administrative user. Because those environment parameters differ from instance to instance, a simple way to switch between different environments for the same OS user is desirable. To achieve this switching, administrators can adapt the default environment contained in the sapenv.sh and dbenv.sh scripts for each instance, and rename the script sapenv_*INSTANCENAME*.sh. Then, they can create a scriptlet—a reusable script element—containing a source statement such as the following:

```
#!/bin/sh
source sapenv_INSTANCENAME.sh
```

```
[root@ls3220 root]# ls -l /sapmnt/clu_export/readonly/RAC_sapsystem/
total 8
drwxrwxr-x    3 racadm   sapsys       4096 Jan 28 03:16 profile

[root@ls3220 root]# ls -l /sapmnt/clu_export/readwrite/RAC_sapsystem/
total 24
drwxrwxr-x    6 racadm   sapsys       4096 Jan 26 18:30 DVBGS00
drwxrwxr-x    2 racadm   sapsys       4096 Jan 28 02:46 global
drwxr-xr-x   11 racadm   sapsys       4096 Nov 11 11:00 trans

[root@ls3220 root]# ls -l /sapmnt/clu_export/readwrite/RAC_sapsystem/DVBGS00/
total 16
drwxrwxr-x    2 racadm   sapsys       4096 Jan 28 06:38 data
drwxrwxr-x    2 racadm   sapsys       4096 Jan 27 11:01 log
drwxrwxr-x    2 racadm   sapsys       4096 Jan 26 18:53 sec
drwxrwxr-x    2 racadm   sapsys       4096 Jan 28 08:36 work
```

Figure 13. Directories used by an SAP system

```
RAC_app_server   DVBGS00  ls3216  10.17.64.22
RAC_message_server   DM01 ls3221 10.17.64.27
RAC_enqueue_server   E02  ls3222    10.17.64.28
R02  no address. Bound to the public IP address
   of the owning member.
```

Figure 14. Virtual IP addresses and cluster services for SAP instances

```
RAC_app_server     /etc/init.d/sapappserver-RAC
RAC_message_server    /etc/init.d/sapmsgsrv-RAC
RAC_enqueue_server   /etc/init.d/sapenserver-RAC
```

Figure 15. Corresponding init scripts for cluster services

By using the command source, administrators can make changes to the environment variables effective for the current shell session. If administrators source this scriptlet, the user has the matching parameters for *INSTANCE_NAME* in the environment. This can also be seen in the `start()` and `stop()` functions of the initscripts package, because the scriptlet is sourced before executing the command.

Integrating the SAP instances as a cluster service

Administrators must create virtual IP addresses for all of the SAP instances, except for the enqueue replication server. The enqueue replication server always runs on the host not owning the enqueue server and attaches itself to the enqueue server (as opposed to the enqueue server trying to contact it). It can be bound to the public IP address of the respective cluster member, even if this means that it changes IP address with every service relocation.

The DVBGS, DM, and E instance each require a virtual IP address, so that they are always present under the same address from outside the cluster. Because DVBGS and DM also appear in the instance list (SM51), administrators should adapt the instance name to show the host name belonging to the service IP address, not the currently active cluster member IP address. Administrators can do this by setting rdisp/myname to *virtualhostname_SID_SAPSYSTEM*. In this manner, the instance names remain stable after relocation of the service from one cluster member to another.

Administrators should create cluster services for the SAP instances and give each service a virtual IP address as a child. Figure 14 shows this configuration for the example scenario, and Figure 15 shows the corresponding init scripts.

As shown in Figure 14, R02 does not have a service. This configuration is used in the example cluster because a service order cannot be defined, nor can services be set in a relationship. Each service is independently monitored and treated without regard to the other configured services. Because the enqueue server and the enqueue replication server are dependent and must be started and stopped on opposite hosts and in a specific order, administrators must start and stop the enqueue replication server from inside the sapenserver-*SID* script.

Next, administrators can enter the scripts as "user scripts" in the service definitions and configure a check interval, which typically varies from 30 to 60 seconds. Before transferring control of the services to the cluster, administrators should run the scripts manually to test their functionality.

Building a reliable platform for SAP

Oracle9*i* RAC for SAP on Linux can provide a stable, flexible, and scalable environment, provided administrators follow proper planning and installation procedures. By using the SAP enqueue mechanism with Linux, administrators not only can help protect the database from unplanned downtime, but they also can set up the SAP environment to avoid disruptions to end users.

David Detweiler is the Dell SAP Alliance Manager in Europe, the Middle East, and Africa (EMEA) and a member of the Dell SAP Competence Center in Walldorf, Germany, which helps ensure that current and future Dell technologies work together with SAP solutions.

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

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

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

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

Introduction to Online Diagnostics for Dell PowerEdge Servers

Dell™ OpenManage™ Online Diagnostics is a comprehensive, cross-platform diagnostics program designed to enhance operation of Dell PowerEdge™ servers and help reduce service costs. This article introduces Dell OpenManage Online Diagnostics, describes its features, and discusses administration scenarios.

BY PRATHAP THATHIREDDY AND SRIKRISHNA SRIDHAR MURTHY

Related Categories:

Dell OpenManage

Dell PowerEdge servers

Diagnostics

Systems management

Troubleshooting

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

Diagnostics that help identify hardware failures and changes in a system's condition can be critical for administrators who must minimize server maintenance and maximize uptime and reliability. Deploying an efficient, effective diagnostics program can help reduce system downtime considerably.

The Dell OpenManage Online Diagnostics program is an integral part of Dell OpenManage Server Administrator software. This program consists of a suite of diagnostic test modules that run locally on a Dell PowerEdge server and can be accessed remotely over a network. Diagnostic tests can be selected from a hierarchical menu representing the hardware that the Online Diagnostics program discovers on a Dell PowerEdge server. Tests can be run simultaneously or sequentially in a single session. In addition, administrators can view progress and results for each selected test or hardware component. Figure 1 shows the Diagnostic Selection screen in Dell OpenManage Server Administrator.

Benefits of Dell OpenManage Online Diagnostics

Administrators can install and use Dell OpenManage Online Diagnostics to diagnose performance issues with their hardware. For instance, a Dell-supplied serial modem might not be performing at the specified speed. An administrator can use the device tree in Online Diagnostics and run diagnostics on the entire subset of devices that might be causing the modem issue.

In this example, the administrator would run diagnostics on the modem and its parent, the serial port. If the modem is at fault, the modem diagnostic program would fail one or more of the commands written to it and flag an error to the administrator. However, if the problem is caused by an improper serial port setting (such as a low baud rate), then the serial port diagnostics would fail, indicating a meaningful error to the administrator. In this manner, the Online Diagnostics program can help administrators and Dell tech support staff to isolate hardware issues and prevent false service dispatches, thus helping to reduce service costs.

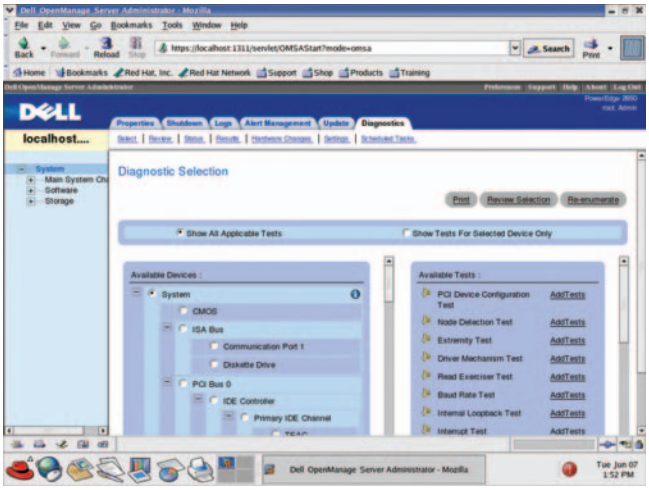


Figure 1. Dell OpenManage Server Administrator Diagnostic Selection screen

Devices supported by Dell OpenManage Online Diagnostics

The Dell OpenManage Online Diagnostics program provides diagnostics for several Dell-supplied and Dell-certified hardware devices.

CMOS RAM diagnostics. CMOS (complementary metal-oxide semiconductor) memory contains system configuration information. The CMOS diagnostic program performs a checksum test of the CMOS memory to determine whether any bytes are corrupt.

CD/DVD diagnostics. The CD/DVD diagnostic test identifies drive-related mechanical problems such as those affecting the drive door, spindle motor, and fault-sector and read functions.

Serial port diagnostics. Serial port diagnostics identify any issues, such as baud rate, related to serial port configurations. They also cover communication issues such as internal loop back, interrupt handling, and internal registers. These diagnostics can be used to diagnose performance-related issues that result from improper configuration of serial devices.

Parallel port diagnostics. This test diagnoses parallel port configurations and communication-related issues. It can be used to diagnose performance-related issues that result from improper configuration of parallel devices.

Modem diagnostics. Modem diagnostics analyze communication registers on the modem. This also helps to diagnose any modem hardware issues in sending and receiving commands.

Network interface controller diagnostics. This test is used to diagnose any network communication and configuration-related issues. These advanced diagnostics can be very helpful to administrators in resolving issues with network configuration on Dell servers.

Memory diagnostics. Memory diagnostics are designed to test the system memory's storage integrity and its ability to store data accurately. This test verifies that data paths, error-correction circuits, and memory devices are working correctly.

Dell Remote Access Controller (DRAC) diagnostics. The DRAC diagnostic test provides IT administrators with continuous access to remote Dell servers. These diagnostics analyze DRAC hardware and communication issues.

USB controller diagnostics. USB controller diagnostics are designed to identify hardware and communication-related issues of USB controllers and any attached devices.

Floppy disk diagnostics. Floppy disk diagnostics detect problems with floppy disk controllers and their related components, such as the motor, read/write mechanism, and floppy disk sectors.

PCI diagnostics. PCI diagnostics detect any driver-related errors or interrupt request (IRQ) sharing warnings for PCI devices in the system.

RAID controller diagnostics. Dell PowerEdge RAID controller diagnostics report problems with RAID controller hardware, batteries, and attached disks.

SCSI controller diagnostics. SCSI controller diagnostics detect problems with SCSI controller hardware and attached devices such as SCSI hard drives, tape drives, and autoloaders.

Features of Dell OpenManage Online Diagnostics

The Dell OpenManage Online Diagnostics program provides IT administrators with various features to perform diagnostic tests.

CLI- and GUI-based tests

Diagnostic tests can be performed using OS-supported command-line interfaces (CLIs) locally or remotely over Telnet or Secure Shell (SSH). This feature can be helpful to administrators when they are scripting diagnostics using well-known scripting tools. Graphical user interface (GUI)–based diagnostics can be performed via HTTP over Secure Sockets Layer (HTTPS) ports using well-known browsers such as Microsoft Internet Explorer or Mozilla Firefox.

Device enumeration and test/device inventory

This feature enables system administrators to re-inventory all supported devices after hardware reconfiguration. This feature can prove helpful after adding components such as plug-and-play devices, installing drivers for supported hardware components, or implementing hot-swappable devices. The test/device selection feature allows administrators to select desired tests and specify the devices on which to run them.

Diagnostic scheduler

The diagnostic scheduler feature can help increase system uptime by allowing administrators to select diagnostic tests to run at specific times and frequencies. This can help administrators schedule server maintenance and run appropriate diagnostics without affecting

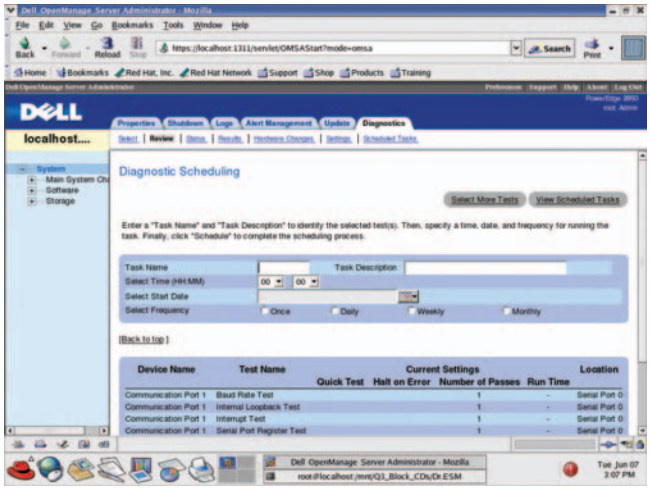


Figure 2. Dell OpenManage Server Administrator Diagnostic Scheduling screen

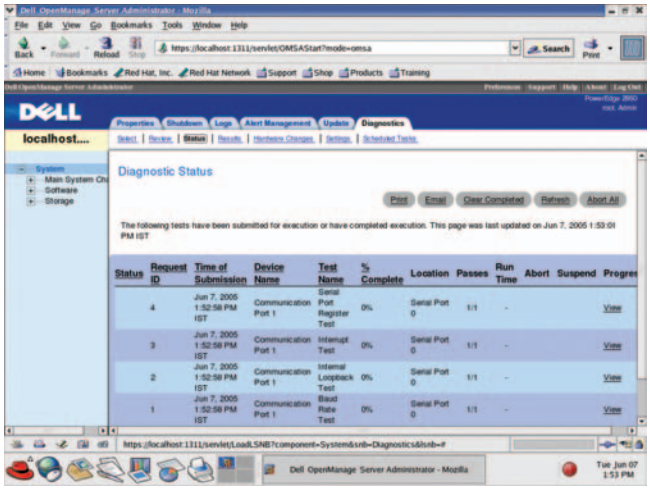


Figure 4. Dell OpenManage Server Administrator Diagnostic Status screen

business deliverables. Figure 2 displays the Dell OpenManage Server Administrator Diagnostic Scheduling screen.

Diagnostic test review, test status, and result history

The diagnostic test review option lets administrators review the selected diagnostic tests before submission for execution. This means that administrators can change settings that are predefined. Other options include choosing test-specific settings such as halt-on-error, specifying the number of iterations a test should run, or even scheduling a test to run at a later time. Figure 3 displays the Diagnostic Selection Review screen.

The diagnostic test status feature enables administrators to monitor the status of the diagnostic tests that are running, allowing them to view the progress of each test under execution. Figure 4 displays the Diagnostic Status screen.

