DELL™ POWER SOLUTIONS

NOVEMBER 2003 · $12.95

THE MAGAZINE FOR DIRECT ENTERPRISE SOLUTIONS

The Scalable Enterprise: Building Blocks for Cluster Computing

Inside This Issue:

Evaluating performance characteristics of Intel-based HPC servers

Scaling Microsoft SQL Server database applications

Building a highly available data environment for Oracle9*i* RAC

Scaling Out the Dell Services Infrastructure

Dell entered the service market very much the way most hardware vendors do. We wanted to ensure a positive customer experience both before and after sales. Over the last few years, however, Dell has explored ways of making a difference in the IT services sector. The key points of our differentiated service strategy directly address the feedback we consistently hear from IT executives:

"I feel like I am held hostage by inflexible contracts that no longer meet my business or operational needs."

"I never know what I am paying for. The scope of my third-party service providers—when it matters most to me—is always ambiguous."

"My users are dissatisfied."

"I like the idea of outsourcing. I need to do something to standardize and reduce costs. But I don't want to lose control of my infrastructure."

"Service pricing is not keeping pace with hardware pricing."

In response, the challenge for Dell became one of building flexibility, process rigor, defined statements of work, and performance measurement into a service delivery strategy that also provides outstanding value. To achieve this we drew upon the Dell™ direct model.

The economics of efficient and high-velocity supply chains apply to service delivery as well as systems production and delivery. In a traditional service support organization the "warehouse" is full of inventory: a standing army of engineers awaits work orders. The management challenge is to efficiently allocate work to free engineers. In this traditional model, because of geographic and base coverage requirements, Dell estimates that the level of inevitable inefficiency can be as high as 20 percent.

By using partners for most of our service work, we dramatically reduce standing-army costs—and Dell can pass these savings on to our customers. Our partner network also allows us to allocate work among our certified partners in a way that best leverages their excess capacity. Moreover, as in systems development, we can immediately leverage industry best practices. Because we are not burdened with a standing army, Dell can allow much more flexibility—both in our statements of work and in our contractual time frames.

Comprehensive but flexible service arrangements that address operational and business requirements have repeatedly proven attractive to our customers. By compiling standard service elements in custom ways, Dell can optimize solution fit and minimize costs. Our menu-driven approach to services focuses on the customer, not on our need to utilize a base of field engineers or consultants. As a result the risk—read: cost—of statement of work changes can be dramatically reduced for both our customers and ourselves.

However, we must emphasize that although Dell partners provide much of the professional and break-and-fix services, Dell owns the service design and the service level commitment. This arrangement provides a strong single point of control and contact and drives the operational performance of the contract.

In EMEA alone, 1,500 Dell employees form an envelope of support around our customers. As in systems production and delivery, we can provide the Dell customer with an integrated service solution—built to order—that provides excellent value.

Josh Claman
Vice President, Services
Dell Europe, Middle East, and Africa (EMEA)

The Economics of Scaling: Cluster-based Computing

In today's business-driven IT environment, administrators must scale the enterprise infrastructure quickly, flexing to meet dynamic business challenges in real time. To survive, IT organizations must reshape their budgets so they can increase return on investment (ROI) and contribute to the bottom line. Scalable architectures built using cost-effective Intel® processor–based systems can perform high-availability, fault-tolerant processing tasks that once required costly supercomputers.

The data center is being transformed rapidly by open source and industry-standard computing components. Recent Dell™ studies indicate that performance goals can be realized more cost-effectively by deploying smaller servers in distributed and clustered architectures rather than by investing significant amounts of capital in large, proprietary systems that may never fully be utilized.[1] In this issue of *Dell Power Solutions*, we focus on how IT organizations can scale out the enterprise using clusters, including high-performance computing (HPC) clusters for research as well as application clusters for increased availability or improved workload management.

Enterprises, universities, and research institutions around the world crank out terabytes of data each year on every topic imaginable—from the origins of the universe to human behavior and even three-legged frogs in Minnesota. Advances in open source and low-cost industry-standard systems have altered the landscape of compute-intensive research. In this issue, we delve into the architecture that supports such efforts, addressing the following topics in particular:

- Performance characteristics of the latest Intel-based servers
- Impact of Intel Hyper-Threading Technology on high-performance computing
- Current and future developments of the Open Source Cluster Application Resources (OSCAR) toolkit
- Best practices and network considerations for building an HPC cluster environment

On the data center side, we explore building blocks and tools that can enhance rapid deployment in Microsoft® and Linux® operating system environments. In addition, we explain methods for scaling Microsoft SQL Server and Oracle 9*i*™ Real Application Clusters (RAC) to support the exponential growth in data access, along with techniques for improving availability and disaster recovery.

Looking ahead, industry efforts to develop the data center will focus on convergence, simplification, and virtualization. In this issue, we examine ways to optimize the virtual data center, including approaches for implementing load-balancing and resource-balancing features so they can be transparent to client applications. Because workload redistribution is the key to load and resource balancing, high-performance interconnects will be critical to the success of the virtual data center.

We hope this issue of *Dell Power Solutions* provides fresh insights into the economics of scaling, as well as new strategies for building out enterprise systems and applications with cluster computing. After all, we are here to capture solutions that work and improve the quality of IT. Your feedback has been valuable to us. Keep sending your comments and let us know about your successful Dell deployments.

Eddie Ho

Eddie Ho
Editor-in-Chief
www.dell.com/powersolutions

[1] For an example, see "Comparing Oracle9*i* Database Performance on Dell and Sun Servers" by Dave Jaffe, Ph.D., and Todd Muirhead in *Dell Power Solutions,* August 2003.

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

EDITORIAL
EDITOR-IN-CHIEF | Eddie Ho
MANAGING EDITOR | Debra McDonald
CONTRIBUTING EDITORS | Stephanie Dobler Cerra, Liza Graffeo, Cathy Luo, Vicki Van Ausdall

ART
ART DIRECTOR | Mark Mastroianni
DESIGNERS | Phu Tran, Cynthia Webb
COVER DESIGN | Phu Tran

ONLINE
WEB PRODUCTION MANAGER | Brad Klenzendorf

SUBSCRIPTION SERVICES
SUBSCRIPTION SERVICES | April Ho

Subscriptions are free to qualified readers who complete the subscription card found in each issue. To subscribe or change your address, complete and return the business reply card in this issue to Dell Inc., c/o April Ho, One Dell Way, RR #5, Box 8503, Round Rock, TX, 78682, USA, or visit us at **www.dell.com/powersolutions_subscribe**.

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 40,000 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., One Dell Way, Round Rock, TX 78682. This publication is also available online at www.dell.com/powersolutions. No part of this publication may be reprinted or otherwise reproduced without permission from the Editor-in-Chief. Dell does not provide any warranty as to the accuracy of any information provided through *Dell Power Solutions*. Opinions expressed in this magazine may not be those of Dell. The information in this publication is subject to change without notice. Any reliance by the end user on the information contained herein is at the end user's risk. Dell will not be liable for information in any way, including but not limited to its accuracy or completeness. Dell does not accept responsibility for the advertising content of the magazine or for any claims, actions, or losses arising therefrom. Goods, services, and/or advertisements within this publication other than those of Dell are not endorsed by or in any way connected with Dell Inc.

November 2003

DELL POWER SOLUTIONS

THE MAGAZINE FOR DIRECT ENTERPRISE SOLUTIONS

NOVEMBER 2003

EXECUTIVE NOTES

EDITOR'S COMMENTS

EXECUTIVE INSIGHTS

HIGH-PERFORMANCE COMPUTING

DATA CENTER ENVIRONMENT

ENTERPRISE CLUSTER ENVIRONMENT

STORAGE ENVIRONMENT

HPC Clusters: Building Blocks for the Enterprise

One year ago, *Dell Power Solutions* asked Reza Rooholamini, Ph.D., director of enterprise solutions at Dell Inc., about the potential for inexpensive, standards-based components to revolutionize high-performance computing (HPC). This year, the answer has materialized in powerful Linux®- and Microsoft® Windows®–based 64-bit HPC clusters from Dell.

Spurred by the dramatic price reductions, performance improvements, and scalability of high-performance computing (HPC) clusters, IT organizations are rapidly deploying parallel systems built from standards-based computing components. In an industry ranking of the 500 most powerful computing systems in the world, the number of clusters comprising Intel® architecture–based servers more than doubled from 56 in November 2002 to 119 in June 2003. Three of the top 10 computing systems in the world now comprise clusters of Intel processor–based servers.[1]

In addition, the affordability and manageability of relatively small HPC clusters have contributed to the large-scale adoption of parallel computing in many diverse customer applications. In this interview, Reza Rooholamini explains the Dell strategy for deploying HPC clusters that use standard processing, storage, and interconnect components as building blocks for the enterprise IT infrastructure.

What is driving HPC cluster growth in the enterprise?

It has become very practical to build parallel computing systems using low-cost, high-density nodes such as Intel processor–based Dell™ PowerEdge™ servers. The number of Intel-based HPC clusters has increased dramatically in the past year largely because their total cost of ownership (TCO) can be much lower than that of proprietary supercomputing systems. Now, the Dell PowerEdge 3250 server, based on the Intel Itanium® 2 processor, can enable enterprises to migrate existing 64-bit applications from proprietary 64-bit processors to the cost-effective, standards-based IA-64 architecture. Moreover, the Dell approach of testing and partnering with best-of-class hardware and software companies can help simplify customer HPC cluster deployments and manage their growth—helping remove barriers to entry.

What types of applications can Dell HPC clusters perform?

The National Center for Supercomputing Applications (NCSA) has selected Dell to provide more than 1,450 PowerEdge servers, which will be networked to form an HPC cluster that offers a theoretical peak performance of 17.7 trillion floating-point operations per second. The NCSA cluster will help researchers explore major scientific and engineering questions, such as studying the evolution, size, and structure of the universe; investigating theories on the life cycle of stars like the Sun; modeling severe storms; studying the human genome and biological processes; and advancing the drug design process.

Meanwhile, Compagnie Générale de Géophysique, a global services oil company, has linked more than 3,000 Dell PowerEdge servers into HPC clusters that analyze seismic data, helping to identify and model oil and gas reservoirs around the world. Another example is the Stanford University Bio-X program, which has

[1] TOP500 Supercomputer Sites, http://www.top500.org: November 2002 ranking, http://www.top500.org/lists/2002/11; June 2003 ranking, http://www.top500.org/lists/2003/06.

networked more than 300 PowerEdge servers in a supercomputing cluster that enables researchers to simulate the effects of medicines for Alzheimer's disease and cancer, and to help improve surgical and physical rehabilitation techniques.

How is Dell helping enterprise customers to adopt HPC clusters?

Dell offers a single point of contact for the purchase, deployment, and management of HPC clusters. Bundled configurations offer the low cost, high availability, scalability, and disaster recovery features that customers associate with HPC clusters, along with affordable Dell service and support. HPC cluster bundles provide a complete set of components in a single package, which simplifies the design, ordering, and deployment process for parallel computing. Moreover, Dell validates each HPC cluster deployment to help ensure achievable performance.

Four high-density servers comprise Dell turnkey HPC cluster configurations: the PowerEdge 1750, PowerEdge 2650, PowerEdge 3250, and PowerEdge 1655MC servers. Each server supports a variety of network interconnects—Fast Ethernet, Gigabit Ethernet,[2] and Myricom® Myrinet®—in various scalable configurations designed to address a wide range of enterprise applications. These PowerEdge servers support configurations of 8, 16, 32, 64, and 128 compute nodes, except for the PowerEdge 1655MC HPC cluster, which supports configurations of 6, 18, 36, 66, and 132 compute nodes.

> The demand from our enterprise customers indicates that HPC clusters are fast becoming a computing standard in the corporate data center.

The software packages that Dell HPC cluster partners offer for these bundles are supported on Red Hat® Enterprise Linux® AS and WS 2.1 operating systems. Dell HPC bundles not only facilitate quick deployment and installation of the cluster, but also provide the essential programming environment, which consists of compilers, math libraries, message-passing middleware, and job schedulers.

How does Dell package Windows-based clusters?

Dell currently offers bundled HPC clusters with the Microsoft® Windows® Server 2003 operating system. Packaged software kits—such as the Microsoft Computational Clustering Technical Preview (CCTP) kit, 2003 Edition—help enable organizations to design, deploy, configure, and manage their Windows-based HPC clusters.

Dell also can provide custom configurations for Windows-based HPC and Microsoft .NET applications through its partnerships with Microsoft, Intel, and the Cornell Theory Center (CTC). The CTC is currently developing an infrastructure to support database-centric parallel computing, distributed Web services, and real-time data generation to control the execution of long-running, resource-intensive programs on Windows Server 2003.