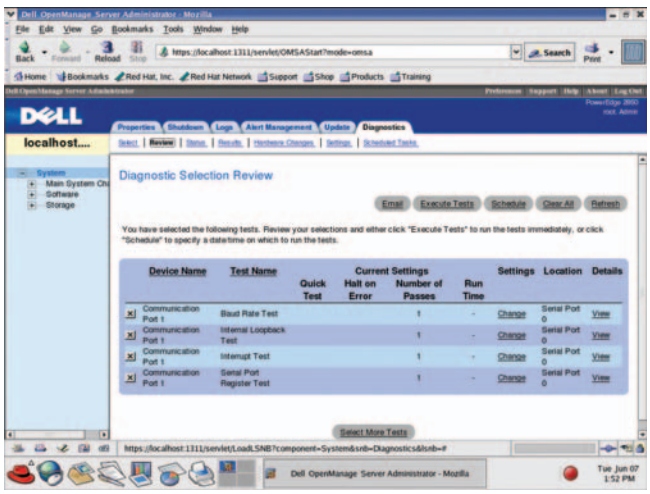


Figure 3. Dell OpenManage Server Administrator Diagnostic Selection Review screen

The diagnostic result history feature allows administrators to view the result history log. This log file contains the results of previously run diagnostics tests. This log can help administrators monitor events such as warnings, device failures, and the working condition of the device under test. The log file size can be configured to a maximum of 5 MB. Figure 5 shows the Diagnostic Result History screen.

Hardware configuration changes and change history

The hardware configuration changes feature gives administrators the option to view changes that have occurred to testable devices on the system since the last reboot, restart of the secure port server, or re-enumeration. It also reports changes in system configuration such as the addition or removal of a hard drive. Figure 6 displays the Diagnostic Hardware Configuration Changes screen.

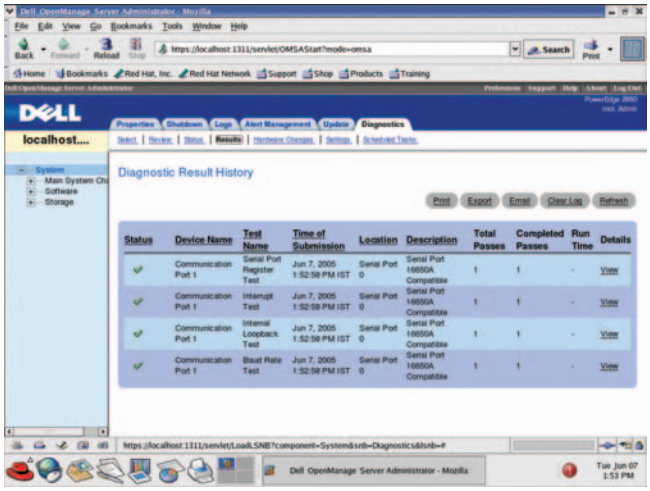


Figure 5. Dell OpenManage Server Administrator Diagnostic Result History screen

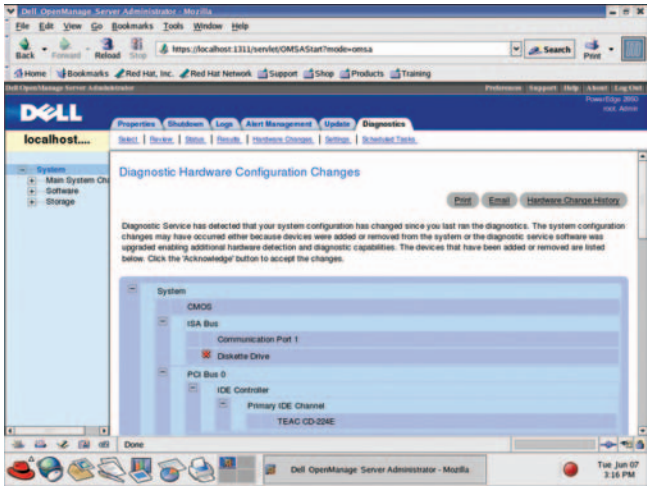


Figure 6. Dell OpenManage Server Administrator Diagnostic Hardware Configuration Changes screen

The change history feature enables administrators to view a log file that contains a history of hardware configuration changes. The log file size can be configured to a maximum size of 5 MB. Figure 7 shows the Diagnostic Hardware Configuration Change History screen.

Configuration options

The Dell OpenManage Online Diagnostics program can specify options for running diagnostic tests. Both application settings and test-execution settings can be specified. Figure 8 shows the Diagnostic Application Settings screen.

Options available under application settings include remote method invocation (RMI) registry port, result history log size, and hardware change history log size. Options available under test-execution settings include halt execution on first error, quick test, the number of passes, and runtime.

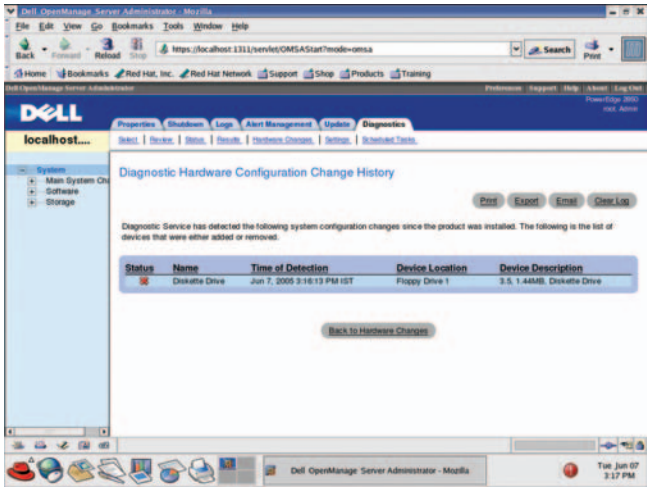


Figure 7. Dell OpenManage Server Administrator Diagnostic Hardware Configuration Change History screen

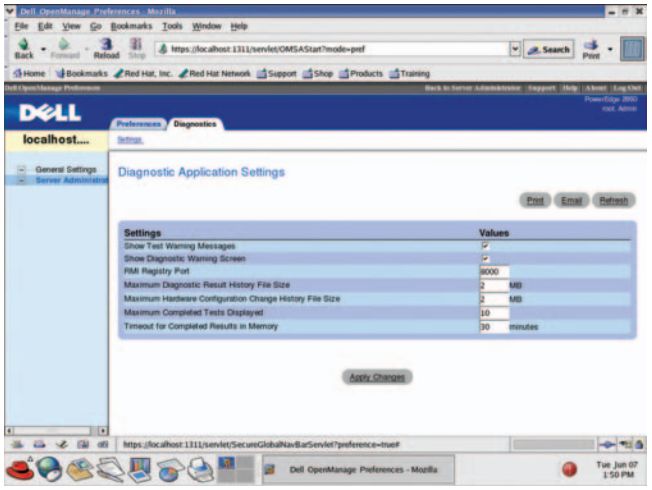


Figure 8. Dell OpenManage Server Administrator Diagnostic Application Settings screen

A powerful systems management tool

The Dell OpenManage Online Diagnostics program is an integrated component within Dell OpenManage Server Administrator enterprise server management software. It provides a rapid method by which administrators can diagnose hardware malfunctions and identify solutions to such problems. IT administrators can use the Online Diagnostics program to help reduce management costs and enhance server management.

Prathap Thathireddy is a senior engineering analyst in the Product Group Test Engineering Group at Dell. He has a B.S. in Computer Maintenance and Engineering from Osmania University in Hyderabad, India. He has seven years of IT experience as a system administrator, technology consultant for storage software, and test engineer.

Srikrishna Sridhar Murthy is an engineering analyst in the Online Diagnostics Group at Dell. He has a B.E. in Computer Science from Birla Institute of Technology and Science in Pilani, India.

FOR MORE INFORMATION

Dell OpenManage:
www.dell.com/openmanage

Automated OS Deployment

Using the Dell OpenManage Deployment Toolkit and Microsoft WinPE

Rapid OS deployment has always been a challenging task in enterprise IT environments. In particular, deploying a server OS can be complicated because it may require first configuring BIOS, RAID, and remote access settings. This article shows how Microsoft® Windows® Preinstallation Environment can be combined with the Dell™ OpenManage™ Deployment Toolkit to help deploy Windows operating systems on Dell PowerEdge™ servers across the enterprise.

BY RAVIKANTH CHAGANTI AND JATIN N. MUDDU

Related Categories:

Dell OpenManage

Enterprise management

Microsoft Windows

System deployment

Systems management

Windows pre-installation environment (WinPE)

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

IT administrators can efficiently install operating systems on multiple servers by scripting most of the deployment operations. Automating pre-OS installation activities, such as the configuration of hardware-specific settings, can help streamline server deployment and often can be completed using original equipment manufacturer (OEM)–supplied utilities.

The Dell OpenManage Deployment Toolkit (DTK) 2.0 is designed to provide a simple interface for scripted and automated configuration of Dell PowerEdge servers before OS deployment. The DTK contains a set of tools and sample scripts that can help accomplish most deployment tasks. The DTK 2.0 for Microsoft Windows Preinstallation Environment (WinPE) is designed to be integrated into a WinPE environment[1] and is available for download as a self-extractable zip file on the Dell support Web site (support.dell.com). This zip file contains DTK tools, sample scripts, template answer files, drivers, and documentation.

Activities involved in a typical server deployment scenario include the following:

- Configuration of the BIOS and baseboard management controller (BMC) settings of the server
- Configuration of the remote access controller (RAC) if a RAC is present
- Configuration of the RAID controller to create virtual disks if a RAID controller is present
- Configuration of the hard disk (creation of partitions)
- Deployment of the OS

[1] The DTK 2.0 has been developed and validated using WinPE 2005 only.

Tool	Purpose
Syscfg	Configure and replicate BIOS and BMC settings
Racadm	Configure and replicate RAC settings
Raidcfg	Create virtual disks on a RAID controller

Figure 1. Tools available within the Dell OpenManage Deployment Toolkit

The DTK can help administrators accomplish these tasks. Furthermore, by integrating the DTK and customized scripts with WinPE, administrators can automate OS deployments.

Configuring BIOS, BMC, RAC, and RAID settings with the DTK

Figure 1 lists the tools available within the DTK. These tools—when run from the WinPE command-line interface (CLI)—can be used to set or retrieve specific settings on a Dell PowerEdge server and are designed to be used in scripts to automate deployment operations. For example, executing the following command retrieves the boot sequence settings from the system:

```
syscfg --bootseq>
```

Executing the following command sets the first alert destination of the BMC to the specified IP address:

```
syscfg lcp --alertdest=1 --destipaddr=ipaddress
```

When several options need to be set, administrators can write a simple batch script that contains all the commands for those options. When executed, the script will automatically set the required options. For a list of supported commands and options, refer the *Dell OpenManage Deployment Toolkit Command Line Interface Reference Guide* at support.dell.com/support/edocs/software/dtk/1.4/en/CLI/index.htm.

Another method for configuring settings on multiple PowerEdge servers is to manually configure a server, referred to as the master server, with all the required optimal settings and then use the `--outfile` option to capture these settings to a file. The captured settings can then be replicated on another PowerEdge server, referred to as the target server, by using the `--infile` option.

To help IT administrators perform this activity, the DTK contains two scripts for capturing the BIOS, BMC, and RAC settings to a file (SYSCAP.BAT and RACCAP.BAT) and two more scripts to replicate the captured settings from the master server to the target server (SYSREP.BAT and RACREP.BAT). *Note:* The DTK currently does not support replication of RAID settings.

The sample scripts provided with the DTK can perform very basic (atomic) operations and should be customized. This may involve changing a few environment variables or even rewriting the entire script. A master deployment script that can initiate and control the deployment process should also be written. This script can be very simple—calling the sample scripts provided with the DTK one after another, depending on the requirements—or it can be very complicated, performing more than just server configuration. The master deployment script should be capable of mounting a network share so that the captured settings from the master server can be stored in a file on the network share, which is then accessible when replicating the same settings on the target systems. This script should also have logic coded into it so that it can differentiate the master server from the target servers. This will make the script usable for both capture and replication purposes. Figure 2 shows an example code snippet demonstrating this logic.

The raidcfg utility included in the DTK can help IT administrators create virtual disks on the target server. Administrators can write scripts to achieve this goal by using the raidcfg utility and the RAIDCFG.BAT sample script. The DTK also contains sample scripts to partition the hard disk on a target server (PARTCFG.BAT) and to initiate the OS installation on a server (WININST.BAT). Best practices recommend modifying these scripts to meet specific deployment needs.

Running DTK scripts from WinPE

After writing the required scripts and determining the appropriate deployment strategy, administrators should integrate the scripts into a WinPE image. WinPE is designed to help administrators perform Windows OS deployment and maintenance activities such as virus scanning, data backup, disk imaging, and hardware configuration and diagnostics.[2] WinPE is based on the Windows kernel running in protected mode and includes a minimal Win32 subsystem with limited services. This section outlines the general procedure for creating a customized WinPE image, integrating DTK components, and finally creating an ISO image that can be used with Microsoft Remote Installation Services (RIS) to boot servers into WinPE over a network.

```
for /F %%i in ('X:\Dell\Toolkit\Tools\SYSCFG
   --svctag') set %%i
If "%svctag%" == "service tag of the Master
   server" goto capturesettings
.... code to perform deployment .....
```

Figure 2. Code snippet from an example master deployment script

[2] For more information about WinPE, visit www.microsoft.com/whdc/system/winpreinst/default.mspx.

Creating a WinPE image

Software requirements for creating a WinPE image include the following:

- WinPE CD
- Microsoft Windows OS installation CD[3]
- Additional mass-storage drivers that are not part of the Windows OS installation CD
- Additional drivers (for the network adapter, instrumentation, storage, and RAC) that are not a part of the Windows OS installation CD

The next step is to build the WinPE directory structure. Administrators can do so by performing the following steps:

1. Place the Windows OEM Preinstallation Kit (OPK) CD in the CD drive of the management station.
2. Create a directory on the hard disk of the management station to store the WinPE build tools, denoted as *build_location*. For example:

```
md c:\build_x86
```

3. Remove the Windows OPK CD and place the WinPE CD in the CD drive, denoted as *cd_drive*.
4. Copy *cd_drive*\WinPE and all subdirectories to *build_location*. For example:

```
xcopy e:\WinPE c:\build_x86 /s
```

5. Remove the WinPE CD and place the Windows OS installation CD in the CD drive.
6. Navigate to *build_location*. For example:

```
cd c:\build_x86
```

7. Run the `mkimg` command with the two required options: the path to the OS installation CD and the target WinPE directory. For example:

```
mkimg.cmd e:\ c:\WinPE
```

Integrating DTK components and drivers

The DTK zip file can be extracted onto the hard disk of a management station—preferably the same system where the WinPE image is built. Best practices recommend that the files be extracted to a folder that is shared across the network and that this network share be mounted by the master deployment script, as discussed earlier in this article. The contents of the DTK zip file can be extracted to the WinPE image's root folder as well. Refer to the DTK documentation for more details about the recommended deployment directory structure.

The DTK zip file contains drivers for instrumentation, network adapters, mass storage, and RACs within the Dell\Drivers folder. Most of these drivers are required for the DTK tools to function correctly. The DRIVERINST.BAT script, which is also located in the Dell\Drivers folder, can be used to install these drivers into the WinPE image. Using this script is very simple—for example, executing the following command will install all the drivers into the WinPE image:

```
c:\WinPE\dell\drivers\driverinst.bat c:\WinPE\
    dell\drivers c:\WinPE
```

Note: Simply extracting the drivers into the WinPE image is not sufficient; the drivers need to be installed using the DRIVERINST.BAT script. Also, some services need to be started during boot-up. Refer to the DTK documentation for information about how to start these services automatically at startup.

Next, administrators should copy the master deployment script to the WinPE image and modify winbom.ini (located in the root WinPE image folder—for example, c:\WinPE) to automatically start executing the master deployment script after booting the system into WinPE.

Creating an ISO image

After using `mkimg` to create the WinPE image, administrators can customize this image. They also can use the oscdimg tool (available in the c:\WinPE folder) to create an ISO image file from the customized image, which they can then burn to a CD.

1. Run the `oscdimg` command with the `-b` option, which specifies location, and the `-n` option, which specifies the path to the WinPE directory and the image file path and name. For example:

```
oscdimg -b c:\build_x86\etfsboot.com -n
    c:\WinPE c:\WinPEx86.iso
```

   Note: In this example, the `-b` option specifies the path to the EL Torito boot sector file and the `-n` option enables support for long file names. This command creates an ISO file in the location specified.
2. Use any CD-recording software, such as Roxio, to burn the ISO image file to a blank CD and create a bootable WinPE CD.

[3] WinPE 2005 supports the 32-bit version of Microsoft Windows Server™ 2003 Service Pack 1 (SP1) as well as Windows Server 2003 x64 Editions; however, the DTK 2.0 supports only the 32-bit version of Windows Server 2003 SP1.

```
[SetupData]
BootDevice = "ramdisk(0)"
BootPath = "\platform\System32\"
OsLoadOptions = "/noguiboot /fastdetect /minint /rdexportascd /rdpath=bootimage"
Architecture = "platform"
[OSChooser]
Description = "Dell OpenManage Deployment Tool Kit"
Help = "This option will load WinPE RAMDISK containing DTK tool set & scripts"
LaunchFile = "%INSTALLPATH%\%MACHINETYPE%\templates\startrom.com"
ImageType = Flat
Version = "5.2 (0)"
```

Figure 3. Example Winnt.sif file

Integrating the WinPE ISO image with RIS

Windows RIS provides network administrators with the capability to easily install a base OS onto a new PC or to replace a system that has failed. RIS allows administrators to perform these tasks without having to visit each individual machine. RIS also provides a centralized location to integrate maintenance and troubleshooting tools that are accessible through a network boot.

Combining WinPE with Windows RIS enables the WinPE image to serve as the pre-OS environment for the network boot machines. This implementation is designed to help deploy Windows operating systems over a TCP/IP network. To integrate the WinPE image with RIS, administrators should perform the following steps:

1. Browse to the location where RIS images are installed—for example, *Server_name*\REMINST\Setup*Language*\Images.
2. Create a folder to which to copy the WinPE image.
3. Navigate to the newly create folder.
4. Create another folder named "i386."
5. Copy the WinPE image created earlier to the i386 folder.
6. Create a folder named "templates" within the i386 folder.
7. Navigate to the *build_location* folder created earlier—for example, c:\build_x86.
8. Copy ntdetect.com to the templates folder.
9. Copy setupldr.exe to the templates folder and rename it "ntldr."
10. Create a text file named "Winnt.sif" within the templates folder and modify it to meet the specific requirements of the deployment. Figure 3 shows an example Winnt.sif file. *Note:* In this example file, *platform* refers to the machine architecture (for example, i386 for x86 or IA-32 architecture), and *bootimage* refers to the path to the WinPE ISO image file.
11. Restart the binlsvc service.

Once these steps are completed, administrators should reboot all the required servers using Preboot Execution Environment (PXE). The servers should then boot into the customized WinPE image, and the master deployment script should begin executing automatically either to capture the settings from a master system or to replicate captured settings to target systems.

Enhancing server deployment across the enterprise

The Dell OpenManage Deployment Toolkit used in conjunction with Microsoft Windows Preinstallation Environment can help simplify the process of deploying multiple similarly configured servers. By scripting common tasks with the DTK tools and integrating scripts into a WinPE image, administrators can minimize the amount of time required to deploy servers as well as the potential errors that can occur during a manual deployment.

Ravikanth Chaganti is a senior engineer with the Factory Install Development team at the Dell Bangalore Development Center. He has a bachelor's degree in science from Andhra University in India and has various IT certifications, including Microsoft Certified Systems Engineer (MCSE), Cisco Certified Network Associate (CCNA), and Sun Certified System Administrator (SCSA).

Jatin N. Muddu is an engineer with the DTK development team at the Dell Bangalore Development Center. His areas of interest include system programming and server architecture.

FOR MORE INFORMATION

Microsoft WinPE:
www.microsoft.com/whdc/system/winpreinst/WindowsPE_over.mspx

Deploying Oracle Database 10*g* with

Altiris Deployment Solution for Dell Servers

Using Altiris® Deployment Solution™ software, administrators can provision a Dell™ PowerEdge™ server from bare metal by configuring components such as BIOS and RAID hardware and deploying applications such as Oracle® Database 10*g* software. This article discusses the bare-metal deployment process, focusing on the remote installation of Oracle Database 10*g* on Microsoft® Windows® and Linux® operating systems.

BY MAHMOUD AHMADIAN, CHETHAN KUMAR, AND ERIC SZEWCZYK

Related Categories:

Altiris

Dell PowerEdge servers

Oracle

System deployment

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

Altiris and Dell offer an integrated solution for bare-metal deployment of Dell PowerEdge servers that leverages the Dell OpenManage™ Deployment Toolkit (DTK). Altiris Deployment Solution for Dell Servers, a component of the Altiris Server Management Suite™, offers administrators a single management framework for complete, automated server builds and can help significantly reduce server build times from hours to minutes. Altiris automates the build process of both hardware and software—including scripted or image-based OS installation of both Microsoft Windows and Linux and installation of popular software applications such as Oracle Database 10*g*.

Altiris can also update hardware components as part of the provisioning process using Dell Update Packages to help ensure that all server components are up-to-date. The entire provisioning process can be encapsulated into a single job that can be easily executed in a one-to-many deployment by simply dragging and dropping the job onto the managed computer(s) in the Altiris Deployment Solution console.

Dell and Oracle provide validated configurations of Dell servers and storage with Oracle databases. Extensive system testing is involved in this validation process to help ensure that all components—down to the specific firmware and driver versions—are tested to work in a Dell and Oracle database environment. This article explores the integration of Oracle Database 10*g* with Altiris Deployment Solution for Dell Servers to provide a single, integrated installation.

Altiris Deployment Solution for Dell Servers

Altiris Deployment Solution for Dell Servers lets administrators fully configure hardware components such as the BIOS, Dell Remote Access Controller (DRAC), baseboard management controller, and PowerEdge Expandable RAID Controller (PERC) in a pre-OS environment leveraging the bundled Dell tools. Altiris Deployment Solution can also create the 32 MB Dell Utility partition. The end result is an automated approach to provisioning Dell PowerEdge servers that can help ensure configuration consistency and enable rapid server builds in hours or minutes.

Altiris Deployment Solution for Dell Servers is an add-on to the Altiris Deployment Solution. Altiris offers the flexibility of installing this add-on as part of a bundled installation (including Altiris Deployment Server) for first-time users, or by installing it separately for existing users

who may already have an Altiris Deployment Server infrastructure in place. This Dell-specific add-on provides predefined drag-and-drop jobs for fifth-generation and later PowerEdge models, as supported by the DTK.

Oracle Database 10*g*

Oracle is a leading provider of enterprise-class database and application software. Many hardware components are available for database administrators (DBAs) when choosing the appropriate hardware for a database server. Adding to the complexity is the installation of components that require knowledge that exceeds traditional DBA skills—hardware configuration, operating systems, drivers, patches, and any third-party software required for management of such components. Dell has tested and validated configurations to help reduce this complexity, and the Solution Deliverable List[1] tells administrators exactly which revision of drivers and patches have been validated for a given release.

Dell-recommended database solutions can help enterprises manage vital data. Dell servers use industry-standard hardware components that are designed to provide high performance, reliability, and high availability. These servers can power business-critical applications and enable scalability. Whether enterprises need an affordable entry-level server or a high-performance server for mission-critical applications, Dell offers a range of server products that are designed to meet various business needs. In addition, industry-standard Dell hardware platforms can help ease server installation and management.

Oracle Database 10*g* has been designed for scalability and high availability. A high degree of scalability also can be achieved by adding servers and storage in Oracle Real Application Clusters. Oracle Database 10*g* is available in Standard Edition One (SE1) and Enterprise Edition.

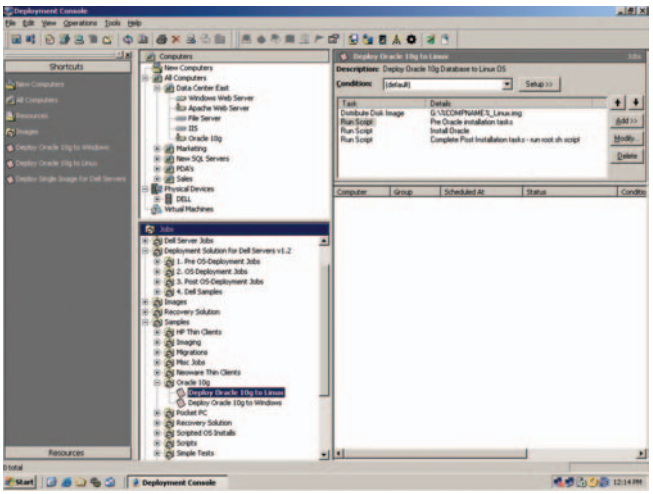


Figure 1. Altiris Deployment Solution console showing Jobs pane

Several features in Oracle Database 10*g* SE1 make it easy to use and manage. One of these features is Automatic Storage Management (ASM), which provides storage fault tolerance, volume management, and disk load balancing. ASM allows enterprises to start with a small storage infrastructure and grow as needed without additional administrative and tuning efforts. Another feature is Flashback Database, which allows error or failure recovery with an automated flashback to a specific point in time. A third feature of Oracle Database 10*g* SE1 is its monitoring and statistics-based self-tuning functionality. All these features are designed to significantly reduce the complexity of managing an Oracle database solution, thereby helping reduce total cost of ownership and increase return on investment.

Hardware deployment: Configuring the Dell PowerEdge servers

After installing the product, a folder named "Deployment Solution for Dell Servers" displays in the Jobs pane of the Altiris Deployment Solution console. This folder contains the predefined Dell jobs, which are organized into four subfolders: Pre OS-Deployment Jobs, OS Deployment Jobs, Post OS-Deployment Jobs, and Dell Samples (see Figure 1). Hardware provisioning tasks are handled by the pre-OS deployment jobs.

Altiris Deployment Solution manages the RAID configuration automatically based on the type of RAID controller detected and the number of drives in the system. For example, a PowerEdge 1855 server typically has two hard drives, so by default a RAID-1 configuration would be implemented. If an administrator uses a PowerEdge server that accommodates three or more drives, then a RAID-5 configuration would be applied. These defaults can be overridden by the Altiris administrator, but at the time of this writing version 1.4 of the DTK bundled with Altiris Deployment Solution for Dell Servers version 1.2B currently limits available options to RAID-0, RAID-1, or RAID-5.

When administrators deploy Oracle 10*g* from bare metal, Dell best practices recommend using a RAID-10 configuration for all data disks for optimal I/O performance. This RAID level requires a minimum of four disks. Altiris Deployment Solution can use a vendor-specific utility, such as MegaRC from LSI Logic, to configure the needed RAID level.[2] A powerful feature of Altiris Deployment Solution is its ability to execute custom scripts in a DOS, Windows Preinstallation Environment (WinPE), Linux Fedora–based pre-OS environment, or any post-OS environment that Altiris supports. The next release of the DTK version 2.0 and the subsequent release of

[1] The Solution Deliverable List is accessible by visiting www.dell.com/10g, clicking "Oracle Database Standard Edition One" under "Microsoft Windows 2003 SP1" and then clicking "Dell | Oracle Database Standard Edition One version 1.2 on Microsoft Windows 2003 SP1 (updated 7/1/2005)".

[2] For more information and to download the MegaRC utility, visit www.lsilogic.com/downloads/downloads.do?product=2838&download_type=misc.

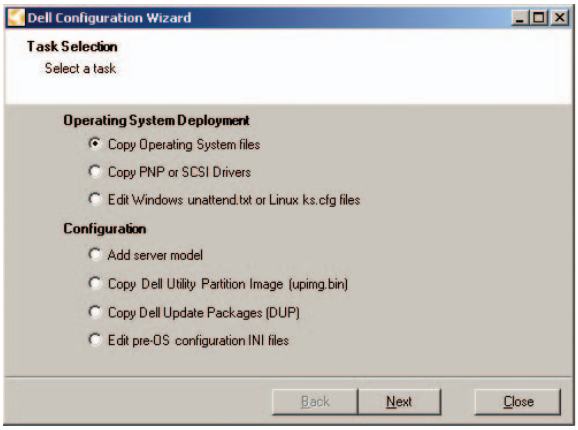


Figure 2. Using the Dell Configuration Wizard for OS installation

Deployment Solution for Dell Servers version 2.0 will offer native RAID-10 and RAID-50 support as well as bare-metal Windows and Linux preinstallation support for eighth-generation servers.

Software deployment: Installing the OS

After the Altiris Deployment Solution job is modified to configure the Dell server hardware and install a Windows or Linux OS, administrators can add software deployment tasks. Any software that can be scripted to run unattended can be deployed with Altiris Deployment Solution. Oracle 10*g* supports unattended installation.