What does the Dell 64-bit HPC cluster server offer?

The PowerEdge 3250 is a rack-optimized (2U) server that supports up to two Intel Itanium 2 processors, up to 16 GB of double data rate (DDR) memory, three Peripheral Component Interconnect Extended (PCI-X) slots, and two integrated Gigabit Ethernet network interface cards (NICs). Given the dense form factor of the PowerEdge 3250 server and the performance capabilities of the Itanium 2 processor, this system is an excellent hardware building block for the Dell 64-bit HPC cluster configurations.

What opportunities do 64-bit nodes create for HPC clusters?

The Dell PowerEdge 3250 is well suited as a compute node for Linux- or Windows-based HPC clusters because IA-64 architecture enables the PowerEdge 3250 to accommodate applications that have large memory footprints. To overcome the limitations of 32-bit architecture, the PowerEdge 3250 was designed to support higher memory bandwidth and higher floating-point memory, allowing the server to address more than 4 GB of memory. This 64-bit Dell server enables organizations that are already utilizing 64-bit applications on proprietary 64-bit processors to migrate to the cost-effective, standards-based IA-64 architecture.

When we spoke one year ago, we asked you to summarize trends in HPC clusters. Would you respond differently today?

One year ago, we said that clustering would revolutionize traditional high-performance computing. The unprecedented increase in standards-based clusters among the 500 most powerful computing systems in the world shows clearly that our prediction has come true. The demand from our enterprise customers indicates that HPC clusters are fast becoming a computing standard in the corporate data center. In response, Dell will continue to offer the bundled packages, low prices, and excellent service required to help enterprises scale out their IT infrastructure and meet ever-changing business needs.

[2] This term indicates compliance with IEEE® standard 802.3ab for Gigabit Ethernet, and does not connote actual operating speed of 1 Gbps. For high-speed transmission, connection to a Gigabit Ethernet server and network infrastructure is required.

Performance Characteristics of Intel Architecture–based Servers

High-performance computing (HPC) has increasingly adopted the use of clustered Intel® architecture–based servers. This article compares the performance characteristics of three Dell™ PowerEdge™ server models with Intel processors—Intel Pentium® III, Intel Xeon™, and Intel Itanium® 2—running benchmarks commonly used in HPC environments.

BY RIZWAN ALI; JENWEI HSIEH, PH.D.; AND ONUR CELEBIOGLU

Computing clusters built from standard components using Intel® processors are becoming the fastest growing choice for high-performance computing (HPC). Twice yearly, the 500 most powerful computing systems in the world are ranked on the TOP500 Supercomputer Sites Web page. In November 2002, the ranking listed 56 entries using Intel processors; by June 2003, that number reached 119. Today, three of the top 10 computing systems in the world are clusters of Intel architecture–based servers.[1]

Since 1995, Intel has continuously introduced more powerful processors with increased speed, faster frontside bus, and larger cache. To test the effectiveness of Intel processor–based servers in HPC environments, a Dell™ team compared performance characteristics of three Intel architecture–based Dell PowerEdge™ servers using benchmarks commonly employed in HPC. Studies have shown that small-scale symmetric multiprocessing (SMP) systems make excellent platforms for building HPC clusters.[2] Thus, two-processor systems were used for the Dell tests discussed in this article.

Establishing the test environment

The test environment consisted of three Dell PowerEdge server models, each with a different Intel processor: Intel Pentium® III, Intel Xeon™, and Intel Itanium® 2. Figure 1 shows the specifications for each server.

The software environment included the Red Hat® Linux® operating system, the MPI/Pro® Message Passing Interface (MPI) library from MPI Software Technology for running multiple processes, and various Intel compilers.

Benchmarks used to compare performance among the three platforms included High-Performance Linpack (HPL), the NAS (NASA Advanced Supercomputing) Parallel Benchmarks (NPB), and a kernel benchmark called Additive Schwarz Preconditioned Conjugate Gradient (ASPCG).

Using High-Performance Linpack to compare HPC performance

Linpack is a popular benchmark for HPC environments. Its HPL implementation is used to rank supercomputers on the TOP500 Supercomputer Sites list. Linpack employs

[1] TOP500 Supercomputer Sites, http://www.top500.org: November 2002 ranking, http://www.top500.org/lists/2002/11; June 2003 ranking, http://www.top500.org/lists/2003/06.

[2] Hsieh, J., T. Leng, V. Mashayekhi, and R. Rooholamini, "Impact of Level 2 Cache and Memory Subsystem on the Scalability of Clusters of Small-Scale SMP Servers," paper presented at IEEE® International Conference on Cluster Computing (Cluster 2000), Chemnitz, Germany, November 2000.

	PowerEdge 1650	PowerEdge 1750	PowerEdge 3250
CPU	Two Intel Pentium III processors at 1.4 GHz	Two Intel Xeon processors at 3.06 GHz	Two Intel Itanium 2 processors at 1.5 GHz
Frontside bus	64 bits at 133 MHz	64 bits at 533 MHz	128 bits at 400 MHz
Cache size	512 KB L2 cache	512 KB L2 cache, 1 MB L3 cache	256 KB L2 cache, 6 MB L3 cache
Memory bandwidth	1.06 GB/sec	4.3 GB/sec	6.4 GB/sec
Maximum memory	4 GB SDRAM	8 GB double data rate (DDR) PC2100	16 GB DDR PC1600
Rack units (U)	1U	1U	2U
Peripheral Component Interconnect (PCI) slots	Two 64-bit/66 MHz PCI slots	Two 64-bit/133 MHz PCI Extended (PCI-X) slots	Two 64-bit/100 MHz and one 64-bit/133 MHz PCI-X slots

Figure 1. Dell PowerEdge specifications for the test environment

several linear algebra routines—such as Gaussian elimination—to measure the time required to solve a dense system of linear equations in double-precision (64-bit) arithmetic. The measurement obtained from Linpack is the number of floating-point operations per second (FLOPS). Running HPL requires the MPI library and either the Basic Linear Algebra Subprograms (BLAS) or the Vector Signal Image Processing Library (VSIPL).

Numerous parameters in HPL can be tuned to increase the overall system performance. The program enables users to run tests for different sets of problem sizes—each problem size represents a certain percentage of aggregate memory. By using the FLOPS values obtained for different problem sizes, testers can plot a graph similar to the one shown in Figure 2.

The Dell team ran HPL on three dual-processor PowerEdge servers. Figure 2 shows the relative performance of the Intel Xeon and Itanium 2 systems compared to the Pentium III–based PowerEdge 1650. The y-axis shows the speedup relative to the PowerEdge 1650; the x-axis shows the problem size. The Itanium 2–based PowerEdge 3250 significantly outperformed the PowerEdge 1650; it also outperformed the PowerEdge 1750, even though it was running at half the clock speed of the PowerEdge 1750.

The Itanium 2 processor can perform four double-precision floating-point operations per clock cycle, the Intel Xeon processor can perform two, and the Pentium III processor can perform just one. Therefore, the theoretical peak of an Itanium 2 processor running at 1.5 GHz is 6 gigaflops (GFLOPS), nearly the same as that of the Intel Xeon processor at 3.06 GHz, which has a theoretical maximum of 6.12 GFLOPS.[3]

Calculating efficiency

To determine processor efficiency, the Dell team calculated the percentage of sustained gigaflops attained by running HPL on a single processor and comparing that number to the theoretical peak performance that can be achieved by that processor. For example, if the sustained peak performance for a single Intel Xeon processor running at 3.06 GHz is 4.5 GFLOPS, the formula for calculating efficiency would be as follows:

$$\frac{\text{sustained peak performance}}{\text{theoretical peak performance}} \times 100$$

In this case, the result would be 73.5 percent (4.5 / 6.12 × 100).

Figure 3 shows the Linpack efficiencies of single and dual processors for each of the three Intel architecture–based platforms—Pentium III, Intel Xeon, and Itanium 2. The Itanium 2 processor–based PowerEdge 3250 offers the best efficiency compared to the other servers tested, for both single- and dual-processor configurations. It is followed by the Intel Xeon processor–based PowerEdge 1750, which outperformed the PowerEdge 1650. Test results indicate that Intel processors show not only increasing performance from generation to generation because of many architectural enhancements, but also increasing efficiency.

Using NPB to test parallel computing performance

The NAS Parallel Benchmarks, developed by NASA Ames Research Center, measure and compare the performance of parallel computers. The standard MPI programming model of NPB version 2.4 enables comparative analysis of a standards-based cluster with various interconnects.

The NPB suite is a set of eight programs derived from computational fluid dynamics (CFD) code. Each of these eight programs—five kernels and three simulated CFD applications—represents a particular function of highly parallel computation for aerophysics applications. The five kernels—EP, FT, MG, CG, and IS—mimic the computational core of different numerical methods used by CFD

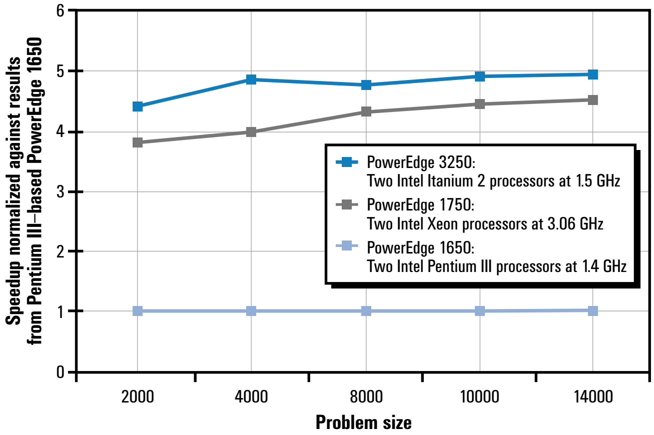


Figure 2. Speedup comparisons for Intel architecture–based servers running the HPL benchmark

[3] One gigaflop equals 1 billion floating-point operations per second (FLOPS).

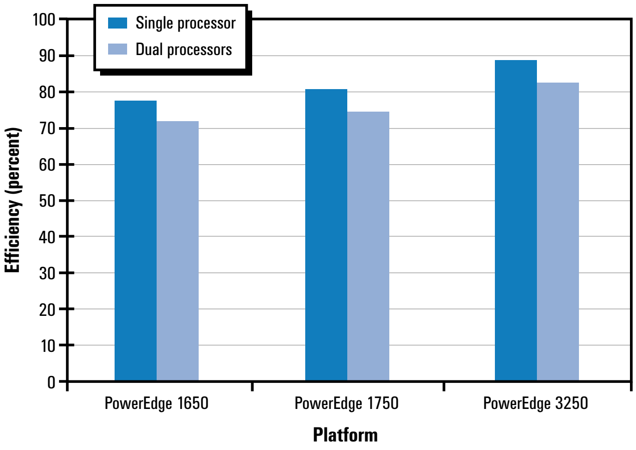


Figure 3. HPL efficiency for single and dual processors on the three Intel architecture–based servers

applications. The NPB benchmark suite outputs its results in millions of operations per second (MOPS).

Each benchmark has six sizes called classes: A, B, C, D, W (workstation), and S (sample). A, B, C, and D represent four different problem sizes; A is the smallest and D is the largest. Here, the problem size is the size of the data set to be benchmarked; larger data sets take longer to process and use more memory. For an exact performance comparison, testers should use problem size B or larger. To compare the three platforms, the Dell team used two of the eight benchmark programs provided with NPB: IS and MG.

Integer Sort (IS): Testing speed and performance

The IS benchmark tests both integer computation speed and data communication performance. It tests a parallel sorting operation important in particle method codes. For example, this type of application is similar to particle-in-cell applications of physics, wherein particles are assigned to cells but may drift out. The sorting operation can reassign particles to the appropriate cells. This type of problem is unique because it requires no floating-point arithmetic, but it does require significant data communication. Integer Sort is sensitive to communication latency; therefore, faster interconnects or faster share memory with low latency and high bandwidth often perform better in communication-intensive applications.

Figure 4 shows the IS benchmark results for the three platforms running on both single- and dual-processor systems. The results are normalized against the result from a single Pentium III processor. The most significant observation from the graph is the one- to two-processor in-the-box scaling (that is, scalability achieved within a server without clustering).

The PowerEdge 3250 showed near-linear scaling from one to two processors. These results for the IS benchmark were achieved by the greater memory bandwidth and the larger L3 cache provided by the PowerEdge 3250. In-the-box scaling for Pentium III and Intel Xeon was poor compared to the Itanium 2 in the PowerEdge 3250. For the single processor run, the speedup of the Intel Xeon processor was nearly twice as fast compared to the Pentium III processor. This result was in line with the clock frequencies for the Pentium III and Intel Xeon processors.