Dell software engineers relied upon Altiris Deployment Solution during their early testing and validation of Oracle Database 10*g* configurations in July 2005. Specifically, the engineers needed to redeploy both Windows and Linux operating systems repeatedly after each test of the Oracle 10*g* installation to help ensure that the OS was ready for the next test. To do this, Dell engineers installed Windows and Linux reference systems and then captured their images using Altiris Deployment Solution. After the initial images had been captured, individual test systems could be rapidly and remotely reset for the next test to ensure the OS was clean. Dell engineers simply dragged and dropped the respective "Deploy Image" jobs for each OS onto various managed test servers displayed in the Altiris Deployment Solution console. The ability to quickly and easily reset the test environment helped reduce the amount of time required to validate Oracle 10*g* installations on Dell server hardware.

For enterprises that choose to implement scripted OS installations rather than imaging installations, the Dell Configuration Wizard function of the Dell Deployment add-on can help administrators complete the necessary setup tasks. Such tasks include copying OS source files and plug-and-play or SCSI drivers to their respective locations on the deployment server and editing the Windows unattend.txt or Linux kickstart files that act as the answer files for the OS builds (see Figure 2).

Imaging can perform OS deployments in minutes, whereas scripted installations take a little longer, permitting administrators to introduce service packs, hot fixes, and upgraded drivers into OS source files for easy maintenance. As mentioned earlier, Altiris supports both installation methods.

Database deployment: Installing Oracle Database 10*g*

Oracle provides a mechanism to install Oracle Database 10*g* silently. Silent installation runs in the background and does not require any input from the DBA. The typical interactive graphical user interface (GUI)–based menus are not displayed during installation. Instead, all options required for installing Oracle 10*g* are supplied to the Oracle Universal Installer (OUI) in a text-based response file. A sample of the response file for the database installation is found in the root of the Oracle 10*g* installation CD.

Installing Oracle Database 10*g* on Windows

Oracle 10*g* installation on Windows is straightforward. Dell created a custom Oracle 10*g* on Windows installation job for the Altiris Deployment Solution console, which performed the following tasks:

- Deployed an OS with the Altiris agent (AClient) preinstalled
- Created a local group—ORA_DBA—on the target server
- Added the local administrator account to the group
- Mapped a drive to the Altiris eXpress share to initiate the silent installation

The response file for the Oracle 10*g* installation is supplied to the OUI as a command-line option in the Altiris Run Script task. The `nowait` option is used to instruct the OUI not to wait for user prompts at the end of the installation, making this a truly unattended installation.

The sections of the Run Script task shown in Figures 3 and 4 initiate the Oracle 10*g* installation for Windows. Figure 3 shows the

```
net localgroup /add ORA_DBA / Comment "Oracle
   DBA group"
net localgroup ORA_DBA /add administrator
```

Figure 3. Script for creating the ORA_DBA group on Windows

```
net use O:\\Altiris_Server\Express
   /USER:Installer Password

O:\Deploy\Oracle10g_Windows\setup.exe -silent
   -nowait -responseFile
O:\Deploy\Oracle10g_Windows\se1.rsp
```

Figure 4. Script for mapping the drive and installing Oracle 10*g* on Windows


```
Create a mount point for the shared drive on the
   Altiris DS machine
mkdir /mnt/Altiris

Mount the shared drive on the client machine
mount -t smbfs //Altiris_Server/Express
   /mnt/Altiris -o username=installer,
   password=password

Change the current path to the folder containing
   the scripts for implementing pre-installation
   tasks
cd /mnt/Altiris/Deploy/Oracle10g_Linux/pre-install

Run the scripts
./users.py
./kernel_config.py
```

Figure 5. Script for performing Oracle preinstallation tasks on Linux

```
su - oracle "-c /mnt/Altiris/Deploy/Oracle10g_
   Linux/runInstaller - silent -waitforcomple-
   tion -responseFile <response file>
```

Figure 6. Script for installing Oracle 10*g* on Linux from the Altiris eXpress mount point

```
Change the path to the folder containing Oracle
   database files (e.g., / if the default path
   is used for installation)
cd /opt/oracle/product/10.1.0/db_1/

Rename the original root.sh script
mv root.sh root_promptuser.sh

Copy the customized root.sh script to the
   ORACLE_RDBMS_HOME folder
cp /mnt/Altiris/Deploy/Oracle10g_Linux/
   post-scripts/root.sh

Execute the customized root.sh script to provide
   answers to user prompts
./root.sh
```

Figure 7. Script for completing the Oracle postinstallation tasks on Linux

script for creating the ORA_DBA group on the target server; Figure 4 shows the script for mapping the drive to the Altiris eXpress share and installing Oracle 10*g*.

Installing Oracle Database 10*g* on Linux

Installing Oracle Database 10*g* on Linux is somewhat more complicated than installing this software on Windows. Tasks are grouped into preinstallation, Oracle installation, and postinstallation categories.

Dell created a custom job for the Altiris Deployment Solution console with tasks to deploy their precaptured image of Red Hat® Enterprise Linux with the Altiris Deployment Agent for Linux (ADLAgent) preinstalled. Other tasks included execution of preinstallation Python scripts to create a local Oracle user as well as the Oinstall and DBA groups. Creating directories for the Oracle binaries and mapping points to the Altiris eXpress share and setting permissions on the directories are also required.

As part of preinstallation, Oracle recommends setting shell limits and configuring certain kernel parameters.[3] The remaining tasks initiate the Oracle 10*g* installation from the mounted Altiris eXpress share and execute the necessary postinstallation tasks. Figures 5, 6, and 7 show the scripts for Oracle 10*g* installation on Linux.

A simple, cost-effective deployment solution

The integrated server deployment process tested jointly by Dell and Altiris and described in this article verifies that the Altiris Deployment Solution framework can effectively support an Oracle Database 10*g* installation.[4] This process provides administrators with a simple, cost-effective method for deploying Oracle databases on Dell PowerEdge servers. By integrating and automating the server build and database deployment processes, Altiris Deployment Solution for Dell Servers helps streamline the installation and configuration of Oracle Database 10*g* on Dell hardware—simplifying administrators' duties significantly and allowing them to focus on other business-critical tasks.

Mahmoud Ahmadian is an engineering consultant with the Database and Applications team of the Dell Product Group. He has an M.S. in Computer Science from the University of Houston, Clear Lake.

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

Eric Szewczyk is a Dell Alliance Technical Strategist for Altiris. He helps manage the Dell IT relationship in addition to training Dell systems engineers using Altiris management software. Eric has a B.A. from the University of Central Oklahoma and is an Altiris Certified Engineer.

[3] For more information about Oracle preinstallation recommendations, refer to the appropriate Oracle 10*g* documentation available at www.oracle.com/technology/documentation/database10g.html.

[4] Samples of the jobs used to test and validate Oracle Database 10*g* deployment on Dell PowerEdge servers have been provided as part of Altiris Deployment Solution 6.5.

THE GRID STARTS HERE

This year the power and value of the Dell | Oracle GRID solution has come into sharper focus with the merging of key 64-bit technologies.

With the advent of Dell™ PowerEdge™ servers powered by 64-bit Intel® Xeon® processors and featuring Oracle® 10g and Microsoft® Windows Server™ x64 Editions, enterprises have an affordable, easy-to-deploy, scalable database infrastructure.

With x64 architecture and significant improvements in Windows Server 2003, Dell's customers can now take advantage of Oracle scalability, performance, and reliability never before available in a Microsoft® Windows® environment.

The power of a scalable, standardized 64-bit database solution

In the past, many companies have relied on legacy SMP systems to scale to large user populations. In fact, test results show that there are few limitations on how Oracle memory can be configured for 64-bit databases when a 64-bit Windows operating system is used.* It is possible to support large user populations, with 1300 users connected via Dedicated Server connections in the tests. Even larger user populations may be supported with more memory and the use of Shared Server connections. Although test results for report queries are not yet conclusive, environments that feature long-running report queries and Data Warehouse activity should also benefit from performance improvements due to the 64-bit memory model.

The platform of choice for transitioning your database to 64-bit

Dell PowerEdge servers offer a powerful platform for deploying enterprise class applications, such as Oracle databases. Key performance improvements over previous Dell server lines include increased processor speed, improved parallel processing performance, faster memory, the use of fast DDR-2 RAM, and the use of PCI-Express technology. In addition, Dell fully exploits the 64-bit extension technology offered by Intel® EM64T processors. The Dell PowerEdge server supports Oracle databases on the Microsoft Windows 2000 Server and Windows Server™ 2003 32-bit operating systems. Combined with Microsoft® Windows Server™ 2003 for extended 64-bit architectures and the Oracle Database 10*g* for extended 64-bit architectures, Dell PowerEdge servers provide the perfect platform for transitioning to 64-bit databases.

Expect performance improvements

With the release of the extended 64-bit versions of Oracle and Microsoft Windows Server 2003, the greatest performance gains can be realized by migrating to 64-bit databases. For those companies that are evaluating Oracle databases on Microsoft Windows, the potential scalability improvements should be carefully considered. With Dell PowerEdge servers and Microsoft Windows Server 2003 for the Intel® EM64T architecture, the time to implement enterprise-class Oracle databases on Microsoft Windows has arrived.

"We have effectively lowered the barrier to entry for customers looking to deploy a high-performance server for database environments."

Jeff Clarke, Sr. VP
Enterprise Product Group
Dell Inc.

"Now is the time to get an affordable enterprise-class, clustered, scalable, high-performing database on Microsoft Windows. All from Dell."

TJ Lamphier, Sr. Manager
Enterprise Product Group
Dell Inc.

GET MORE OUT OF YOUR DATABASE

For more information contact your Dell representative and visit www.dell.com/oracle

*All technical content has been drawn from Comparison of 32-bit and 64-bit Oracle Database Performance on the Dell PowerEdge 6850 Server with Microsoft Windows Server 2003, a whitepaper sponsored by Dell, Larry Pedigo, Performance Tuning Corporation (July 2005). Read the complete whitepaper, including test results for scaling both 32-bit and 64-bit versions of Oracle databases for large user populations, at: www.dell.com/oracle

SMASH Command-Line Protocol:

Setup and Configuration Considerations

Systems Management Architecture for Server Hardware Command-Line Protocol (SMASH CLP) relies on Telnet or Secure Shell to enable connectivity between the management client and the managed server. This article discusses the requirements and settings to deploy SMASH CLP on Dell™ PowerEdge™ servers.

BY NATHAN RAKOFF AND JAVIER JIMENEZ

Related Categories:

Change management

Command-line interface (CLI)

Out-of-band (OOB) management

Systems Management Architecture for Server Hardware (SMASH)

Systems management

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

Systems Management Architecture for Server Hardware (SMASH) is the Distributed Management Task Force (DMTF) initiative to unify data center server management by defining architectural semantic specifications.[1] Dell and other major server hardware vendors have been actively participating in the DMTF Server Management Work Group (SMWG) to establish the SMASH specifications. SMASH Command-Line Protocol (SMASH CLP), also known as Server Management CLP (SM-CLP), is one of the standards that has been produced by the SMWG. Using standardized SMASH CLP commands, system administrators can monitor and manage heterogeneous computing environments remotely.

Dell has been actively developing server management systems and tools that support SMASH standards. Because these standards are scheduled to be completed in 2006, Dell SMASH CLP will first be made available as an evaluation product to select customers. The Dell SMASH CLP package includes in-band management, which is standardized systems management on a running OS; and out-of-band (OOB) proxy management, which is remote service-processor systems management driven via an in-band proxy server.

Dell SMASH CLP can be deployed over Telnet or Secure Shell (SSH). However, Telnet lacks security, and thus SSH is recommended because it provides a secure point-to-point communication tunnel. Several implementations of SSH are available, any of which can be used as a secure protocol for SMASH CLP communication. Figure 1 provides a topological view of the SMASH CLP communication capabilities. Administrators can access an in-band Dell PowerEdge server or use an in-band PowerEdge server as a proxy to an OOB service processor such as a Dell Remote Access Controller (DRAC) or a baseboard management controller (BMC) to perform server management. This article describes key configuration settings required for installing SMASH CLP on Dell PowerEdge servers.

Preparing to install SMASH CLP

Prior to installing SMASH CLP, administrators must first install Dell OpenManage™ Server Assistant instrumentation and Telnet or SSH. Dell SMASH CLP supports Microsoft® Windows® and Linux® operating systems. Specifically supported platforms are listed in the SMASH CLP README file.

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

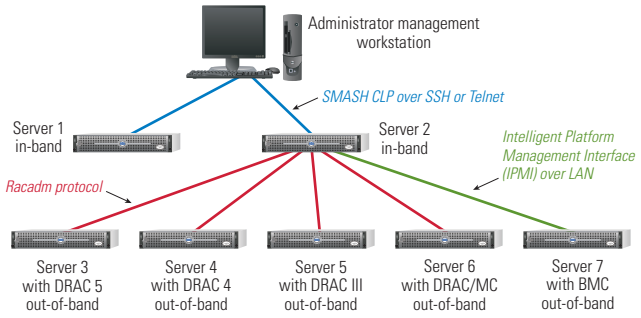


Figure 1. Server communication topology in Dell SMASH CLP environment

Sixth generation or later Dell PowerEdge servers, such as the PowerEdge 2600, PowerEdge 1750, and PowerEdge 1800, are required for running SMASH CLP. SMASH CLP provides systems management support across multiple generations of PowerEdge servers, as well as across multiple operating systems.

Configuring Telnet to support SMASH CLP

The standard Windows Telnet Server[2] can be configured to support SMASH CLP. On Microsoft Windows Server™ 2003, administrators can change the DefaultShell registry entry located under HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\TelnetServer\1.0\ to the path and file name for SMASH CLP.

The standard Linux Telnet server supports SMASH CLP in much the same way it supports the Bash shell. If a system administrator changes a user's startup shell—which is specified in the /etc/passwd file—from a standard shell to the SMASH CLP shell, that user's shell will default to SMASH CLP when connecting over Telnet or SSH.

Configuring SSH to support SMASH CLP

To use SMASH CLP over SSH, administrators must select a version of SSH to use on the PowerEdge server. OpenSSH,[3] a free version of the SSH protocol suite, supports most platforms based on the UNIX® OS—including Novell® SUSE™ Linux Enterprise Server and Red Hat® Enterprise Linux—as well as Cygwin,[4] an open source, Linux-like environment on Windows that is supported by Red Hat.