Multigrid (MG): Testing highly structured communications

The MG benchmark tests both short- and long-distance communication. MG is a simplified multigrid kernel that solves a 3-D Poisson partial differential equation (PDE). Because MG has constant rather than variable coefficients, it simplifies the problem. MG runs a mixture of different operations—floating-point as well as integer.

Figure 5 shows the MG benchmark results for the three platforms running on both single- and dual-processor systems. Similar to those of the IS benchmark, the MG results are normalized against the result from a single Pentium III processor. The 64-bit PowerEdge 3250 outperformed both 32-bit platforms by a significant margin for single- and dual-processor runs. The MG benchmark takes advantage of the Streaming SIMD (Single Instruction Multiple Data) Extensions 2 (SSE2)

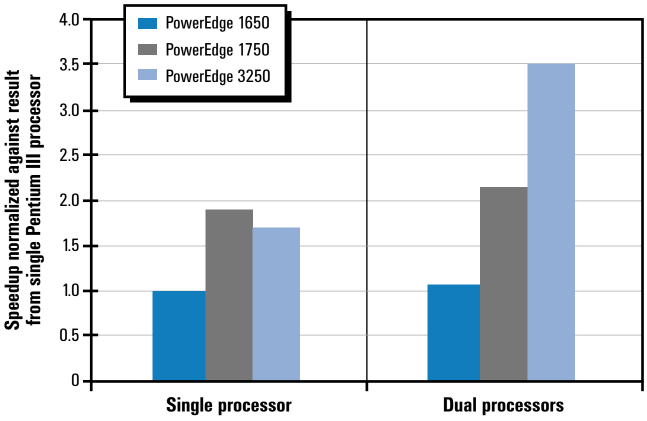


Figure 4. Comparison of the Integer Sort benchmark on three Intel-based platforms

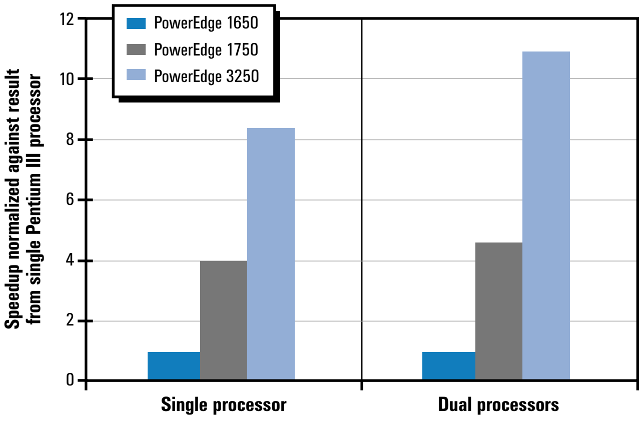


Figure 5. Comparison of the Multigrid benchmark on three Intel-based platforms

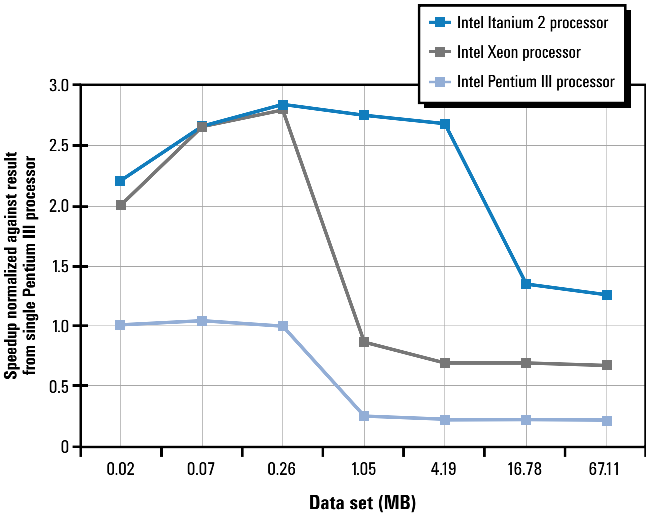


Figure 6. Normalized results of the ASPCG benchmark running on a single Intel processor

instruction set in Intel Xeon processors and also the larger number of floating-point operations that the Itanium 2 processor can perform per clock cycle.

Because MG is also cache-friendly, the larger L3 cache on the PowerEdge 3250 helped to improve performance. When comparing the performance of one to two processors, the PowerEdge 3250—with its greater memory bandwidth—provided over 30 percent more scalability versus the Intel Xeon processor–based PowerEdge 1750 and the Pentium III–based PowerEdge 1650, which provided negligible scalability.

Using ASPCG to evaluate cache performance and memory bandwidth

Developed at Virginia Polytechnic Institute and State University, the ASPCG kernel benchmark solves a linear system of equations generated by a Laplace equation in Cartesian coordinates. The kernel was developed to evaluate cache performance and memory bandwidth of different architectures. ASPCG performs a block decomposition of the full domain and uses a conjugate gradient (CG) solver and a two-level Additive Schwarz preconditioner.

The ASPCG benchmark can be performed on different data set sizes. The results are output in millions of FLOPS (MFLOPS). All test runs on the three platforms used a single processor.

In Figure 6, the x-axis shows the data set sizes, which are based on relevant block sizes. The y-axis shows the speedup factor for each platform normalized against the result from the first data point size of Pentium III. The graph illustrates the performance of each platform when accessing different levels of cache. For example, the Itanium 2 processor reached maximum performance when the code was running within its cache size, but as the data set increased beyond 6 MB (the L3 cache size of the Itanium 2 processor), performance dropped until the processor accessed physical memory and stabilized. The Itanium 2 processor significantly outperformed the other platforms with the help of its large 6 MB L3 cache.

Intel has continually increased frontside bus speed and improved memory bandwidth, thereby enabling hardware integrators to create more balanced systems.

Designing a powerful building block for HPC

Results from the tests performed by the Dell team showed that Intel has not only improved clock speeds for its processors, but it also has increased frontside bus speed and memory bandwidth, thereby enabling hardware integrators to create more balanced systems.

Test results indicate that the Itanium 2–based PowerEdge 3250 with its greater L3 cache, higher memory bandwidth, and larger addressable memory space, outperformed the two other Intel processor–based Dell platforms—making the PowerEdge 3250 an attractive building block for high-performance computing.

Rizwan Ali (rizwan_ali@dell.com) is a systems engineer working in the Scalable Systems Group at Dell. His current research interests are performance benchmarking and high-speed interconnects. Rizwan has a B.S. in Electrical Engineering from the University of Minnesota.

Jenwei Hsieh, Ph.D. (jenwei_hsieh@dell.com) is an engineering manager of the Scalable Systems Group at Dell. Jenwei is responsible for developing high-performance clusters. He has published extensively in the areas of multimedia computing and communications, high-speed networking, serial storage interfaces, and distributed network computing. Jenwei has a Ph.D. in Computer Science from the University of Minnesota and a B.E. from Tamkang University in Taiwan.

Onur Celebioglu (onur_celebioglu@dell.com) is a systems engineer in the Scalable Systems Group at Dell. His current areas of focus are networking and HPC interconnects. Onur has an M.S. in Electrical and Computer Engineering from Carnegie Mellon University.

FOR MORE INFORMATION

ASPCG benchmark:
http://www.hpcfd.me.vt.edu/codes.shtml#ASPCG_CODE

High-Performance Linpack: http://www.netlib.org/benchmark/hpl

IA-32 Intel Architecture Software Developer's Manual, Volume 1: Basic Architecture:
http://developer.intel.com/design/pentium4/manuals/245470.htm

NAS Parallel Benchmark: http://www.nas.nasa.gov/Software/NPB

A High-Performance Computing Cluster

for Parallel Simulation of Petroleum Reservoirs

This article examines compute-intensive benchmark tests performed on an Intel® Xeon™ processor–based Dell™ PowerEdge™ server cluster running the Red Hat® Linux® operating system. To determine how configuration factors affect a standards-based high-performance computing (HPC) cluster, the study explores the scalability of a parallel petroleum reservoir simulator when varying the network interconnect type, test case size, and configuration. Each test was run twice to compare the results of single-processor versus dual-processor nodes.

BY KAMY SEPEHRNOORI, PH.D.; BARIS GULER; TAU LENG, PH.D.; VICTOR MASHAYEKHI, PH.D.; AND REZA ROOHOLAMINI, PH.D.

Petroleum reservoir simulation requires detailed—and computationally intensive—geological and physical models. Traditionally, this work has been performed on supercomputers, mainframes, and powerful workstations. Now developers are experimenting with reservoir simulation on loosely coupled parallel systems known as clusters, enabled by recent hardware advances in the design of standardized processors, memory and I/O subsystems, storage, and network interconnects. Software developments bolstering the use of clusters include the rapid adoption of the Linux® operating system (OS), new compiler and math libraries, and the middleware Message Passing Interface (MPI) specification. Altogether, these advances are helping to make standards-based clusters a legitimate and cost-effective alternative to proprietary high-performance computing (HPC) systems for reservoir simulations.

This article is an update of a previous study in the ongoing research collaboration between the Center for Petroleum and Geosystems Engineering at The University of Texas at Austin and the Scalable Systems Group at Dell.[1] Building upon the previous study, small and large simulation cases were developed to benchmark different cluster configurations using various network interconnects. The goal for the testing reported in this article was to understand how problem size affects simulator performance and to determine which

[1] See "Parallel Simulation of Petroleum Reservoirs on High-Performance Clusters" by the Center for Petroleum and Geosystems Engineering, The University of Texas at Austin, in collaboration with Dell, in *Dell Power Solutions,* Issue 2, 2001.

network connections might be optimal for each simulation case. These simulations were carried out using both single-processor and dual-processor nodes to analyze symmetric multiprocessing (SMP) cluster performance.

Introducing reservoir simulation

Predicting the performance of petroleum reservoirs under a variety of operations is essential for reservoir management. Making oil recovery more productive through the injection of substances such as carbon dioxide, nitrogen, surfactant, and polymers is called Improved Oil Recovery (IOR). By conducting a series of reservoir simulations that address different IOR processes, researchers can assess the risk involved in each process to help determine the best possible recovery method for a given reservoir before it is implemented in the field.

A reservoir simulator consists of a coupled set of nonlinear partial differential equations and constitutive relations that describe the physical process occurring in a petroleum reservoir. The governing partial differential equations are solved using numerical methods. Black oil and compositional simulators are the most commonly used simulators. Black oil simulators use water, oil, and gas phases for modeling fluid flow in a reservoir, whereas compositional simulators use phases with different chemical species for modeling physical processes that occur in a reservoir. Compositional simulators are required in simulations that account for mixing fluids that have drastically different properties and for displacing oil by miscible fluids. Traditionally, black oil simulators have been used more often than compositional simulators because compositional simulators are more complicated and have intensive memory and CPU requirements.

However, advances in memory and CPU technologies in the last decade have enabled researchers to increase the use of compositional simulations. Measurement methods also have evolved, allowing researchers to use more data to characterize reservoirs. Current simulations require hundreds of thousands of cells or millions of unknowns; as simulation complexity increases, future simulations will call for millions of cells with multimillions of unknowns.

Parallel reservoir simulators have the potential to solve larger, more realistic problems than previously possible. Researchers at The University of Texas at Austin have been developing a new-generation code for parallel computers that is designed to perform fast, accurate, and efficient high-resolution simulations of fluid flow in permeable media. Their research scope involves development of new, complex physical and chemical models and accurate numerical methods implemented in a parallel-processing environment.

> Predicting the performance of petroleum reservoirs under a variety of operations is essential for reservoir management.

The simulator

The governing equations for fluid flow in a permeable medium are implemented in a new, parallel, fully implicit compositional simulator called the General Purpose Adaptive Simulator (GPAS). A finite-difference method, which divides a continuous domain into small cells, is used to solve the governing partial differential equations. As the number of cells increases, more accurate results can be obtained; however, a greater number of cells increases the computation time. A fully implicit solution results in a system of nonlinear equations that are solved using Newton's method. Numerical solution of nonlinear equations requires large, sparse linear systems of equations. The linear systems are handled with solvers from the Portable Extensible Toolkit for Scientific Computation (PETSc).[2]

To handle the complicated tasks associated with parallel processing, researchers at The University of Texas have developed an Integrated Parallel Accurate Reservoir Simulator (IPARS) framework.[3] The goal is to separate the physical model development from parallel processing. Communications between the simulator framework and a physical model are carried out through hooks provided within the IPARS. These hooks are composed of FORTRAN subroutine calls, and all the communications among processors for a physical model are performed in these routines. The physical model developers insert these calls into the codes that perform the corresponding tasks. IPARS provides input and output, memory management, domain decomposition, and message passing among processors to update overlapping regions.

Establishing the computational environment

The computational environment for the benchmarks in this study comprised a cluster of 64 Dell™ PowerEdge™ 2650 servers using Fast Ethernet, Gigabit Ethernet,[4] and Myricom® Myrinet® interconnects.

[2] Balay, S., W.D. Gropp, L.C. McInnes, and B.F. Smith, "Efficient management of parallelism in object-oriented numerical software libraries," *Modern Software Tools for Scientific Computing,* edited by E. Arge, A.M. Bruaset, and H.P. Langtangen, Boston: Birkhauser, 1997, pages 163–202.

[3] Wang, P., I. Yotov, M. Wheeler, T. Arbogast, C. Dawson, M. Parashar, and K. Sepehrnoori, "A new generation EOS compositional reservoir simulation: Part I. Formulation and discretization," paper SPE 37979 presented at the 1997 SPE Reservoir Simulation Symposium, San Antonio, Texas, June 1997; and Wang, P., S. Balay, K. Sepehrnoori, J. Wheeler, J. Abate, B. Smith, and G.A. Pope, "A fully implicit parallel EOS compositional simulator for large-scale reservoir simulation," paper SPE 51885 presented at the 1999 SPE Reservoir Simulation Symposium, Houston, Texas, February 1999.

[4] This term indicates compliance with IEEE® standard 802.3ab for Gigabit Ethernet, and does not connote actual operating speed of 1 Gbps. For high-speed transmission, connection to a Gigabit Ethernet server and network infrastructure is required.

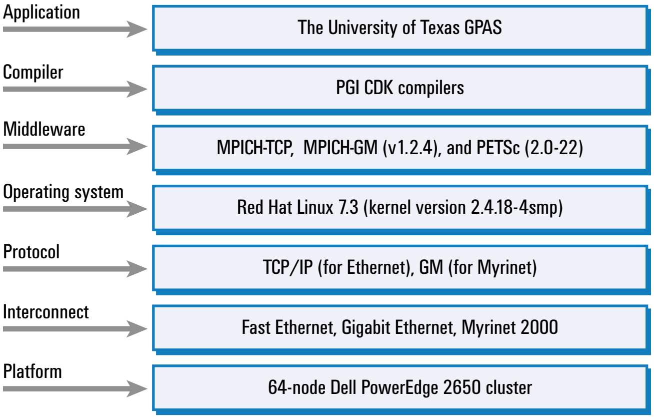

Application
The University of Texas GPAS
Compiler
PGI CDK compilers
Middleware
MPICH-TCP, MPICH-GM (v1.2.4), and PETSc (2.0-22)
Operating system
Red Hat Linux 7.3 (kernel version 2.4.18-4smp)
Protocol
TCP/IP (for Ethernet), GM (for Myrinet)
Interconnect
Fast Ethernet, Gigabit Ethernet, Myrinet 2000
Platform
64-node Dell PowerEdge 2650 cluster

Figure 1. Architectural stack of the test environment

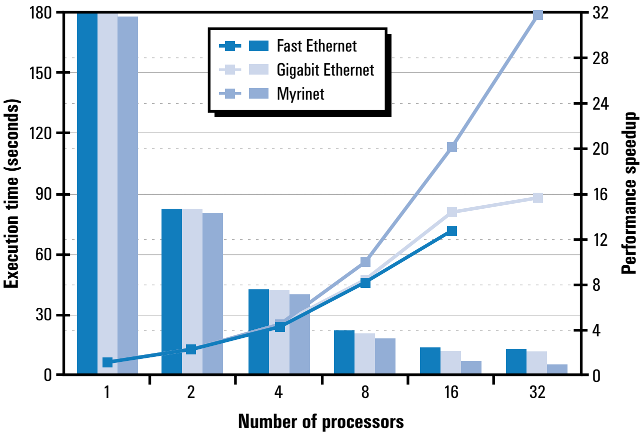


Figure 2. Simulator execution time and speedup plots for small case (16 × 224 × 8) using single-processor nodes

Each PowerEdge 2650 had two Intel® Xeon™ processors running at 2.4 GHz with 512 KB of Level 2 (L2) cache and 2 GB of double data rate (DDR) RAM operating on a 400 MHz frontside bus (FSB).

The PowerEdge 2650 was equipped with Peripheral Component Interconnect Extended (PCI-X) slots capable of supporting the peak network traffic generated by the Myrinet network interface card (NIC). The operating system installed in the cluster was Red Hat® Linux 7.3 (kernel version 2.4.18-4smp). The GPAS was compiled using the PETSc library and the PGI® CDK™ (Cluster Development Kit) C/C++ and FORTRAN 77/90 compilers. Figure 1 shows the architectural stack of the computational environment.

Simulation cases

To evaluate the scalability of the GPAS using different interconnects on the same cluster, two simulation cases were developed. The first, a small simulation case with 3-D grid 16 × 224 × 8 used approximately 350 MB of memory. The case comprised 28,672 grid blocks, requiring 229,376 unknowns to be solved at each time step in the process. A total of 10 time steps was taken, and each time step (ranging from 0.1 to 20 days) consisted of 2 to 5 Newton iterations. The first case simulated 100-day gas injection with one injection well and one production well. The reservoir had homogeneous permeability and porosity fields.

The computational environment for the benchmarks in this study comprised a cluster of 64 Dell PowerEdge 2650 servers using Fast Ethernet, Gigabit Ethernet, and Myricom Myrinet interconnects.

The second simulation case built upon the first, adding complexity by enlarging the reservoir, introducing heterogeneity, and increasing the number of injection and production wells. The reservoir in the larger simulation was divided into 197,120 (77 × 256 × 10) grid blocks, requiring 1.5 million unknowns to be solved simultaneously at each time step. To run the case, the GPAS required approximately 1.7 GB of memory.

Performance metric

In parallel applications, efficiency is usually measured by *speedup*. In this study, the speedup of a cluster with N processors was defined as *speedup* $= t_1 / t_N$, where t_1 is the execution time running on one processor and t_N is the execution time running on N processors. The ideal speedup of a parallel simulation with N processors is N; that is, the program runs N times faster. However, as the number of processors becomes larger, researchers usually observe a speedup of less than N. The performance reduction can be attributed to increased interprocessor communication (known as memory contention) or to network contention arising from a cluster whose nodes are SMP machines, or both. Such overhead is not encountered using only one processor. Sometimes the performance reduction can be caused by an inefficient program that does not decompose the application evenly.

Analyzing the interconnect test results

The study tested three different interconnects: Fast Ethernet, Gigabit Ethernet, and Myrinet. Researchers executed both simulation cases—small (16 × 224 × 8) and large (77 × 256 × 10)—to analyze the performance of the cluster using different interconnects. The initial test used only one processor per compute node. Figure 2 shows the execution times (bars) and speedups (lines) of the small case from 1 node to 32 nodes. The left y-axis in the figure indicates the execution time (in seconds), and the right y-axis indicates the performance speedup of the simulator. As shown in Figure 2, the simulator performed and scaled best on the low-latency, high-bandwidth Myrinet cluster. In fact, the Myrinet cluster performance scaled linearly from 1 to 32 nodes regardless of the simulation size—the result of cache and memory aggregations when more than one node is used.

For configurations using few nodes, the clusters with high-latency Fast Ethernet or Gigabit Ethernet interconnects offered similar performance to the Myrinet cluster.

For configurations using few nodes, the clusters with high-latency Fast Ethernet or Gigabit Ethernet interconnects offered similar performance to the Myrinet cluster. As node count increased, communication among processes increased—while computation performed by each node decreased. The performance differences among the three types of interconnects became increasingly significant after scaling to as few as four nodes. As Figure 2 shows, the most efficient Gigabit Ethernet cluster was the 16-processor configuration, because this interconnect did not scale when the processor count went beyond 16.

Similar observations were made in the large simulation case, as shown in Figure 3. However, performance differences became significant after 16 nodes instead of 4 nodes as in the small simulation case. Because the communication-to-computation ratio of the large case was lower than the communication-to-computation ratio of the small case, the simulator showed better scalability for all interconnects in the large case. In the 64-node configuration, Myrinet provided the best performance and super linear speedup. In addition, both Fast Ethernet and Gigabit Ethernet clusters could be scaled beyond 32 nodes efficiently. This indicates that for reservoir simulation, the importance of the interconnect is truly case dependent; even Fast Ethernet might be a viable solution for large simulations.

Single-processor and dual-processor node performance

In addition to testing the three types of interconnects at two different problem sizes, researchers examined simulator performance using one processor per node and then using two processors per node. Two processes running in a node will compete for resources such as memory and I/O. In particular, the shared memory bus will be a performance bottleneck when the memory is accessed at the same time by both processes. In addition, the communication traffic generated by the two processes may create another potential bottleneck on I/O resources, such as the PCI bus or the NIC, or both.

Researchers could calculate the ratio of time spent in each section of the large simulation case based on the recorded data, using a dual-processor node configuration as compared to a single-processor node configuration. The runs were performed on the Myrinet cluster. Figure 4 shows the ratios of the plotted execution times. A ratio larger than 1.0 indicates some contention inherent to SMPs. For example, highly computational- and data-dependent sections of the simulation such as "Update viscosity and relperm" and "Update dependent variables" suffered the most from shared memory architecture.

All the ratios, except one, decrease as the number of processors increases, which indicates that memory contention diminished as more processors were used. Because the amount of data per processor was reduced as the number of processors increased, less tendency for contention existed when retrieving the data. Conversely, increased communication among processes resulted in network contention on the NIC. For the communication-sensitive sections such as the "Total linear solver time," the ratio

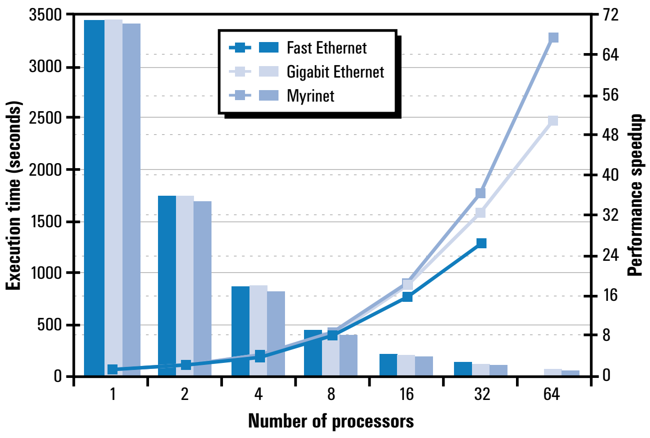


Figure 3. Simulator execution time and speedup plots for large case (77 × 256 × 10) using single-processor nodes

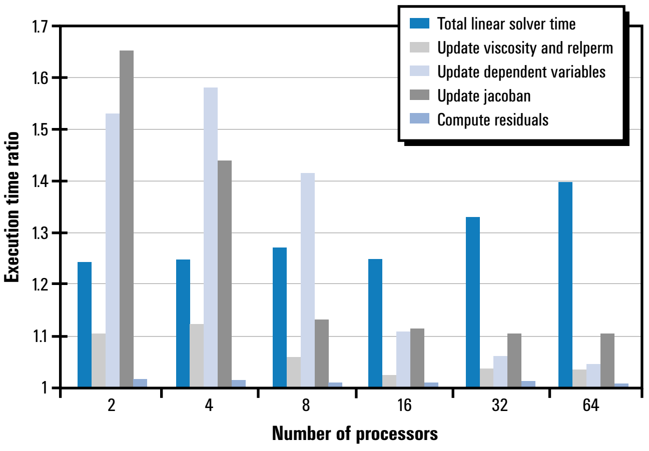


Figure 4. Execution time ratio of dual-processor versus single-processor nodes for sections of the large simulation case (77 × 256 × 10) on the Myrinet cluster

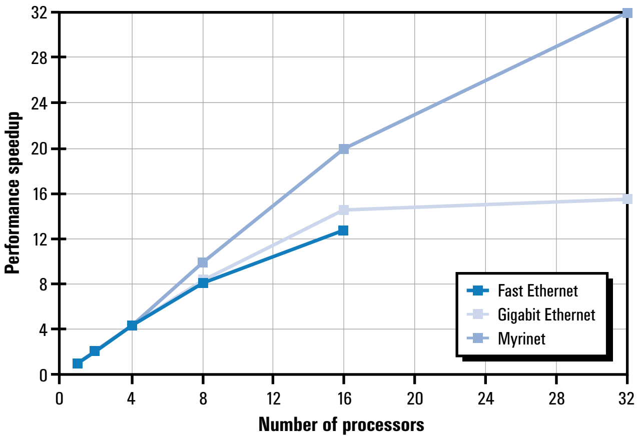


Figure 5. Simulator speedup plots for small case (16 × 224 × 8) using dual-processor nodes

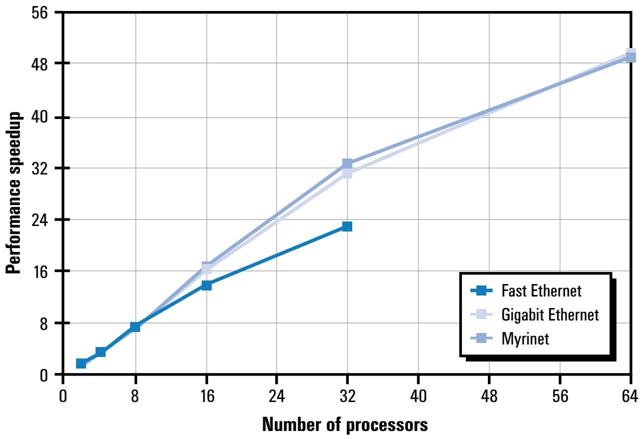


Figure 6. Simulator speedup plots for large case (77 × 256 × 10) using dual-processor nodes

is considerably larger—greater than 1.2—and increases with the processor count as shown in Figure 4.