Settings for Linux systems

Both SUSE Linux Enterprise Server and Red Hat Enterprise Linux have built-in support for OpenSSH. Therefore, an administrator needs to install only the Dell SMASH CLP package. The installer will copy the SMASH CLP binaries to an appropriate directory. Once the binaries are installed, the administrator can configure a system user to enable access to the SMASH CLP shell. After creating the user, the administrator should assign the SMASH CLP binary as the startup shell for the SMASH CLP authorized user—for example, by editing the startup shell field in the system user's /etc/passwd file to point to /usr/bin/smash.

Settings for Windows systems

To use OpenSSH on Windows, administrators must install Cygwin, which uses approximately 1.5 GB of hard drive space for a full installation. However, OpenSSH support requires only a subset of the Cygwin installation—approximately 7 MB of Cygwin tools. The SSHWindows[5] package installs OpenSSH and a minimal amount of Cygwin tools. Another option for packaging OpenSSH and Cygwin is copSSH.[6]

When using OpenSSH on Cygwin, SSHWindows, or copSSH, administrators should be aware that the packages are incompatible with each other. Because SSHWindows and copSSH use a small subset of Cygwin, they depend on the same libraries, registry keys, and environment variables. If a system uses Cygwin, then administrators must install SMASH CLP over the existing Cygwin installation. Alternatively, uninstall the Cygwin environment and remove all registry keys under HKEY_CURRENT_USER\Software\Cygnus Solutions and HKEY_LOCAL_MACHINE\Software\Cygnus Solutions before installing SSHWindows or copSSH.

After installing an SSH package, the administrator should use the tools provided by the SSH package to create an authorized user and assign the SMASH CLP shell as the user's startup shell. Figure 2 shows the OpenSSH commands for setting up an authorized user.

Then, the administrator should edit the startup shell field in the c:\openssh\etc\passwd file to point to smash.exe. This binary is located in the smash/bin subdirectory of the Dell OpenManage Server Assistant directory. The Dell OpenManage Server Assistant directory is defined in the PowerEdge system's PATH environment variable.

```
mkgroup -l >> c:\openssh\etc\group
mkpasswd -l [-u username] >> c:\openssh\etc\passwd
```

Figure 2. OpenSSH commands for setting up an authorized user

[2] For more information about Windows Telnet Server tools, visit www.microsoft.com/technet/prodtechnol/windowsserver2003/library/TechRef/ddf8b035-9035-475c-ae50-1d97bde83dba.mspx.

[3] The latest version of OpenSSH is available at www.openssh.org.

[4] For more information about Cygwin, visit www.cygwin.com.

[5] The latest version of SSHWindows is available at sshwindows.sourceforge.net.

[6] The latest version of copSSH is available at www.itefix.no/phpws/index.php?module=pagemaster&PAGE_user_op=view_page&PAGE_id=12&MMN_position=22:22.

Understanding security issues

Administrators should consider their organizations' security exposure when selecting an SSH package. Three options are available:

- Download and build a custom release of the OpenSSH package
- Download the OpenSSH packaged binaries
- Buy a commercial SSH package

The first option enables security verification by allowing administrators to review the OpenSSH code base for "back doors" and build their own releases. However, unless an administrator has a corporate mandate or a significant amount of spare time to examine the source code, this option is not practical. The second option requires trust in the integrity of the OpenSSH developer who built the binaries that the binaries have not been compromised. The third option requires the least amount of effort from an administrative level but requires putting trust in the integrity of the commercial vendor. Cost, convenience, and security of the solution should be deciding factors when choosing an SSH package.

Authentication

The primary method of authenticating a user is through usernames and passwords, which can be supplied through a secure connection from any SSH client to any SSH server. The problem with using a username and password as a security measure is that passwords can be shared or written down. This compromising of system security can be ameliorated with additional security breach deterrents such as host-based authentication and digital certificates.

OpenSSH has several configuration settings that can increase or decrease the security level of the deployed system:[7]

- **LoginGraceTime:** Sets the amount of time a user has to complete authentication
- **PermitRootLogin:** Enables or disables the root user to log in to the SSH session
- **StrictModes:** Enables or disables the checking of a user's permissions on his or her home directory and rhosts files before accepting authentication
- **RSAAuthentication:** Enables or disables RSA authentication
- **PublicKeyAuthentication:** Enables or disables public-key authentication
- **AuthorizedKeysFile:** Names the path to the authorized key validation file on the SSH server
- **RhostAuthentication:** Enables or disables rhosts authentication for RSH
- **RhostRSAAuthentication:** Enables or disables rhosts authentication from RSA host keys
- **HostbasedAuthentication:** Enables or disables host-based authentication
- **PasswordAuthentication:** Enables or disables password authentication
- **PermitEmptyPasswords:** Enables or disables the ability to use a blank password

As more settings are enabled, the level of security on the SSH server increases—the more settings enabled, the higher the security.

Firewalls

By default, the OpenSSH server listens for traffic on TCP port 22. By editing the /etc/sshd_config file, administrators can change the default setting to a port that the firewall allows. Before circumventing the enterprise's firewall policy, administrators should check that port 22 cannot be opened and that running an SSH server is allowed.

Selecting the appropriate protocol for SMASH CLP connections

The system administrator's dilemma is the trade-off between security and convenience. Unlike SSH, Telnet is simple to configure. However, Telnet is too risky to use in a production environment. Even in an isolated network of Dell Remote Access Controllers, Telnet is insufficient for establishing point-to-point communication for SMASH CLP. If the isolated network is compromised, Telnet offers no secondary security measure, such as encryption.

Encrypted communication and command-line capabilities make SSH preferable to Telnet for SMASH CLP environments. When properly installed with Dell OpenManage Server Assistant instrumentation and an SSH server, Dell SMASH CLP is designed to provide basic enterprise systems management over a secure point-to-point communication tunnel.

Nathan Rakoff is a lead software engineer in the Open Standards Architecture and Management Group at Dell. His principal focus is the development of SMASH standards for Dell enterprise management solutions. He has a bachelor's degree in Computer Science from New Mexico State University.

Javier Jimenez is a lead software architect in the Open Standards Architecture and Management Group at Dell. His principal focus is the adoption of open industry standards and the development of applications to support those standards for Dell enterprise management solutions. He has a bachelor's degree in Computer Engineering from Nova Southeastern University.

[7] For more information about installing SSH securely, see *Implementing SSH: Strategies for Optimizing the Secure Shell* by Himanshu Dwivedi, published by John Wiley & Sons, 2003.

Link Aggregation Interoperability

of the Dell PowerConnect 5316M Switch and Cisco Switches

This article explains how to configure the Dell™ PowerConnect™ 5316M Gigabit Ethernet switch, which resides within the Dell Modular Server Enclosure, to interoperate and connect with Cisco IOS-based and CatOS-based switches by using industry-standard link aggregation groups that adhere to the IEEE 802.3ad standard.

BY BRUCE HOLMES

Related Categories:

Blade servers

Cisco

Command line interface (CLI)

Data networking

Dell PowerConnect switches

Dell PowerEdge blade servers

Ethernet

Internetworking

Scripting

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

The Cisco IOS and CatOS operating systems run on Cisco network switches and provide Cisco EtherChannel, Fast EtherChannel (FEC), and Gigabit EtherChannel (GEC) technologies, which enable network administrators to group ports on Cisco switches together to increase available throughput. Dell PowerConnect switches offer a similar technology known as link aggregation groups (LAGs), which are designed to increase the overall bandwidth between two Dell switches by aggregating multiple ports to act as a single, logical connection between the switches. Dell PowerConnect 5316M switches implement IEEE 802.3ad–based link aggregation, which is interoperable with Cisco EtherChannel technology.[1]

Link aggregation on Dell PowerConnect switches can be configured as either dynamic or static. The dynamic configuration uses the IEEE 802.3ad standard, which is also known as Link Aggregation Control Protocol (LACP). LACP enables a Gigabit Ethernet switch to confirm that the external switch is also configured for link aggregation. Static configuration is used when connecting the Dell PowerConnect 5316M Gigabit Ethernet switch to an external Gigabit Ethernet switch that does not support LACP. In a static configuration, a cabling or configuration mistake involving the PowerConnect 5316M or the external switch could go undetected and thus could cause undesirable network behavior. Both static and dynamic LAGs (via LACP) can detect physical link failures within the LAG and continue forwarding traffic through the other connected links within that same LAG. LACP can also detect switch or port failures that do not result in the loss of a link, helping provide a more resilient LAG. Best practices suggest using dynamic link aggregation instead of static link aggregation.

The examples presented in this article use the command-line interface (CLI) of the Dell PowerConnect 5316M to configure the switch.[2] These example configurations also can be implemented via the Web-based graphical user interface (GUI) of the PowerConnect 5316M.[3]

[1] In dynamic configurations, this interoperability is possible only via LACP, not the proprietary Cisco Port Aggregation Protocol (PAgP). Link aggregation interoperability for all Dell PowerConnect products is tested at the University of New Hampshire InterOperability Lab. This lab tests products for the Bridge Functions Consortium, which includes leading vendors of switch and networking products. For more information, visit ftp://ftp.iol.unh.edu/pub/bfc/testsuites/la.io.test.suite.pdf.

[2] For more information about configuring LAGs via the CLI., see the "Port Channel Commands" section of the *Dell PowerConnect 5316M CLI Reference Guide* at support.dell.com/support/edocs/network/PC5316M/en/CLI/portchan.htm#1016308.

[3] For more information about configuring LAGs via the GUI, see the "Defining LAG Parameters" section of the *Dell PowerConnect 5316M Ethernet Switch Module User's Guide* at support.dell.com/support/edocs/network/PC5316M/en/UG/switch.htm#1125197.

Link aggregation with Gigabit Ethernet switches

The following examples show minimal configurations necessary to establish a LAG between a Cisco IOS-based Gigabit Ethernet switch (Catalyst 3750), Cisco CatOS-based Ethernet switch (Catalyst 6509), and the Dell PowerConnect 5316M Gigabit Ethernet switch. These commands should work properly when using the default configuration of each switch. *Note:* These commands will erase any configuration data previously configured and reboot the switch.

To set the Dell PowerConnect 5316M to the default configuration, administrators should issue the following commands:

```
5316M# delete startup-config
5316M# reload
```

To set the Cisco IOS-based Catalyst 3750 switch to the default configuration, administrators should issue the following commands:

```
3750# delete flash:/config.text
3750# reload
```

To set the Cisco CatOS-based Catalyst 6509 switch to the default configuration, administrators should issue the following commands:

```
Cat_6509 (enable) clear config all
```

Please see the "Configuration limitations" section in this article for scenarios in which resetting the switches to factory defaults would be impractical.

The Dell PowerConnect 5316M can support up to eight LAGs. A port channel can have from zero to six of the external ports as members. Internal ports cannot be members of a LAG. The examples in this article use different numbers of ports in a LAG.

Best practices recommend that the ports to be aggregated on both the Cisco and Dell switches be disconnected during configuration. This will avoid any network loops being formed before the LAG is set up.

Configuring the PowerConnect 5316M external ports for dynamic link aggregation

The following example shows the Dell PowerConnect 5316M CLI commands for configuring the six external ports on the Gigabit Ethernet switch for LACP:

```
5316M(config)# interface range ethernet g11-16
5316M(config-if)# channel-group 1 mode auto
```

The first command sets the CLI mode to configure the six external Gigabit Ethernet ports (referred to in the command as `g11-16`, which represents Gigabit Ethernet ports 11 through 16). All 6 ports do not have to be selected; a LAG can have from zero to six ports, depending on the requirements of the network. The number of ports in the LAG correlates to the amount of bandwidth and redundancy achievable in the network—that is, the more ports, the more bandwidth and redundancy. A LAG can even be configured without any member ports. When ports are added to the LAG, they will be set to the configuration of the LAG.

The second command aggregates the six ports into a LAG (referred to in the command as `channel-group`), which will use LACP (referred to in the command as `mode auto`). The channel-group number, which is 1 in this example, has meaning only within the switch and is used to differentiate up to eight unique channel-groups. For each LAG created, administrators must designate it with a number between one and eight for up to eight groups. Only the external ports (11 through 16) can be part of a LAG.

Configuring a Cisco IOS-based Gigabit Ethernet switch for dynamic link aggregation

The following example shows the Cisco IOS-based switch CLI commands for configuring six ports for LACP:

```
3750(config)# interface range GigabitEthernet
   1/0/1 - 6
3750(config-if)# channel-protocol lacp
3750(config-if)# channel-group 1 mode active
```

The first command sets the CLI mode to configure six Gigabit Ethernet ports (referred to in the command as `GigabitEthernet 1/0/1 - 6`, which represents Gigabit Ethernet ports 1 through 6). The second command sets ports to use LACP as the LAG protocol (and not PAgP). The third command aggregates the six ports into a LAG (referred to in the command as `channel-group`), which will use LACP (referred to in the command as `mode active`). The channel-group number, which is 1 in this example, has meaning only within the switch and is used to differentiate unique channel-groups.

Configuring a Cisco CatOS-based Gigabit Ethernet switch for dynamic link aggregation

The following example shows the Cisco CatOS-based switch CLI commands for configuring six ports for LACP:

```
Cat_6509(enable) set channelprotocol lacp 2
Cat_6509(enable) set port lacp-channel 2/1-6
   mode active
```



```
5316M# show interfaces port-channel 1
Channel   Ports
.......   .....
ch1       Active: g(11-16)
```

Figure 1. Using the show interfaces command to confirm a LAG connection for the Dell PowerConnect 5316M

The first command sets the LAG dynamic protocol to LACP on module 2 (in this example, this module in the switch is used to validate the examples in this article; other switches may be configured differently). The second command aggregates six ports on module 2 (referred to in the command as `2/1-6`, which represents ports 1 through 6 on module 2) into a LAG (referred to in the command as `lacp-channel`), which will use LACP (referred to in the command as `mode active`).

Note: Only the "active" Cisco mode is supported for LACP interoperability with the Dell PowerConnect 5316M. The other modes ("passive," "auto," "on," and "desirable") should not be used when using LACP between a Cisco switch and the PowerConnect 5316M. This is a common configuration error.

Confirming a successful dynamic link aggregation connection with the PowerConnect 5316M

Figure 1 provides an example usage of the `show interfaces` command, which can be issued to help ensure that the Dell PowerConnect 5316M switch has established a LAG and that the LAG is connected. The output of this command shows that ports 11 through 16 are active. This confirms that there is physical link on all ports in the LAG and that the PowerConnect 5316M has communicated with the Cisco switch to successfully establish an aggregated link on ports 11 through 16 with LACP.

Confirming a successful dynamic link aggregation connection with Cisco IOS

Figure 2 provides an example of the Cisco IOS `show interfaces` command, which can be used to help ensure that the Cisco IOS-based switch has established a LAG and that the LAG is connected. The output of this command shows that ports 1/0/1 through 1/0/6 are active. This confirms that there is physical link on all ports in the LAG and that the Cisco switch has communicated with the PowerConnect 5316M switch to successfully establish an aggregated link on ports 1/0/1 through 1/0/6 with LACP.

Confirming a successful dynamic link aggregation connection with Cisco CatOS

Figure 3 provides an example usage of the Cisco CatOS `show lacp-channel info` command, which can be used to help ensure that the Cisco CatOS-based switch has established a LAG and that the LAG is connected. The output of this command shows that the status of ports 2/1 through 2/6 is "connected" and the channel mode of these ports is "active." This confirms that there is physical link on all ports in the LAG and that the Cisco switch has communicated with the PowerConnect 5316M switch to successfully establish an aggregated link on ports 2/1 through 2/6 with LACP.

Configuring the PowerConnect 5316M external ports for static link aggregation

The following example shows the Dell PowerConnect 5316M CLI commands for configuring three external ports of the Gigabit Ethernet switch for static aggregation:

```
5316M(config)# interface range ethernet g13-15
5316M(config-if)# channel-group 1 mode on
```

Note that a LAG can be configured with zero to six ports (this example uses three ports), but a port can be part of only a single LAG. The first command sets the CLI mode to configure three external Gigabit Ethernet ports (13 through 15). The second command aggregates the three ports into a static LAG. Static LAGs do not use LACP and are defined in the CLI by setting the channel-group mode to "on." The channel-group number, which is 1 in this example,

```
3750# show interfaces port-channel 1 etherchannel
Port-channel1   (Primary aggregator)

Age of the Port-channel  = 00d:01h:11m:34s
Logical slot/port        = 10/1
Number of ports          = 6
HotStandBy port          = null
Port state               = Port-channel Ag-Inuse
Protocol                 = LACP

Ports in the Port-channel:

Index  Load     Port      EC state     No of bits
------+------+---------+------------+------------
  0      00     Gi1/0/1   Active           0
  0      00     Gi1/0/2   Active           0
  0      00     Gi1/0/3   Active           0
  0      00     Gi1/0/4   Active           0
  0      00     Gi1/0/5   Active           0
  0      00     Gi1/0/6   Active           0
```

Figure 2. Using the show interfaces command to confirm a LAG connection for a Cisco IOS-based switch

```
Cat_6509> (enable) show lacp-channel info
Chan Port  Status      Channel    Admin  Speed   Duplex  Vlan
id                     mode       group
---- ----- ---------- --------- -----  -----   ------  ----
801  2/1   connected   active     395   a-1Gb   a-full   1
801  2/2   connected   active     395   a-1Gb   a-full   1
801  2/3   connected   active     395   a-1Gb   a-full   1
801  2/4   connected   active     395   a-1Gb   a-full   1
801  2/5   connected   active     395   a-1Gb   a-full   1
801  2/6   connected   active     395   a-1Gb   a-full   1
                          . . .
```

Figure 3. Using the show lacp-channel info command to confirm a LAG connection for a Cisco CatOS-based switch

has meaning only within the switch and is used to differentiate up to eight unique channel-groups. For each LAG created, administrators must designate it with a number between one and eight for up to eight groups. The internal ports that connect to the servers do not support LAGs.

Configuring a Cisco IOS-based switch for static link aggregation

The following example shows the Cisco IOS CLI commands for configuring three ports of the Cisco switch for static link aggregation:

```
3750(config)# interface range GigabitEthernet
   1/0/9 - 11
3750(config-if)# channel-group 1 mode on
```

The first command sets the CLI mode to configure three Gigabit Ethernet ports (1/0/9 through 1/0/11). The second command aggregates the three ports into a static LAG. Static LAGs do not use LACP and are defined in the Cisco CLI by setting the channel-group mode to "on." The channel-group number, which is 1 in this example, has meaning only within the switch and is used to differentiate channel-groups. The number of channel-groups supported by Cisco switches depends on the switch model.

Configuring a Cisco CatOS-based switch for static link aggregation

The Cisco CatOS CLI allows the configuration of static LAGs via LACP or PAgP commands. The following example shows the Cisco CatOS CLI LACP `channelprotocol` commands for configuring three ports of the Cisco switch for static link aggregation:

```
Cat_6509(enable) set channelprotocol lacp 2
Cat_6509(enable) set port lacp-channel 2/9-11 mode on
```

The first command sets module 2 to use the LACP commands to configure LAGs. Because a static LAG is being defined, the setting for the `channelprotocol` command does not matter. The second command configures the three Ethernet ports (2/9 through 2/11) into a static LAG. Static LAGs do not use LACP and are defined in the Cisco CLI by setting the lacp-channel mode to "on."

The following example shows the Cisco CatOS CLI PAgP `channelprotocol` commands for configuring three ports of the Cisco switch for static link aggregation using the PAgP command:

```
Cat_6509(enable) set channelprotocol pagp 2
Cat_6509(enable) set port channel 2/9-11
   mode on
```

The first command sets module 2 to use the PaGP commands to configure LAGs. As mentioned before, the setting for the `channelprotocol` command does not matter because a static LAG is being defined. The second command configures the three Ethernet ports (2/9 through 2/11) into a static LAG. Static LAGs do not use PAgP and are defined in the Cisco CLI by setting the channel mode to "on."

Confirming a successful static link aggregation connection

When LACP is not being used, only careful inspection of the Cisco and PowerConnect 5316M configurations can confirm that a static LAG has been established. Administrators can take the following steps to help confirm the connection:

1. Check that the cabling is connected to the correct ports on both switches.
2. Check that all the LAG ports have a link.
3. Use the `show running-config` command to confirm that the desired ports are in the LAG:
   - **PowerConnect 5316M:** `5316M# show running-config`
   - **Catalyst 3750:** `3750# show running-config`
   - **Catalyst 6509:** `Cat_6509(enable) show running-config`

Link aggregation with Cisco Fast Ethernet switches

Some enterprise IT organizations use Cisco Fast Ethernet (100 Mbps) network switches. In this case, they may not want to incur the expenses to replace the Cisco Fast Ethernet switches to match the high speed of the Dell PowerConnect 5316M Gigabit Ethernet switch, but they probably still want to achieve the most bandwidth possible. Because the PowerConnect 5316M switch supports auto-negotiation, administrators do not need to perform any additional steps to connect aggregated links to a Cisco Fast Ethernet switch if the Cisco switch's link aggregation ports are also set to auto-negotiation.

The ports in a Dell PowerConnect 5316M LAG are set to auto-negotiation by default. If the negotiation setting of the LAG has been changed because of a previous switch configuration, administrators can use the following command to set the LAG ports back to auto-negotiation:

```
5316M(config)# interface port-channel 1
5316M(config-if)# negotiation
```

To set the ports on a Cisco IOS-based switch to auto-negotiation, administrators can use the following commands:

```
2950(config)# interface range FastEthernet 0/1 - 3
2950(config-if)# speed auto
2950(config-if)# duplex auto
```

To set the ports on a Cisco CatOS-based switch to auto-negotiation, administrators can use the following command:

```
Cat_6509> (enable) set port speed  2/9-11  auto
```

If auto-negotiation cannot be used, both the Dell PowerConnect LAG and the Cisco switch ports in the LAG must be set to the same speed and duplex. Intermittent link failures may occur if one switch is in auto-negotiation mode and the other is forced to a certain speed and duplex.

The Dell PowerConnect 5316M LAG can be forced to 100 Mbps with the following commands:

```
5316M(config)# interface port-channel 1
5316M(config-if)# no negotiation
5316M(config-if)# speed 100
```

In this example, the LAG is referred to in the command as `port-channel 1`. The `no negotiation` command means that there is no auto-negotiation on the ports in the LAG. The `speed 100` command specifies all the ports in the LAG to be 100 Mbps. Because this is a LAG configuration, and the 802.3ad standard requires all ports in a LAG to be full duplex, administrator do not need to set the duplex to full (and in fact, cannot do so via the PowerConnect 5316M CLI). The duplex is set to full by default on LAG ports.

Note: This process differs from the Cisco IOS and CatOS methods, which require that all the ports in the LAG be configured to 100 Mbps and full duplex rather than setting the LAG to 100 Mbps. Configuring all ports in a LAG to 100 Mbps and full duplex on the Dell PowerConnect 5316M switch would have no effect because the LAG configuration takes precedence over individual port configurations.

Administrators can use the following commands to set the ports on the Cisco IOS-based switch to 100 Mbps and full duplex:

```
2950(config)# interface range FastEthernet 0/1 - 3
2950(config-if)# speed 100
2950(config-if)# duplex full
```

Administrators can use the following commands to set the ports on the Cisco CatOS-based switch to 100 Mbps and full duplex:

```
Cat_6509> (enable) set port speed 2/9-11 100
Cat_6509> (enable) set port duplex 2/9-11 full
```

Configuration limitations

Ports to be aggregated must be configured so that they are compatible with the link aggregation feature and with the switch to which they will be connected. For the Dell PowerConnect 5316M, the following limitations apply to aggregated ports (the commands to remove the configuration are shown immediately after each limitation):

- The port cannot have an IP address defined on it:
```
5316M(config)# interface Ethernet g11
5316M(config-if)# no ip address
```

- The port cannot belong to another LAG:
```
5316M(config)# interface Ethernet g11
5316M(config-if)# no channel-group
```

- The port cannot be a mirrored port:
```
5316M(config)# interface Ethernet g11
5316M(config-if)# no port monitor gxx
```

- The port cannot have GARP (Generic Attributes Registration Protocol) VLAN (virtual LAN) Registration Protocol (GVRP) enabled:
```
5316M(config)# interface Ethernet g11
5316M(config-if)# no gvrp enable
```

- The port cannot belong to an access VLAN other than the default VLAN (1):
```
5316M(config)# interface Ethernet g11
5316M(config-if)# no switchport access vlan
```

- The port cannot belong to a trunk VLAN other than the default VLAN (1):
```
5316M(config)# interface Ethernet g11
5316M(config-if)# no switchport trunk native vlan
```



```
5316M(config-if)# exit
5316M(config)# exit
5316M# show running-config
interface range ethernet g(13-16)
channel-group 1 mode on
exit
interface ethernet g11
gvrp enable
exit
```

Figure 4. Example output for the show running-config command showing GVRP enabled

```
5316M# show running-config
interface port-channel 1
speed 100
no negotiation
exit
interface ethernet g11
speed 10
no negotiation
exit
interface range ethernet g(11,13-16)
channel-group 1 mode on
exit
```

Figure 5. Example output for the show running-config command showing differing LAG and port speeds

- The internal switch ports (g1 through g10) cannot be part of a LAG. The CLI will prevent adding internal ports to a LAG.

To check the configuration of the ports on the PowerConnect 5316M, administrators can use the `show running-config` command and view the `interface Ethernet gxx` configurations, where *xx* indicates the port number. Figure 4 shows example output after this command has been issued. In this scenario, the `no gvrp enable` command would have to be issued on port g11 before this port could be added to a LAG.

Cisco and Dell port configuration differences

On the Dell PowerConnect 5316M, configurations for the LAG take precedence over the configuration of the ports. In Figure 5, example output from the `show running-config` command shows that port g11 is actually set to 100 Mbps (and not 10 Mbps) because the LAG is set to 100 Mbps. If g11 is removed from the LAG, the port configuration will be applied (that is, g11 would be set to 10 Mbps).

On Cisco IOS- and CatOS-based switches, ports must be configured identically to be included in a LAG. Cisco IOS-based switches can use the "desirable" and "passive" mode options for the LAG setting. The PowerConnect 5316M does not support this implementation, and thus administrators should not use these modes when configuring a LAG with a Dell PowerConnect switch. Instead, they should use only the "active" mode (for LACP configuration) or the "on" mode (for static configuration).

Switches can control the distribution only of outgoing traffic on LAG ports. The PowerConnect 5316M uses a static distribution method based on source and destination Media Access Control (MAC) addresses to decide which port or LAG a packet will travel through.[4]

Cisco IOS- and CatOS-based switches provide configuration options for changing the distribution of traffic on LAG ports. The Cisco IOS commands shown in Figure 6 can be used if the Cisco IOS-based switch performs poorly in the LAG. These commands allow administrators to configure the switch to distribute packets to ports in a LAG based on the following settings: destination IP address, destination Ethernet address, a combination of source and destination IP addresses, a combination of source and destination Ethernet addresses, source IP address, or source Ethernet address.

The Cisco CatOS commands shown in Figure 7 can be used if the Cisco CatOS-based switch performs poorly in the LAG. These commands allow administrators to configure the switch to distribute packets to ports in a LAG based on the following settings: destination IP address, destination Ethernet address, a combination of source and destination IP addresses, a combination of source and destination Ethernet addresses, source IP address, or source Ethernet address.

```
3750(config)#port-channel load-balance dst-ip
3750(config)#port-channel load-balance dst-mac
3750(config)#port-channel load-balance src-dst-ip
3750(config)#port-channel load-balance src-dst-mac
3750(config)#port-channel load-balance src-ip
3750(config)#port-channel load-balance src-mac
```

Figure 6. Cisco IOS commands for configuring packet distribution in a LAG

[4] For an in-depth discussion of this algorithm and network design considerations, see "Network Link Aggregation Practices with the Dell PowerEdge 1855 Blade Server" by Bruce Holmes in *Dell Power Solutions,* May 2005; www.dell.com/downloads/global/power/ps2q05-20040286-Holmes-OE.pdf.


```
Cat_6509> (enable) set port channel all distribution ip destination
Cat_6509> (enable) set port channel all distribution mac destination
Cat_6509> (enable) set port channel all distribution ip both
Cat_6509> (enable) set port channel all distribution mac both
Cat_6509> (enable) set port channel all distribution ip source
Cat_6509> (enable) set port channel all distribution mac source
```

Figure 7. Cisco CatOS commands for configuring packet distribution in a LAG

Interoperability between Dell and Cisco switches

The standards-based link aggregation feature of the Dell PowerConnect 5316M Gigabit Ethernet switch is designed to interoperate easily with Cisco IOS- and CatOS-based switches. By understanding the differences in the Dell PowerConnect 5316M and Cisco CLIs and building on the examples presented in this article, system administrators can help integrate the PowerConnect 5316M switch into their Cisco-based networks.

Bruce Holmes is a senior test engineer in the Dell PowerConnect Group. He has worked at Dell for two years and supports PowerConnect switches in all phases of product development and testing. He has a B.S. in Electrical Engineering from The University of Texas at Austin.