For reservoir simulation, the importance of the interconnect is truly case dependent; even Fast Ethernet might be a viable solution for large simulations.

Figures 5 and 6 show the performance speedup curves using dual-processor nodes for the small and large simulation cases and varying the interconnect. Overall, single-processor configurations (see Figures 2 and 3) outperformed dual-processor configurations anywhere from 2 percent to 27 percent. As seen in both Myrinet runs, the aggregated cache and memory improvements were offset by both memory contention and network contention; this performance degradation was caused by routing communication traffic for two processors through a single NIC. In the Myrinet cluster, the performance degradation was 18 percent and 27 percent for small and large cases, respectively. The Gigabit Ethernet cluster was the least affected by the use of SMPs as compute nodes, showing 9 percent performance degradation for the small case and 2 percent for the large. In fact, using 64 processors with 49.8 speedup, the Gigabit Ethernet cluster performed and scaled slightly better than the Myrinet cluster. Using SMPs affected Fast Ethernet cluster performance similarly for both small and large cases, exhibiting 13 percent performance degradation.

Configuring standards-based clusters to scale optimally

Examining both small and large simulation cases using Fast Ethernet, Gigabit Ethernet, and Myrinet cluster interconnects enabled conclusions about the scalability of the simulator. Additionally, comparing single-processor node to dual-processor node cluster performance led to the following observations:

- A low-latency, high-bandwidth interconnect, Myrinet performed best in all simulations.
- For the small case simulation, when relatively few computations were performed on each processor, low-latency interconnects provided better scalability.
- As the problem size increased, a high-bandwidth interconnect became more important than a low-latency interconnect to produce consistent scalability.
- The aggregations of cache and memory helped computational performance.
- I/O resource contention may introduce additional communication overhead in an SMP cluster.
- Memory contention that can arise in SMP nodes may contribute to overall performance degradation ranging from 2 percent to 27 percent when comparing dual-processor nodes to single-processor nodes.
- The Myrinet and Gigabit Ethernet clusters performed almost the same when using 32 dual-processor nodes (64 processors).

The results of this study and others like it can help system administrators make wise infrastructure choices. By determining the most effective network interconnects and number of processors per node for their specific computational problem sets, administrators can improve the scalability of clusters built using Dell PowerEdge servers running Linux—thereby leveraging inexpensive, standards-based HPC clusters as a cost-effective alternative to proprietary supercomputers, mainframes, and workstations for many computational purposes.

Acknowledgments

This research was supported in part by the U.S. Department of Energy, the sponsors of the Joint Industry Project for Reservoir Simulation at the Center for Petroleum and Geosystems Engineering at The University of Texas at Austin, and Dell Inc.

References

Balay, S., W.D. Gropp, L.C. McInnes, and B.F. Smith. "Efficient management of parallelism in object-oriented numerical software libraries." *Modern Software Tools for Scientific Computing.* Edited by E. Arge, A.M. Bruaset, and H.P. Langtangen. Boston: Birkhauser, 1997, pages 163–202.

Killough, J.E., and C.A. Kossack. "Fifth comparative simulation project: Evaluation of miscible flood simulators." Paper SPE 16000 presented at the Ninth SPE Symposium on Reservoir Simulation. San Antonio, Texas, February 1987.

Uetani, T., B. Guler, and K. Sepehrnoori. "Parallel Reservoir Simulation on High Performance Clusters." *The 6th World Multi-Conference on Systemics, Cybernetics, and Informatics.* Orlando, Florida, July 2002.

Wang, P., I. Yotov, M. Wheeler, T. Arbogast, C. Dawson, M. Parashar, and K. Sepehrnoori. "A new generation EOS compositional reservoir simulation: Part I. Formulation and discretization." Paper SPE 37979 presented at the 1997 SPE Reservoir Simulation Symposium. San Antonio, Texas, June 1997.

Wang, P., S. Balay, K. Sepehrnoori, J. Wheeler, J. Abate, B. Smith, and G.A. Pope. "A fully implicit parallel EOS compositional simulator for large-scale reservoir simulation." Paper SPE 51885 presented at the 1999 SPE Reservoir Simulation Symposium. Houston, Texas, February 1999.

Kamy Sepehrnoori, Ph.D. (kamy@mail.utexas.edu) is the Bank of America Centennial Professor in the Department of Petroleum and Geosystems Engineering at The University of Texas at Austin. His research interests include computational methods, reservoir simulation, parallel computations, applied mathematics, and enhanced oil recovery modeling. He has a B.S. in Mechanical Engineering, an M.S. in Aerospace Engineering, and a Ph.D. in Petroleum Engineering from The University of Texas at Austin.

Baris Guler (baris_guler@dell.com) is a systems engineer and advisor in the Scalable Systems Group at Dell. His current research interests are parallel processing, diskless HPC clusters, performance benchmarking, reservoir engineering and simulation, and numerical methods. Baris has a B.S. in Petroleum and Natural Gas Engineering (PNGE) from the Middle East Technical University in Turkey and an M.S. in PNGE from Pennsylvania State University. He is currently a Ph.D. candidate in Petroleum and Geosystems Engineering at The University of Texas at Austin.

Tau Leng, Ph.D. (tau_leng@dell.com) is an engineering manager of the Scalable Systems Group at Dell. His current research interests are parallel processing, distributed computing systems, compiler optimization, and performance benchmarking. Tau has a B.S. in Mathematics from the Fu Jen Catholic University in Taiwan, an M.S. in Computer Science from Utah State University, and a Ph.D. in Computer Science from the University of Houston.

Victor Mashayekhi, Ph.D. (victor_mashayekhi@dell.com) is a senior technical manager of the Scalable Systems Group at Dell. His product development responsibilities at Dell have included all the cluster product offerings from Dell. His current research interests are distributed systems, database management systems, computer-supported cooperative work, concurrent software engineering, multimedia systems, clustering, and interconnect technologies. Victor has a B.A., M.S., and Ph.D. in Computer Science from the University of Minnesota.

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

FOR MORE INFORMATION

Center for Petroleum and Geosystems Engineering at The University of Texas at Austin: http://www.cpge.utexas.edu

Myricom Myrinet products: http://www.myri.com

Dell HPCC products: http://www.dell.com/hpcc

Using Intel Hyper-Threading Technology to Achieve Computational Efficiency

The effects of Intel® Hyper-Threading Technology on system performance vary according to application characteristics and the system's configuration. This article describes Dell™ tests of Hyper-Threading Technology used in several cluster configurations—varying the interconnect and processor cache size—and provides results and recommendations.

BY ONUR CELEBIOGLU; TAU LENG, PH.D.; RIZWAN ALI; AND JENWEI HSIEH, PH.D.

Hyper-Threading Technology, a feature of Intel® Xeon™ and Intel Pentium® 4 processors, makes a single physical processor appear as two logical processors to the operating system. Hyper-Threading duplicates the architectural state on each processor, while sharing one set of execution resources.[1] This duplication allows a single physical processor to execute instructions from different threads in parallel rather than in serial, potentially leading to better processor utilization and overall performance.

However, sharing system resources, such as cache or memory bus, may degrade system performance. Previous studies have shown that Hyper-Threading can improve the performance of some applications, but not all.[2] Performance gains may vary depending on the cluster configuration, such as communication fabric or cache size, and on the applications running on the cluster.

In high-performance computing (HPC) clusters, software developers often use standard message-passing systems such as Message Passing Interface (MPI) or Parallel Virtual Machine (PVM) to achieve parallelism in applications. For optimal performance, in most cases the number of processes spawned is equal to the number of processors in the cluster. Therefore, parallelized applications can benefit from Hyper-Threading, because doubling the number of processors means the number of processes spawned is doubled, allowing parallel tasks to execute faster. Applying Hyper-Threading—and thus doubling the processes that simultaneously run on the cluster—also increases the utilization rate of the processors' execution resources. Although performance may improve, doubling simultaneous processes may introduce overhead in the following ways:

- **Cache access:** Logical processes of the same physical CPU may compete for access to the caches, which potentially generates more cache-miss situations.
- **Memory contention:** More processes running on the same compute node may increase memory contention if the processes access the memory bus or communicate through shared memory simultaneously.

[1] For more information, see "Hyper-Threading Technology Architecture and Microarchitecture" by D.T. Marr et al in *Intel Technology Journal,* February 2002.

[2] For more information, see "A Study of Hyper-Threading in High-Performance Computing Clusters" by Tau Leng, Ph.D.; Rizwan Ali; Jenwei Hsieh, Ph.D.; and Christopher Stanton in *Dell Power Solutions,* November 2002.

- **Communication traffic:** More processes on each node increase the message passing within and between nodes, which can oversubscribe the communication capacity of the shared memory, the I/O bus, or the interconnect networking, and thus create performance bottlenecks.

This article discusses how the Dell™ HPC cluster team tested clusters with various interconnects, cache sizes, and configuration parameters to understand the performance and adaptability of using Hyper-Threading in HPC clusters.

Understanding how interconnects affect cluster performance

The interconnect networks individual compute nodes together, converting a group of computers into a single system: an HPC cluster that can execute parallel jobs. Interconnect performance can significantly affect cluster performance when the applications running on the cluster are communication intensive. When Hyper-Threading is applied, cluster performance may suffer if the interconnect cannot accommodate increased communications among the doubled processes. Conversely, if the interconnect can handle the increased traffic, cluster performance may benefit from Hyper-Threading.

Comparing three interconnects

The Dell team used three interconnects—Gigabit Ethernet,[3] Myricom® Myrinet®, and Quadrics™ QsNet—to understand the performance impact on each cluster when Hyper-Threading is enabled versus disabled.

Ethernet is the most popular networking configuration today. However, it uses the conventional TCP/IP protocol, which can be too inefficient for communication-intensive jobs in a dedicated system area network such as an HPC cluster. Nevertheless, familiarity and low cost make Fast Ethernet or Gigabit Ethernet the popular choice for many applications.

Myrinet and QsNet are two leading interconnects for HPC clusters. Both provide low-latency and high-bandwidth end-to-end communication between two nodes in a cluster. Myrinet, a connectionless interconnect, uses packet-switching technologies derived from experimental massively parallel processing (MPP) networks. For communication among the nodes in a cluster, Myrinet employs the GM message-passing system, an efficient, low-level communication protocol. This system gives user processes direct access to the network interface, avoiding immediate copies of data and bypassing the operating system networking stack.

The QsNet interconnect employs a similar technique—using a network interface based on the Quadrics Elan application-specific integrated circuit (ASIC)—to efficiently transfer messages between the communicating nodes. The protocol implemented by Elan provides low-latency, protected access to a global virtual memory through remote direct memory access (RDMA) operations.

Interconnect	Gigabit Ethernet	Myrinet	QsNet
Network interface card (NIC)	Broadcom BCM5701 64-bit/66 MHz Peripheral Component Interconnect (PCI)	Myrinet-2000 M3F-PCI64C 64-bit/66 MHz PCI	Elan III 64-bit/66 MHz PCI
Network switch	Gigabit Ethernet switch	Myrinet-2000 32-port switch	Quadrics QM-S16 16-port switch
Link speed	1 Gbps	2 Gbps	2.72 Gbps
Topology	Non-blocking	Clos network	Fat tree
Protocol	TCP/IP	GM	Elan

Figure 1. Specifications for three interconnects used in Dell testing

Figure 1 summarizes the specifications of the three interconnects the Dell team used in its testing.

Establishing the test environment

The testing environment consisted of a 16-node cluster of Dell PowerEdge™ 6650 servers interconnected with Gigabit Ethernet, Myrinet, and QsNet. Each PowerEdge 6650 had two Intel Xeon processors MP at 1.6 GHz with 512 KB level 2 (L2) cache, 1 MB L3 cache, 4 GB of double data rate (DDR) RAM, and a 400 MHz frontside bus. The PowerEdge 6650 uses the ServerWorks™ Grand Champion™ HE chipset, which accommodates up to 16 registered DDR 200 (PC1600) dual in-line memory modules (DIMMs) with a four-way interleaved memory architecture.