FOR MORE INFORMATION

Holmes, Bruce. "Network Link Aggregation Practices with the Dell PowerEdge 1855 Blade Server." *Dell Power Solutions,* May 2005. www.dell.com/downloads/global/power/ps2q05-20040286-Holmes-OE.pdf

Dell PowerConnect 5316M Ethernet Switch Module User's Guide:
support.dell.com/support/edocs/network/PC5316M/en/UG

Understanding USB-based Virtual Media in the Dell PowerEdge 1855 Blade Server

The virtual media feature of the Dell™ PowerEdge™ 1855 blade server employs USB technology when presenting connected devices to the target host. To properly use this feature, administrators should understand how operating systems and BIOSs recognize and communicate with floppy disks, USB flash memory keys, and CD/DVD media in the virtual media environment.

BY JAKE DINER AND SANJEEV SINGH

Related Categories:

Blade servers

Dell OpenManage

Dell PowerEdge servers

Storage management

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

In Dell PowerEdge 1855 blade servers, virtual media capability is supplied through the Avocent® Digital Access KVM (keyboard, video, mouse) module. This module uses USB technology to enable virtual media access. As conceived, USB technology was meant to be seamless across operating systems, architectures, and multiple USB devices. However, proper operation across a specific OS requires custom configuration, because each OS treats USB devices differently. This article explores the behavior of the client and target systems for proper virtual media operation in Dell blade server environments.

Client-side system requirements and operation

The virtual media feature of the Avocent Digital Access KVM module creates a connection between storage media devices on the client system (the management system) and the target platform (the managed system—in this case, the Dell PowerEdge 1855). The procedure requires Java Runtime Environment (JRE), version 1.4.2 or later, installed on the client system; otherwise, the virtual media application will fail to execute.

After updating the Dell Remote Access Controller/Modular Chassis (DRAC/MC) management module or the Avocent Digital Access KVM firmware, administrators must clear the temporary Java files from the Java control panel and clear the temporary Internet files from the Web browser to avoid potential compatibility issues.

Currently, virtual media can be launched from the DRAC/MC Web page in supported Web browsers.[1] Linux® operating systems require administrators to disable writable storage devices such as floppy drives and recordable CD/DVD drives before launching the virtual media application. Failure to do so prevents the virtual media application from obtaining exclusive write access to the devices.

Security enhancements introduced in some versions of Microsoft® Internet Explorer prevent Java applications

[1] For more information about supported Web browsers for the DRAC/MC, see the *Dell Remote Access Controller/Modular Chassis User's Guide* at support.dell.com/support/edocs/software/smdrac3/dracmc/1.1/en/UG/dracugc1.htm#wp34352.

from running. To avoid this conflict, administrators should disable the Microsoft Windows Server™ 2003 Internet Explorer Enhanced Security Configuration component.

Target-side system requirements and operation

The Avocent Digital Access KVM module supports all Dell-branded USB media formatted using the M-Systems Dell USB Memory Key boot utility.[2] With this utility, administrators can make the Dell USB Memory Key bootable from MS-DOS (see Figure 1).

This type of formatting is often called hard-drive or hard-disk formatting because the memory key is seen by the BIOS as another hard drive (see Figure 2). In some instances, when the application is designed to access only the floppy disk on drive A:\, the virtual floppy disk will not be detected. This formatting works identically on Microsoft Windows® and Linux platforms.

Booting the system using the Dell USB Memory Key

After the Dell USB Memory Key has been formatted, it is ready to be used with the Avocent Digital Access KVM module—provided that the BIOS settings are adjusted and the minimum BIOS version of the Dell PowerEdge 1855 target system is A04. First, the hard disk drive sequence must be altered to promote the virtual media to the primary position, as shown in Figure 2. After the Dell USB Memory Key has been promoted in the drive priority, the boot sequence must be adjusted to reflect the change.

Accessing virtual media from the OS

Virtual media devices are detected by the target OS as a USB composite device. The Avocent Digital Access KVM module supports both a CD drive and a floppy drive or USB key simultaneously. Operating systems handle USB devices differently. Windows 2000 and Windows Server 2003 automatically detect and enumerate virtual media devices without administrator intervention. Linux may require administrators to mount virtual media devices as CD drives and hard disk drives.

Before mounting the drive, administrators may need to create a connection to the media at the managed station. The Linux OS automatically creates mount points, which are listed in the /etc/fstab file, for virtual floppy drives and virtual CD drives. For example, if a mount point such as /media/cdrom has already been created by the OS,

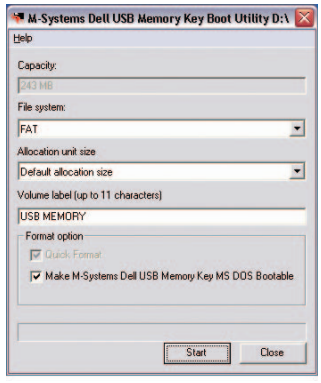


Figure 1. M-Systems Dell USB Memory Key Boot Utility screen

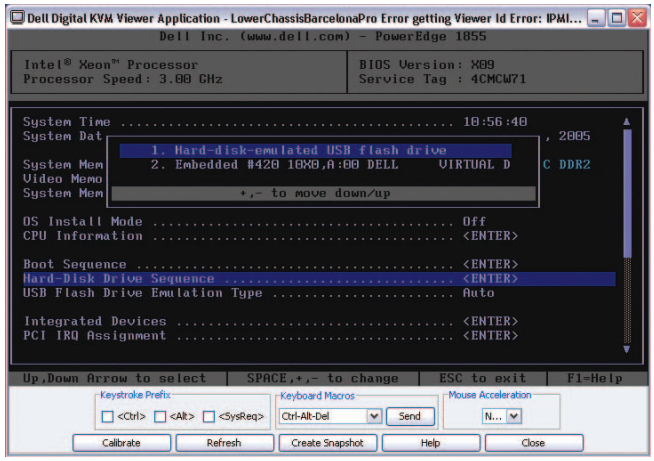


Figure 2. Dell Digital KVM Viewer Application screen: Drive sequence option box

then the administrator can type `mount /media/cdrom` to mount the virtual media and make it accessible.

If the mount point has not been created, then the device name needs to be used to mount the drive. For example, if the device /dev/scd0 has been created by the OS and a directory /mnt1 is present on the hard drive, then administrators can issue the following command to mount the virtual media:

```
mount /dev/scd0 /media/cdrom
```

Meeting virtual media OS requirements

Minimum OS requirements for successful deployment of virtual media and its proper functioning include Microsoft Windows 2000 Server with Service Pack 4 (SP4), Microsoft Windows Server 2003 with SP1 or later, and Red Hat® Enterprise Linux 3 Update 4 or later (Red Hat Enterprise Linux 4 does not require updates).

The power of virtual media access

Administrators can control the OS and BIOS of Dell PowerEdge 1855 blade servers remotely using virtual media access. Understanding this virtual media feature can help administrators efficiently use data center resources and maximize the power and manageability of blade server environments.

Jake Diner is a software engineer in the Dell Enterprise Systems Management Software organization. He has a B.S. in Computer Science from Michigan State University.

Sanjeev Singh is a senior software engineer at Dell. He has a B.E. in Electrical Engineering from Shri Govindram Seksaria Institute of Technology and Science in Indore, India, and an M.S. in Computer Engineering from North Carolina State University.

[2] For instructions on how to make the Dell USB Memory Key bootable, visit support.dell.com/support/edocs/storage/p72969/en/usage.htm#Bootability.

Understanding DRAC/MC Alerts

The Dell™ Remote Access Controller/Modular Chassis (DRAC/MC) provides various alerting mechanisms by which administrators can monitor and manage the components within a Dell Modular Server Enclosure—the chassis that houses the Dell PowerEdge™ 1855 blade server. This article describes the alerts and their related configurations in the DRAC/MC.

BY BABU CHANDRASEKHAR AND STEVEN GRIGSBY

Related Categories:

Blade servers

Dell PowerEdge blade servers

Dell Remote Access Controller (DRAC)

Enterprise management

Systems management

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

The Dell Remote Access Controller/Modular Chassis (DRAC/MC) monitors all modules within the Dell Modular Server Enclosure, including the Dell PowerEdge 1855 server blades, I/O modules, fans, power supply units, and the DRAC/MC modules themselves. When events occur, the DRAC/MC creates entries in one of two logs: the DRAC/MC log or the system event log (SEL). Administrators can access the DRAC/MC log and the SEL using a Web interface, a telnet remote command-line interface (CLI), or a local serial console.

The DRAC/MC log is intended to provide a recent history of DRAC/MC activities on the Dell Modular Server Enclosure. Alerts captured by the DRAC/MC log are triggered either by administrative actions or by the DRAC/MC interacting with other chassis components. The alerts are time-stamped and include the user name and a brief description.

System-critical events occurring on modules in the Dell Modular Server Enclosure are captured in the SEL. Log entries are time-stamped and contain a brief description of the event. SEL entries can be monitored in real time by means of Simple Network Management Protocol (SNMP) traps and Simple Mail Transfer Protocol (SMTP) e-mail messages sent to administrators.

Understanding the DRAC/MC log

The DRAC/MC log is an activity log that contains information such as login activity, session status, firmware update status, and module interaction status. It is a circular log of 512 entries; after it fills up, the oldest alerts are overwritten by new alerts.

DRAC/MC login and session alerts

Login attempts from the Web, telnet, and local serial interfaces, along with the user's IP address, are captured in the DRAC/MC log. Logouts and session timeouts are logged as well. Figure 1 lists the types of DRAC/MC login and session alerts.

ID	Description	Severity
1	Login successful	Informational
2	Login authentication failed	Warning
3	Logout	Informational
4	Session cancelled due to user inactivity	Informational
5	Session cancelled due to DRAC/MC IP address change	Informational
6	Session cancelled due to invalid session ID	Informational

Figure 1. DRAC/MC login and session alerts

ID	Description	Severity
10	User requested server hard reset	Informational
11	User requested server power-cycle	Informational
12	User requested server power-down	Informational
13	User requested server power-up	Informational
14	User requested Advanced Configuration and Power Interface (ACPI)-graceful server shutdown	Informational
15	User requested server nonmaskable interrupt	Informational
44	User requested chassis power-cycle	Informational
45	User requested chassis power-down	Informational
46	User requested chassis power-up	Informational
47	User requested graceful chassis shutdown	Informational
48	User requested switch power-cycle	Informational
49	User requested graceful reboot	Informational

Figure 2. Enclosure and server module power alerts

ID	Description	Severity
16	DRAC/MC IP address changed	Informational
17	DRAC/MC powered up	Informational
18	DRAC/MC Secure Sockets Layer (SSL) certificate expired	Warning
20	DRAC/MC firmware update started	Informational
21	DRAC/MC reset	Informational
22	DRAC/MC assumed master role due to resource failover	Warning
23	DRAC/MC set time	Informational
24	DRAC/MC SEL log cleared	Informational
25	DRAC/MC log cleared	Informational
27	DRAC/MC changed role due to Ethernet disconnection	Warning
28	DRAC/MC firmware update failed due to unavailable image	Warning
29	DRAC/MC firmware update failed due to invalid image	Warning
30	DRAC/MC firmware update failed due to Trivial FTP (TFTP) server time-out	Warning
31	DRAC/MC firmware update successful	Informational

Figure 3. DRAC/MC configuration change alerts

For example, if the "root" user logs in to the Web interface from a system with an IP address of 192.168.1.100, and the session expires from inactivity, the following alerts would be generated in the DRAC/MC log:

```
Root Web Login successful. (192.168.1.100)
Root Web Session cancelled due to
   inactivity. (192.168.1.100)
```

Server module and enclosure power alerts

When an administrator powers up or powers down server modules or the entire Dell Modular Server Enclosure, the activity is logged in the DRAC/MC log. Figure 2 lists the types of server module and enclosure power alerts.

For example, if the "root" user requested server 1 to power down and then requested it to power up, the following alerts would be generated:

```
Root  Requested server-1 powerdown
Root  Requested server-1 powerup
```

DRAC/MC configuration change alerts

When an administrator makes configuration changes to the DRAC/MC or updates the firmware, the activity is recorded in the DRAC/MC log. Figure 3 lists the types of DRAC/MC configuration change alerts.

KVM module alerts

Alerts generated by the KVM (keyboard, video, mouse) module are recorded in the DRAC/MC log. Figure 4 lists the types of KVM-related alerts.

Understanding the system event log

The DRAC/MC constantly monitors the modules within the enclosure for system-critical events. Some events are captured by temperature, voltage, or fan sensors. Other events are captured by the DRAC/MC when configuration changes and power status changes are detected. When an event occurs, the DRAC/MC makes an entry in the SEL, which can be viewed either from the Web interface or the CLI.

Chassis temperature alerts

The DRAC/MC monitors several temperature sensors located in the chassis. These sensors monitor the ambient air temperature inside the chassis.

ID	Description	Severity
50	KVM firmware update started	Informational
51	KVM firmware update was successful	Informational
52	KVM firmware update failed due to unreachable TFTP server	Warning
53	KVM firmware update failed due to unavailable image	Warning
54	KVM firmware update failed due to TFTP server time-out	Warning
55	KVM firmware update failed due to invalid image	Warning
56	KVM firmware update failed due to authentication error	Warning
57	KVM firmware update failed due to unknown error	Warning
58	KVM firmware update failed due to open virtual media session	Warning
59	KVM firmware file transfer complete	Informational
60	KVM configuration reset to default	Informational

Figure 4. KVM module alerts

ID	Description	Severity
9	Sensor returned to normal range	Informational
10	Sensor failure detected	Warning
11	Sensor returned to below critical threshold	Warning
12	Sensor detected a nonrecoverable event	Critical

Figure 5. Chassis environmental sensor alerts

ID	Description	Severity
3	Module sensor returned to normal range	Informational
4	Module sensor unavailable or failure detected	Warning
5	Module sensor returned to below critical threshold	Warning
6	Module sensor detected a nonrecoverable event	Critical

Figure 6. Server module environmental sensor alerts

A warning alert is generated when the temperature exceeds the normal operating range. If the temperature continues to climb and reaches the critical threshold, a nonrecoverable event alert is generated. If the temperature falls back below the critical threshold but remains higher than the warning threshold, an alert is generated indicating the temperature has returned to the warning temperature range. When the temperature falls back below the warning threshold, an alert is generated indicating the temperature has returned to the normal range.