The PowerEdge 6650 also is equipped with two integrated Broadcom® Gigabit Ethernet adapters. The clustered servers ran the Red Hat® Linux® 7.3 operating system, kernel version 2.4.18-4smp. To use the QsNet interconnect, the team upgraded the kernel to a special version based on kernel 2.4.18. The benchmark programs were compiled using the Intel C Compiler and the Intel Fortran Compiler.

The Dell team used the NAS (NASA Advanced Supercomputing) Parallel Benchmarks (NPB) to test the performance of the clusters.[4] These benchmarks—derived from computational fluid dynamics (CFD) applications—consist of five kernels and three simulated CFD applications that are designed to gauge parallel computing performance.

Evaluating the effect of cache size on cluster performance

The FT benchmark requires solving a 3-D partial differential equation using Fast Fourier Transforms (FFTs). The Dell team used the FT benchmark to understand the cache effect of Hyper-Threading

[3] This term indicates compliance with IEEE® standard 802.3ab for Gigabit Ethernet, and does not connote actual operating speed of 1 Gbps. For high-speed transmission, connection to a Gigabit Ethernet server and network infrastructure is required.

[4] For more information about the NAS Parallel Benchmarks, visit http://www.nas.nasa.gov/Software/NPB.

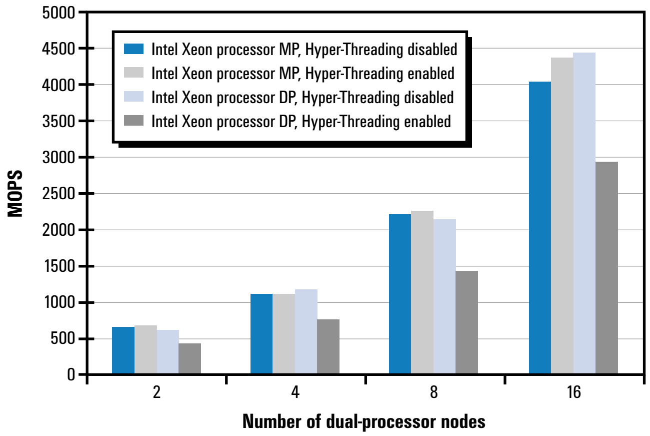


Figure 2. Comparing results for the Fast Fourier Transform benchmark, FT, between Intel Xeon processor MP and Intel Xeon processor DP clusters using Myrinet interconnects, with Hyper-Threading enabled and disabled

because FT performance is sensitive to cache size—the larger the cache size, the better the performance.

The team ran FT on the 16-node test cluster equipped with 32 Intel Xeon processors MP and compared this performance to results found in a previous study, which tested a 16-node cluster with 32 Intel Xeon processors DP.[5] Both clusters used the Myrinet interconnect, but had different processor cache sizes. Intel Xeon processors DP have no L3 cache; Intel Xeon processors MP have 1 MB of L3 cache. Comparing results from clusters using the same interconnect, but different processor cache sizes, helps demonstrate the effect of varying cache size on cluster performance when Hyper-Threading is enabled.

The cluster with Intel Xeon processors MP showed improved performance when Hyper-Threading was enabled, while the Intel Xeon processor DP cluster showed reduced performance with Hyper-Threading (see Figure 2). In the Intel Xeon processor MP cluster, the L3 cache was large enough to accommodate the memory accesses for both logical processors sharing the caches. As a result, the L3 cache hits compensated for the increased L2 cache misses, improving performance because of better CPU resource utilization. For the Intel Xeon processor DP cluster, the performance drop occurred because the cache miss rate increased considerably without L3 cache support.

Testing the effect of interconnects on cluster performance

The Dell team also ran tests on the Intel Xeon processor MP cluster using the Integer Sort (IS) benchmark. IS performs an integer sorting operation that is important in particle method codes and tests both integer computation speed and communication performance. This benchmark is unique because it does not use floating-point arithmetic, but does require significant data communication. Hence, IS is particularly useful for testing the communication capability of clusters. In this study, the IS results showed the communication performance for each interconnect. In addition, the benchmark indicated whether doubling the processes by using Hyper-Threading helped the sorting to complete faster—or decreased performance by increasing traffic.

The IS results in Figure 3 show that the three interconnects performed as expected. QsNet performed best, followed by Myrinet, then Gigabit Ethernet. Figure 3 also shows that Hyper-Threading increased IS performance slightly for Gigabit Ethernet and Myrinet configurations. However, the QsNet configuration, despite giving the highest overall performance, did not benefit from Hyper-Threading as much.

For applications to benefit from Hyper-Threading, the CPU must contain underutilized shared resources. With a very low-latency interconnect such as QsNet, the CPU is occupied performing computations, minimizing the amount of idle time while a process waits for message-passing activities. Thus, enabling Hyper-Threading for IS in the QsNet configuration realized no performance improvement. Furthermore, for QsNet, the team observed a slight performance drop across large node counts because of contention for memory and use of the interconnect.

Enabling Hyper-Threading helped the other configurations, resulting in an overall performance gain. Although contention for I/O was higher for the slower interconnects, utilizing the idle CPU cycles caused by high message-passing latency negated the effect of I/O contention and resulted in a net performance increase with Hyper-Threading enabled.

Matching interconnect, application, and Hyper-Threading to maximize performance

The benchmark testing on the three interconnects yielded several conclusions:

- **Communication-intensive applications:** In the IS benchmark, QsNet gave the highest overall cluster performance for communication-intensive tasks, but this cluster benefited least from Hyper-Threading. On the other hand, the slowest of the three interconnects, Gigabit Ethernet, benefited more from Hyper-Threading in communication-intensive benchmarks.

- **Compute-intensive applications:** Fine-tuned arithmetic operations are less likely to see improved performance from Hyper-Threading because the CPU resources are already highly utilized.

[5] For more information, see "A Study of Hyper-Threading in High-Performance Computing Clusters" by Tau Leng, Ph.D.; Rizwan Ali; Jenwei Hsieh, Ph.D.; and Christopher Stanton in *Dell Power Solutions,* November 2002.

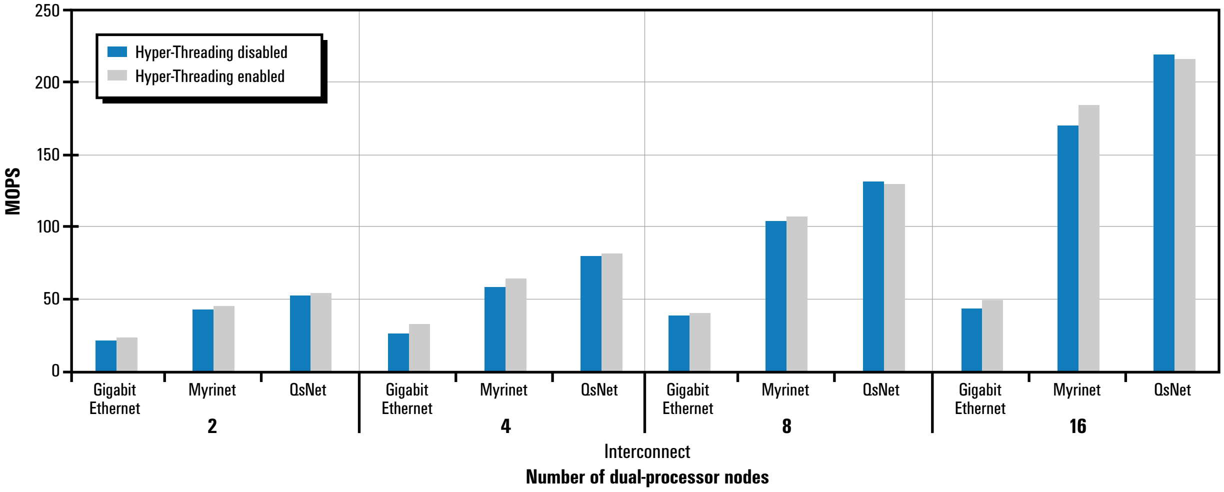


Figure 3. Comparing Integer Sort benchmark results among Gigabit Ethernet, Myrinet, and QsNet interconnects, with Hyper-Threading enabled and disabled

- **Communication- and compute-intensive applications:** As modeled by the IS benchmark, the overhead and latency associated with communications may cause the CPU resources to be underutilized. For interconnects like Myrinet and QsNet, this overhead is negligible because the interconnects use efficient message-transfer mechanisms such as RDMA. The CPU involvement in message transfer for these interconnects is already minimal.

For applications that require intensive communication, enabling Hyper-Threading and doubling the number of processes in the cluster node creates I/O contention for shared memory and the interconnect subsystems. This performance degradation can be offset by Hyper-Threading that claims underutilized CPU cycles because of high message-passing latency, as is the case with Gigabit Ethernet. For a very low-latency interconnect such as QsNet, the computation already occupies CPU cycles, yielding little offset from Hyper-Threading; consequently, resource contention leads to a performance drop. This behavior often occurs in higher node count configurations.

The decision whether to use Hyper-Threading Technology in a cluster depends not only on the characteristics of the application, but also on the capacities of the CPU caches and interconnect communication. Even in the same type of application, Hyper-Threading may affect clusters with various interconnects and cache configurations differently. The guidelines in this article can help IT organizations choose the best type of interconnect for their applications. In addition, the guidelines can show how to enable or disable Hyper-Threading for improved cluster performance and computational efficiency.

Onur Celebioglu (onur_celebioglu@dell.com) is a systems engineer in the Scalable Systems Group at Dell. His current areas of focus are networking and HPC interconnects. Onur has an M.S. in Electrical and Computer Engineering from Carnegie Mellon University.

Tau Leng, Ph.D. (tau_leng@dell.com) is an engineering manager of the Scalable Systems Group at Dell. His current research interests are parallel processing, distributed computing systems, compiler optimization, and performance benchmarking. Tau has a B.S. in Mathematics from the Fu Jen Catholic University in Taiwan, an M.S. in Computer Science from Utah State University, and a Ph.D. in Computer Science from the University of Houston.

Rizwan Ali (rizwan_ali@dell.com) is a systems engineer working in the Scalable Systems Group at Dell. His current research interests are performance benchmarking and high-speed interconnects. Rizwan has a B.S. in Electrical Engineering from the University of Minnesota.

Jenwei Hsieh, Ph.D. (jenwei_hsieh@dell.com) is an engineering manager of the Scalable Systems Group at Dell. Jenwei is responsible for developing high-performance clusters. He has published extensively in the areas of multimedia computing and communications, high-speed networking, serial storage interfaces, and distributed network computing. Jenwei has a Ph.D. in Computer Science from the University of Minnesota and a B.E. from Tamkang University in Taiwan.

FOR MORE INFORMATION

Intel Xeon processor DP:
http://www.intel.com/design/Xeon/prodbref

Intel Xeon processor MP:
http://www.intel.com/products/server/processors/server/xeon_mp

NAS parallel benchmarks:
http://www.nas.nasa.gov/Software/NPB

The OSCAR Toolkit:

Current and Future Developments

The Open Source Cluster Application Resources (OSCAR) software helps simplify cluster management by offering the best-known standards-based tools in one toolkit. This article discusses recent and proposed improvements to develop an OSCAR framework, such as increased modularity and support for diskless and high-availability clusters.

BY THOMAS NAUGHTON; STEPHEN L. SCOTT, PH.D.; YUNG-CHIN FANG; PHIL PFEIFFER, PH.D.; BENOÎT DES LIGNERIS, PH.D.; AND CHOKCHAI LEANGSUKSUN, PH.D.

Managing computing clusters is a challenging, often daunting task. A clustered system can take hours to install, configure, and test—even for experts. After installation, more hours of work follow: managing accounts, configuring applications, upgrading software, and solving or working around problems. The Open Source Cluster Application Resources (OSCAR) toolkit was developed to simplify the installation, configuration, and management of computing clusters.

The OSCAR toolkit, an open source software project, is the product of the OSCAR working group—industry, academic, and research organizations that contribute to OSCAR design and development. The group's fundamental goal is to simplify cluster management by integrating the best-known standards-based tools.

Since the November 2002 release of OSCAR 2.0, the toolkit has evolved toward a more modular framework with enhancements to the build system and package facilities.[1] Two new projects build on this framework: Thin OSCAR (a diskless version) and High-Availability OSCAR.

Increasing modularity: The OSCAR framework

OSCAR has evolved from a simple bundle of standard clustering software into a framework for configuration management. This framework, like the original bundle, supports the fundamental goal of using the best-known standards-based tools to simplify cluster computing. The framework was developed in response to the need for a more modular distribution, which became apparent as users requested new packages and OSCAR grew in popularity. Discrepancies among the various Linux® distributions also served to hasten the change.

Moreover, because of increasing demand, the number of popular high-performance computing (HPC) architectures has grown to include the following three classes: Intel® Pentium® processor–based architecture (IA-32); Intel Itanium® processor–based architecture (IA-64); and AMD® Opteron™ processor–based architecture (x86-64). Although OSCAR fully supports only IA-32 (and experimentally supports IA-64), the OSCAR working group plans to fully support all the successful HPC architectures.[2]

[1] For more information, see "Looking Inside the OSCAR Cluster Toolkit" by Thomas Naughton; Stephen L. Scott, Ph.D.; Brian Barrett; Jeff Squyres; Andrew Lumsdaine, Ph.D.; Yung-Chin Fang; and Victor Mashayekhi, Ph.D., in *Dell Power Solutions,* November 2002. The "References" section at the end of this article lists additional information resources.

[2] At the time of publication, the current release of OSCAR was version 2.3.

To increase modularity, developers began reworking OSCAR by decoupling the procedures for building and configuring packages from the software ultimately installed on clusters. This decoupling has allowed developers to reuse the framework for other purposes, including diskless and high-availability cluster system configurations. The framework has since evolved to include a build and configuration system, a database, and an environment management tool.

Build system expedites cluster setup

A characteristic cluster management tool is a *build system,* a mechanism for configuring and installing cluster software. The OSCAR framework's build system, the System Installation Suite (SIS), uses the SystemImager® tool to support cluster node initialization. SystemImager generates a representative directory structure that can be traversed and manipulated just as a standard file system. This structure, or *image,* is downloaded to a cluster's nodes during system installation and then dynamically configured with host-specific values like IP and network interface card (NIC) addresses. Administrators can then change and redistribute the master image to a cluster's nodes as part of ongoing cluster maintenance. SystemImager also supports multicast-based installation, a new feature included in the OSCAR 2.3 release.

The OSCAR build system includes a graphical user interface-(GUI-) based wizard that guides users through cluster setup and enables them to download and select packages for installation. Once a cluster's nodes have been initialized, OSCAR runs a series of simple tests to verify the installation. Administrators can later add or remove nodes by using this wizard.

All components (except the build system and database) of the OSCAR framework use the Oak Ridge National Laboratory (ORNL) Cluster Command and Control (C3) tools to perform actions across the entire cluster in parallel.[3] The C3 tools facilitate parallel execution and scatter/gather operations that are essential for administrators and users alike, and the tools can operate on single or multiple clusters simultaneously.

OSCAR packages facilitate application installation

An OSCAR package provides a simple way to wrap an application for use in a target system. A package comprises an application; a meta file; and an optional set of scripts, tests, and documentation. Administrators configure packages by using the OSCAR package application programming interface (API). This API enables a package to query for data during cluster installation and execute configuration and setup routines specific to the application. The applications themselves are bundled using the standard RPM™ (Red Hat® Package Manager) format.

> The Open Source Cluster Application Resources (OSCAR) toolkit was developed to simplify the installation, configuration, and management of computing clusters.

OSCAR's modular packaging facility enables the toolkit to install software obtained from either the standard release or online package repositories. The OSCAR Package Downloader (OPD) aids in the acquisition of updated or third-party packages. OPD, a command-line tool, is similar in nature to tools that access the Comprehensive Perl Archive Network (CPAN). The installation wizard uses a GUI version of OPD called OPDer.

The package API supports package-specific interactive operations during installation. Administrators can, for example, configure packages to prompt for a default version if more than one instance of a package is selected. A package also can convey its platform dependencies to OSCAR by using the XML meta file. OSCAR uses these dependencies to generate the list of installable packages for the current platform and omits any packages whose dependencies are not met.

OSCAR database provides central repository

The OSCAR database (ODA) provides a central, package-updatable repository of information about a cluster's nodes and packages. Users access ODA through a command-line interface (CLI), which abstracts the underlying database engine (MySQL®) and provides a simpler interface for data access, including shortcuts for common package operations. The API scripts use these shortcuts to gather information such as available packages and number of nodes. In turn, any OSCAR package can use this information to initialize configuration files, such as the /etc/c3.conf configuration file in C3, and other package-specific settings.

OSCAR framework supports environment management

The task of managing user execution environments is a critical part of normal system administration. Several common applications such as Parallel Virtual Machine (PVM), Message Passing Interface (MPI), and Portable Batch System (PBS) use environment variables to integrate their operations with their host platforms. More generally, easy command-line access to any application depends on the proper setting of a shell's command-path environment variable. If the relevant variables are missing or undefined, the system becomes extremely difficult to use.

[3] Cluster Command and Control (C3) uses a configuration file to characterize the logical cluster structure. In OSCAR, a default configuration file is created for the user. See the C3 documentation cited in the "References" section for details on how to customize the configuration to achieve better scalability.

Users commonly employ two methods to maintain environment variables. The first method, which requires users to maintain their own environments, is cumbersome and error-prone. The second, which requires users to source a system-wide file at login, is superior to the first—but must be done well to help ensure proper system operation.

The OSCAR framework supports environment management with Env-Switcher, an extension of the Modules utility for cluster environments. Env-Switcher adds support for environment variable persistence over successive logins, and for cluster-wide environment consistency through an administrator-configurable script. Env-Switcher uses Modules to update system shells, propagating these modifications throughout a cluster. A package that must add items to the environment may do so by providing an Env-Switcher script that is loaded into the user's shell upon login.

Env-Switcher supports user-configurable defaults by providing a CLI for querying and modifying system-wide and user-level settings. Administrators could, for example, set a path for a system-wide default MPI implementation and let users inherit this default or override it with alternative user-level selections.

Developing OSCAR: New working groups

Since November 2002, the Open Cluster Group, OSCAR's parent group, has added two new working groups that are developing OSCAR-based configurations: Thin OSCAR for diskless environments and High-Availability OSCAR for high-availability system operation (see Figure 1).

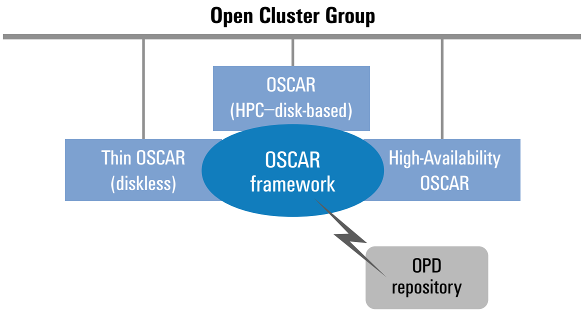


Figure 1. The OCG working groups share the OSCAR framework for cluster installation and management

Thin OSCAR supports diskless clustering technique

The three classic rationales for diskless operation are decreased cost from eliminating the local hard drive, which is no longer needed to host the system's initial image; increased reliability from eliminating the local hard drive; and speed, if the network connection can be accessed more quickly than the local disk. Thin OSCAR supports diskless operation by extending standard OSCAR with three new classes of nodes:

- **Diskless:** Has no local disk; uses the master node for long-term storage over the network
- **Diskful:** Supports local disks
- **Systemless:** Uses local disks only for temporary storage or swap space, not to store resident operating systems

Currently, Thin OSCAR directly supports only diskless and diskful nodes; systemless nodes require manual configuration.

Diskless clusters use the Dynamic Host Configuration Protocol (DHCP) for image distribution to manage system initialization. At startup, a diskless node uses its network interface to download a small boot image. The diskless bootstrap procedure then uses this image to load a complete run image on the target node.

The current Thin OSCAR implementation uses a collection of Perl scripts and supporting libraries to transform a regular SIS image into the two RAM disks necessary for diskless and systemless operation. The first RAM disk—the boot image—ensures that a node has network connectivity. The second RAM disk—the run image—is built directly from the SIS image and contains the complete system that will run on the node. Some material is directly copied from the SIS image, while other parts are exported to the nodes through the Network File System (NFS) in a read-only mode, directly from the SIS image.

Currently, Thin OSCAR is provided through an OSCAR package. The standard OSCAR installation wizard handles most, but not all, of the steps needed for Thin OSCAR installation. Administrators can use the CLI to build the two RAM disks necessary for diskless and systemless operation; this step occurs outside the actual installation wizard. The OSCAR installer (currently a simple wizard approach) is being redesigned to be more modular, as was the case with the package API. Once the installer development is complete, the external steps will be integrated into the installation process.

Most of the Thin OSCAR concepts have been validated; they are now being used in a production environment—a 180-node diskless cluster at the Université de Sherbrooke in Québec, Canada. This fully functional cluster is expected to grow to approximately 300 nodes by the end of 2003.

High-Availability OSCAR helps eliminate single points of failure

High availability has become critical to the fundamental mission of high-performance computing. HPC systems are being tasked to run very large and complex applications, whose runtimes exceed their host systems' aggregated mean time between failure (MTBF) rate. To effectively run code on HPC machines, high-availability computing techniques are necessary to prevent *code thrashing*—the rerunning of many code segments when attempting to recover from a failure. Instead, systems must maximize the underlying HPC environment.

High-availability computing differs somewhat from fault-tolerant computing. The former is proactive by sensing and preventing potential failures, whereas the latter is usually costly and reactive. In fault-tolerant computing, replicated components execute the same instructions at the same time, so even if one fails the application keeps running, with no difference other than perhaps a reduced performance level based on the percent of resources lost. A significant cost of fault tolerance is the lockstep redundant execution and frequent checkpointing that must occur regardless of failure state.

To support high-availability requirements, clustered systems must eliminate single points of failure. During its first phase of operation, the new HA-OSCAR working group will seek to eliminate single points of failure in the current HPC release of OSCAR through active/hot-standby configurations and eventually through implementing $n+1$ active/active distributions. These strategies involve hardware duplication and network redundancy, common techniques for improving the reliability and availability of computer systems.

Initial efforts in this area will focus on supporting a duplicate cluster master node. Various techniques currently exist for implementing such an architecture, which includes active/active, active/hot standby, and active/cold standby. Figure 2 shows the High-Availability OSCAR cluster system architecture.

Members of the HA-OSCAR group who have experimented with these techniques plan to incorporate Linux Virtual Server and heartbeat mechanisms into an initial active/hot-standby High-Availability OSCAR distribution. This architecture will be extended to support active/active high availability after release of the hot-standby distribution. The active/active architecture will provide for better resource utilization, because both master nodes will be simultaneously active and providing services.

The dual master nodes will run redundant OpenPBS, Maui, DHCP, Network Time Protocol (NTP), Trivial FTP (TFTP), NFS, rsync, and Simple Network Management Protocol (SNMP) servers. If a master node fails, all functions provided by that node will fail over to the second, redundant master node. All service requests will continue to be met, although at a reduced performance rate (in theory, at 50 percent of the master node's peak or busy hours).

HA-OSCAR also will support a high-availability network through redundant Ethernet ports on every machine, and duplicate switching fabrics (such as network switches and cables) for the entire network configuration. This functionality is designed to enable every node in the cluster to be present on two or more data paths within the cluster networks. Backed with this Ethernet redundancy, the cluster will achieve higher network availability. Furthermore, when the entire network is up, techniques such as channel bonding of messages across the redundant communication paths may improve communication performance.

Enhancing OSCAR: Future developments

The current OSCAR development path includes enhancements to the package API and a simplified procedure for package addition and removal. These enhancements will support more precise characterizations of package and target information (such as supported distribution and architecture), which in turn will support package updates between full OSCAR releases. The developers also are reworking the toolkit to simplify upgrading packages between releases. These enhancements will eliminate dependencies between packages and full OSCAR releases, streamlining package management.

Developers are making the OSCAR Installer more flexible by adding support for a user-configurable front end—for example, a CLI interface—along with a user-configurable build engine specific to Thin or High-Availability OSCAR. The Scientific Discovery Through Advanced Computing (SciDAC) Scalable System Software (SSS) project is using OSCAR as a deployment vehicle to simplify the evaluation and eventual adoption of the SSS suite. The OSCAR framework itself also is being fitted with an SSS interface for use as a build and configuration system.

The need for better cluster management software has spurred the development of OSCAR over the past three years and has led to the Thin OSCAR and HA-OSCAR working groups. The efforts to modularize the toolkit for better reusability have enabled these new groups to reuse the existing framework for cluster installation and management. In addition to the modularized package system, the build system has added multicast (SIS) support and improved scalability for the parallel toolset (C3).