The affected module and sensor type are logged with each alert. For example, if the left housing temperature probe indicates a warning temperature and then returns to normal, the following alerts would be generated:

```
Housing-Left temperature sensor failure was
   detected.
Housing-Left temperature sensor returned to
   normal.
```

Figure 5 shows the types of alerts that are generated by environmental sensors (temperature, voltage, and fan).

Chassis voltage alerts

Some I/O modules contain voltage sensors that are monitored by the DRAC/MC. The alerting mechanism is the same as for temperature probes, except the sensor type is set to "voltage." There are warning and critical thresholds and corresponding alerts, as shown in Figure 5. For example, if the voltage sensor for I/O module 3 exceeds the critical threshold, the following alerts are generated:

```
Switch-3 voltage sensor failure was detected.
Switch-3 voltage sensor detected a nonrecoverable
   event.
```

When the voltage returns to normal, the following alert is generated:

```
Switch-3 voltage sensor returned to normal.
```

Chassis fan alerts

The fan modules and power supply modules in the rear of the chassis are instrumented with sensors to indicate when a fan is not operating at the correct speed. Each fan module contains two fans, and each power supply module contains three fans. The types of alerts listed in Figure 5 are also used for fan alerts; however, the affected module and fan are included in the alert message. For example, if fan 1 in power supply module 4 stops operating, the following alert is generated:

```
PS-4 Fan-1 RPM fan sensor detected a nonrecoverable
   event.
```

When the fan RPM returns to normal, the following alert is generated:

```
PS-4 Fan-1 RPM fan sensor returned to normal.
```

Miscellaneous alerts

The DRAC/MC reports other server module conditions. These include server module environmental sensors, chassis configuration changes, and invalid chassis configurations.

Server module environmental sensors. In addition to the chassis-wide environmental sensors, server modules have their own temperature and voltage sensors. However, the server module's baseboard management controller (BMC) is responsible for monitoring these sensors and recording alerts in the BMC SEL. Each server module contains a BMC that logs alerts specific to that server in its SEL.

When the server module's BMC detects a voltage or temperature out-of-range condition, the BMC logs the event to the server's SEL and notifies the DRAC/MC that a server module sensor alert was detected. The DRAC/MC in turn generates an alert and logs it to the DRAC/MC SEL. The types of alerts triggered by the server's environmental sensors and generated by the DRAC/MC are listed in Figure 6.

For example, if the server in slot 5 of the chassis detects a failure caused by a critical temperature condition, the server's BMC would log the event to its SEL and notify the DRAC/MC. The DRAC/MC would generate the following alert:


```
Server-5 module sensor detected a nonrecoverable
    event.
```

The alert generated by the DRAC/MC indicates that a critical failure occurred on server 5 but does not specify the details of the failure. The administrator must then go to the server module's SEL to obtain details about the failure.

Chassis configuration change alerts. The DRAC/MC detects when modules are added or removed from the Dell Modular Server Enclosure and generates corresponding event alerts. Figure 7 describes these types of alerts and the corresponding severities. The alert severity for adding modules is *informational;* however, the alert severity for removing modules may reach *warning* or *critical* status, depending on which type of module is removed.

When modules are powered up and down, the DRAC/MC generates an informational alert. The DRAC/MC also detects when power is lost and restored to power supply modules. Figure 8 describes the type of power status change alerts generated by the DRAC/MC.

Chassis invalid configuration alerts. Because of the modular design of the Dell Modular Server Enclosure, administrators could configure a chassis with I/O modules that are not compatible with existing I/O devices in the server modules. The DRAC/MC monitors the chassis configuration and does not allow any module to power up if it is incompatible with the rest of components in the chassis. Figure 9 lists the types of alerts relating to an invalid chassis configuration.

ID	Module type	Description	Severity
1	All	Module sensor presence detected	Informational
2	DRAC/MC	Module sensor removed	Informational
2	Server module	Module sensor removed	Warning
2	I/O module	Module sensor removed	Warning
2	Fan module	Module sensor removed	Critical
2	Power supply module	Module sensor removed	Critical

Figure 7. Chassis configuration change alerts

ID	Description	Severity
7	Module sensor detected power-up	Informational
8	Module sensor detected power-down	Informational
13	Power supply sensor power lost	Warning
14	Power supply sensor power restored	Informational

Figure 8. Power status change alerts

Configuring the DRAC/MC to send SEL alerts

Because of the possibly critical nature of SEL alerts, the DRAC/MC provides two mechanisms for sending alerts to remote destinations. The DRAC/MC can be configured to send e-mail alerts to as many as 16 different recipients. It can also be configured to send SNMP traps to as many as 16 network management stations such as Dell OpenManage IT Assistant (ITA). E-mail alerts and SNMP traps cannot be sent for DRAC/MC alerts.

E-mail alerts

Administrators must enable e-mail alerts on the DRAC/MC console so that the system can generate them. To do so, administrators should navigate to the Configuration > Network page of the DRAC/MC Web interface (see Figure 10) and click the "Enable E-mail Alerts" checkbox in the "E-mail Alert Settings" section of the page. The IP address of the SMTP (e-mail) server must also be entered.

ID	Description	Severity
18	New I/O module incompatible with current I/O type	Warning
19	Slave I/O module installed before master installed	Warning
20	Non–Gigabit Ethernet fabric I/O module installed in I/O module group 1	Warning
21	Fabric type of group 2 I/O module incompatible with daughtercard in current blade	Warning
22	Daughtercard in new server blade incompatible with I/O module	Warning
23	Daughtercard in new server blade incompatible with daughtercard(s) in existing blades	Warning
24	Chassis misconfiguration event log	Warning

Figure 9. Chassis invalid configuration alerts

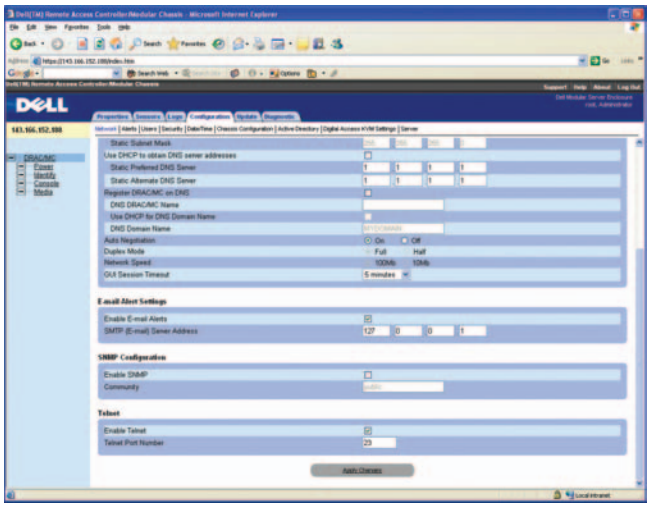


Figure 10. Network setup and e-mail alert settings screen in the DRAC/MC Web interface

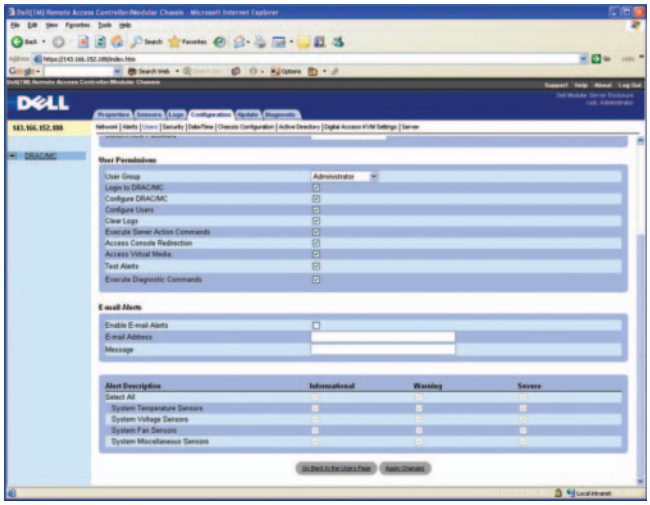


Figure 11. Alert permissions and preferences screen in the DRAC/MC Web interface

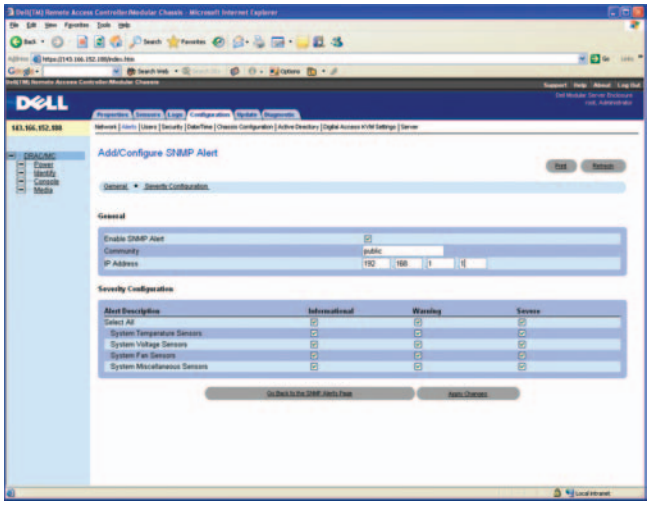


Figure 12. SNMP alert configuration screen in the DRAC/MC Web interface

The DRAC/MC permits administrators to control alert functions for as many as 16 local recipients. Recipients may specify their e-mail addresses and whether they wish to receive e-mail alerts. Recipients can also specify which SEL alerts they wish to receive.

To receive e-mail alerts, a recipient must navigate to the Configuration > Users page of the DRAC/MC Web interface and click on an available or existing username. The Configuration > Users page should then appear (see Figure 11). The "E-mail Alerts" section allows the alert recipient to enable and disable e-mail alerts, configure alert filters, and specify e-mail addresses. The "Alert Description" section lets alert recipients specify which SEL alerts and severities should generate e-mail alerts to them.

SNMP traps

Besides sending e-mail alerts to specific recipients, the DRAC/MC can be configured to send SNMP traps to up to 16 destinations. These destinations are typically management systems such as ITA, which can decode the traps and display the appropriate alert message. The DRAC/MC must be configured to send SNMP traps. Administrators can do this by navigating to the Configuration > Network page of the Web interface (see Figure 10) and scrolling to the "SNMP Configuration" section. They should click the "Enable SNMP" checkbox and designate the SNMP recipient community for the DRAC/MC.

After SNMP is configured for the DRAC/MC, SNMP trap destinations can be configured. Administrators can do so by navigating to the Configuration > Alerts page of the Web interface and clicking on an existing or available SNMP alert destination. The Add/Configure SNMP Alert page should then appear (see Figure 12). Administrators can enable SNMP alerts for this destination and set the community string and IP address of the management system to receive alerts. The alert filter allows administrators to specify which SEL alerts and severities qualify to trigger SNMP traps to be sent to this destination.

Monitoring modular server environments through comprehensive alerting

The DRAC/MC is designed to detect events and generate alerts in the DRAC/MC log and the SEL. User-triggered events generate alerts captured in the DRAC/MC log. Critical system alerts are captured in the SEL, and the DRAC/MC can be configured to notify multiple recipients through e-mail alerts or SNMP traps. Such alerting capabilities can help administrators monitor the components within a Dell Modular Server Enclosure and enhance management of the Dell PowerEdge 1855 blade server system.

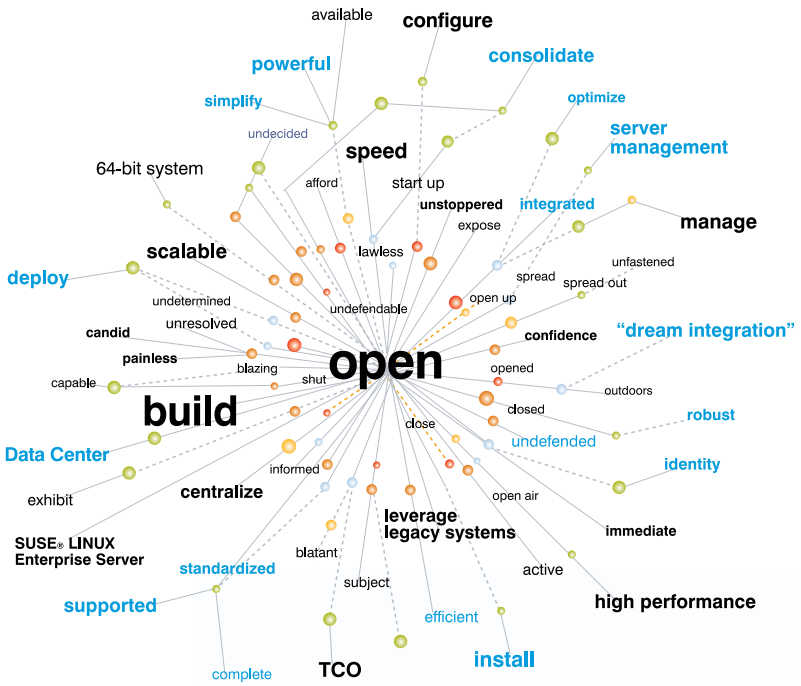
Babu Chandrasekhar is a lead software engineer in the Dell Enterprise Server Group. Before joining Dell, he worked as a software engineer for Digital Equipment Corporation, Intel Corporation, and Bhabha Atomic Research Centre. He has a B.S. in Computer Science and Engineering from the University of Kerala in India.

Steven Grigsby is a product test engineer in the Dell Enterprise Server Group. He has a B.S. in Computer Science from the University of Oklahoma, Norman.

FOR MORE INFORMATION

Dell PowerEdge 1855 blade server:
www1.us.dell.com/content/products/productdetails.aspx/pedge_1855?c=us&cs=04&l=en&s=bsd

Dell Remote Access Controller/Modular Chassis User's Guide:
support.dell.com/support/edocs/software/smdrac3/dracmc/index.htm



Define Your Open Enterprise.™

What does Open mean to you? Community? Security? Risk? Reward? Can it leverage legacy systems? Consolidate and simplify? Do you believe in its power and potential?

Introducing Software for the Open Enterprise™ from Novell®—the only software that makes Open work for you. From desktop and data center to identity management, resource management and collaboration, our flexible combination of open source and commercial software delivers more than you ever imagined. The power to automate IT asset management. Freedom from single vendor lock-in. Security that keeps the right information safe and the right people informed. And the ability to connect people to performance and business to possibilities. So you can build an open enterprise that makes sense for you—and your future. This is Software for the Open Enterprise from Novell. The Open you've wanted all along.

Novell®

This is your open enterprise.™

www.novell.com/defineyouropen