Managing clusters is a time-consuming task. OSCAR helps to reduce the installation, configuration, and management costs of a cluster. The OSCAR framework assists with integrating reusable

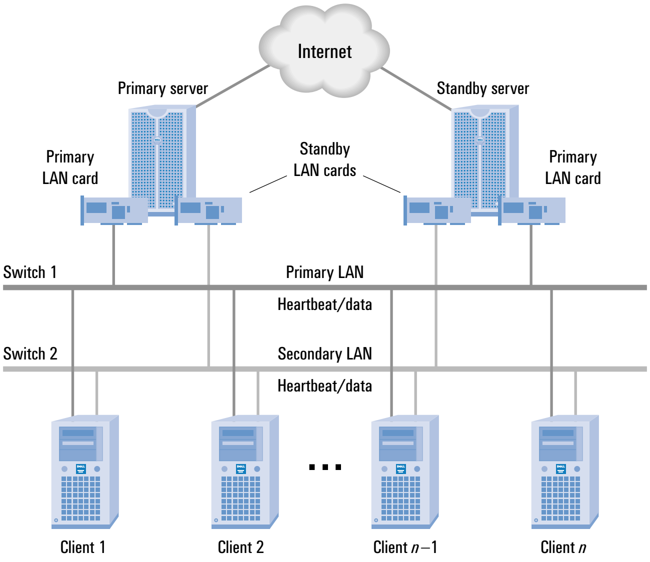


Figure 2. The High-Availability OSCAR cluster architecture facilitates failover

software on a cluster. This reusability and extensibility is a primary goal for the future developments of OSCAR.

Acknowledgments

Work by ORNL was supported by the U.S. Department of Energy under Contract DE-AC05-00OR22725.

The authors would like to thank the OSCAR community, fellow developers, and users, as well as the organizations that support member contributions to the project.

References

"Abstract Yourself with Modules." http://www.usenix.org/publications/library/proceedings/lisa96/pwo.html.

Clusters for High Availability: A Primer of HP Solutions. 2nd ed. Englewood Cliffs, N.J.: Prentice-Hall, 2001.

"C3 Power Tools." *Distributed and Parallel Systems: Cluster and Grid Computing.* Boston: Kluwer Academic Publishers, 2002.

"Development, installation and maintenance of Elix-II, a 180-node diskless cluster running thin-OSCAR." http://hpcs2003.ccs.usherbrooke.ca/papers/desLigneris_03.pdf.

"Open Source Cluster Application Resources (OSCAR): design, implementation and interest for the [computer] scientific community." http://hpcs2003.ccs.usherbrooke.ca/papers/desLigneris_01.pdf.

"OSCAR Clusters." *Proceedings of the Ottawa Linux Symposium.* Ottawa, Canada: Linux Symposium, 2003. See also: http://archive.linuxsymposium.org/ols2003/Proceedings/All-Reprints/Reprint-Scott-OLS2003.pdf.

"The OSCAR Revolution." *Linux Journal,* June 2002. See also: http://www.linuxjournal.com/article.php?sid = 5559.

"The Penguin in the Pail–OSCAR Cluster Installation Tool." *The 6th World MultiConference on Systemics, Cybernetics and Informatics (SCI 2002)*. Orlando, Fla.: International Institute of Informatics and Systemics, 2002.

"System Installation Suite: Massive Installation for Linux." *Proceedings of the Ottawa Linux Symposium.* Ottawa, Canada: Linux Symposium, 2002. See also: http://www.linux.org.uk/~ ajh/ols2002_proceedings.pdf.gz.

"Thin-OSCAR: Design and future implementation." http://hpcs2003.ccs.usherbrooke.ca/papers/desLigneris_02.pdf.

Thomas Naughton (naughtont@ornl.gov) is a research associate in the Computer Science and Mathematics Division, ORNL. Thomas has a B.S. in Computer Science and a B.A. in Philosophy from the University of Tennessee—Martin, and an M.S. in Computer Science from Middle Tennessee State University.

Stephen L. Scott, Ph.D. (scottsl@ornl.gov) is a senior research scientist in the Computer Science and Mathematics Division, ORNL. Stephen leads the cluster computing effort at ORNL. He has a B.A. from Thiel College in Greenville, Pennsylvania, and an M.S. and Ph.D. from Kent State University in Kent, Ohio.

Yung-Chin Fang (yung-chin_fang@dell.com) is a member of the Scalable Systems Group at Dell. Yung-Chin has a bachelor's degree in Computer Science from Tamkang University in Taiwan and a master's degree in Computer Science from Utah State University. He is currently working on his doctorate degree.

Phil Pfeiffer, Ph.D. (phil@etsu.edu) is a professor of computer science at East Tennessee State University (ETSU) and is currently working with Dr. Scott on an ORNL-sponsored non-instructional leave. Phil has a B.S. in Computer Science from Yale University, and an M.S. and Ph.D. from the University of Wisconsin-Madison.

Benoît des Ligneris, Ph.D. (benoit.des.ligneris@ccs.usherbrooke.ca) is a postdoctoral fellow in the Scientific Computing Center of the Université de Sherbrooke. He has a B.Sc. and M.Sc. from Pierre & Marie Curie University in Paris, and an M.Sc. and Ph.D. from the Université de Sherbrooke.

Chokchai Leangsuksun, Ph.D. (box@latech.edu) is an associate professor of computer science and an affiliate of the Center for Entrepreneurship and Information Technology (CEnIT) at Louisiana Tech University. He also is a director of the eXtreme Computing Research (XCR) Group. He has a B.Eng. from Khon Kean University, Thailand, and an M.S. and Ph.D. in Computer Science from Kent State University in Kent, Ohio.

FOR MORE INFORMATION

C3 power tools:
http://www.csm.ornl.gov/torc/C3

Open Cluster Group:
http://www.openclustergroup.org

OSCAR Cluster User's Guide:
http://oscar.sourceforge.net/docs/oscar_user_2.3.pdf

OSCAR project:
http://oscar.sourceforge.net

System Installation Suite (SIS):
http://www.sisuite.org

Thin OSCAR working group:
http://thin-oscar.ccs.usherbrooke.ca

Leveraging Commercial Parallel Technologies and Open Source Support to

Scale Out Enterprise Linux Clusters

High-performance clusters of standard computing and networking components can be a cost-effective way to execute massively parallel processing tasks. Two products from MPI Software Technology, Inc. (MSTI)—ChaMPIon/Pro™ MPI-2.1 middleware for the Linux® operating system and Felix™ cluster deployment and management software—help enable IT organizations to create reliable clusters of all sizes, providing particularly effective support for high-performance computing (HPC) configurations that comprise more than 1,000 nodes (or 2,000 processors).

BY ANTHONY SKJELLUM

High-performance computing (HPC) clusters enable cost-effective parallel processing by combining standards-based hardware components to aggregate computing power across many computational processors. The basic infrastructure of such clusters is often provided by open source technologies: notably the Linux® operating system and related tools. Enterprises can further lower their total cost of ownership (TCO) by selectively replacing open source components with commercial software products that offer high performance, scalability, and support, along with comprehensive cluster management tools to configure, maintain, and upgrade those clusters.

MPI Software Technology, Inc. (MSTI) ChaMPIon/Pro™ middleware, the first commercial MPI-2.1 middleware release for the Linux operating system, is designed to maximize parallel programming in large and midsize clusters. ChaMPIon/Pro, which supports Red Hat® Linux as well as other Linux distributions, offers a level of middleware support previously available only in supercomputers costing substantially more per node than Dell™ PowerEdge™ cluster nodes.

Another MSTI offering, Felix™ cluster deployment and management software, provides comprehensive cluster management and support. Felix functions as both a platform for software deployment and a management tool for maintenance and upgrade of large-scale Red Hat Linux–based clusters. Felix works in concert with the Dell OpenManage™ Server Administrator server monitoring and management tool.

MSTI, in collaboration with Dell Inc., recently deployed two terascale clusters using Felix: a 2,008-node cluster at the University at Buffalo (UB), a campus of The State University of New York, and a 1,500-node cluster at the National Center for Supercomputing Applications (NCSA). The newly completed NCSA cluster uses Felix for day-to-day maintenance and ChaMPIon/Pro for parallel programming.

Enhancing terascale cluster support using ChaMPIon/Pro middleware

Terascale clusters contain thousands of processors and offer theoretical peak performance in the teraflop range.[1] Terascale clusters are designed to solve extremely large-scale computational problems. However, as enterprises continue to grow HPC clusters, the complexity increases exponentially for achieving top performance from applications while maximizing cluster resource utilization and uptime. Easy-to-use middleware can help improve the scalability, performance, and robustness of a terascale cluster.

In an HPC cluster, users expect a common interface even though each computing node runs a separate instance of the operating system. Presenting a single, parallel supercomputer–like interface to system administrators, programmers, and other cluster users is the task of middleware such as ChaMPIon/Pro.

ChaMPIon/Pro provides portable, high-throughput, low-latency, and low-overhead data communication for high-performance parallel applications. The middleware implements the complete MPI-2 standard for efficient message passing among tasks that are cooperating to solve a parallel problem. ChaMPIon/Pro is designed to enhance the underlying performance of end-user applications, rather than simply to provide the fastest communication path to a small subset of particular-size messages. This approach enables applications in terascale clusters to run faster.

ChaMPIon/Pro demonstrates superior performance running standard legacy benchmark programs—such as the popular Linpack benchmark, used for measuring the performance of supercomputers—but it also excels in real-world applications (including computational fluid dynamics code) that use asynchronous communication to overlap computation with communication.[2] Therefore, while a cluster's network is moving data, cluster nodes can continue useful computation work. By using an independent message progress engine and a blocking message arrival–detection strategy, ChaMPIon/Pro avoids wasting code resources and reduces the impact of file and network I/O latency associated with parallel computation.

ChaMPIon/Pro uses scalable algorithms to launch and execute parallel processes, thereby overcoming resource limits imposed by the operating system.

On symmetric multiprocessing (SMP) nodes, ChaMPIon/Pro enables applications to exploit fine-grain parallelism using threads for intranode communication, while continuing to use messages to facilitate data transfer between nodes (see Figure 1). Applications that use threads inside nodes and messages between nodes include quantum chemistry codes, but many other applications can benefit from such an approach. Because ChaMPIon/Pro is thread-aware, it enables multithreaded applications to perform well in the presence of multiple communicating threads.

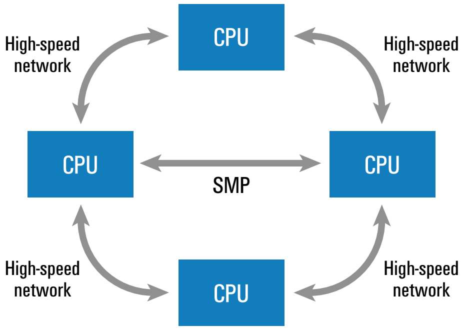


Figure 1. ChaMPIon/Pro offers high-performance communication across nodes

ChaMPIon/Pro uses scalable algorithms to launch and execute parallel processes, thereby overcoming resource limits imposed by the operating system. Any middleware allows parallel programs to work once started; however, getting parallel programs started grows more difficult as clusters scale. ChaMPIon/Pro resource-conscious spanning-tree strategies help parallel programs start. The middleware also provides system administrators with options for executing parallel processes in ways that best meet site requirements for configuration, operation, and security.

ChaMPIon/Pro is tested with an industry-leading correctness test suite to ensure compliance with the MPI-2 standard. This test suite is a combination of previously created public suites with MSTI's own testing methodologies, creating the most coverage available. Evaluation of the MPI-2 standard and this test suite helps ensure that ChaMPIon/Pro is a compliant interface that robustly implements the standard. In addition to test suites, ChaMPIon/Pro has been run against demanding applications such as MPQC—the Massively Parallel Quantum Chemistry program—to validate the MPI-2 functionality demanded by such codes.

ChaMPIon/Pro offers IT organizations interoperability with a wide variety of cluster computing tools, including parallel debuggers and performance-analysis software. In addition, the product supports networking interconnects such as Gigabit Ethernet,[3]

[1] One teraflop = 10^{12} floating-point operations per second.

[2] For Linpack experiment data, see http://www.mpi-softtech.com/company/publications/files/ChaMPIonPerfNumbersSept15_2003.pdf.

[3] This term indicates compliance with IEEE® standard 802.3ab for Gigabit Ethernet, and does not connote actual operating speed of 1 Gbps. For high-speed transmission, connection to a Gigabit Ethernet server and network infrastructure is required.

Myricom® Myrinet®, Quadrics™ QsNet, and the InfiniBand™ architecture, enabling cluster owners to migrate to a different network backbone without changing software providers.

Streamlining terascale cluster deployment and life-cycle maintenance using Felix software

Felix is a commercial-grade Linux clustering application that provides an effective mechanism for system administrators to install, manage, maintain, and upgrade Linux-based clusters—particularly clusters of greater size and complexity. The Felix application manages clusters through a master-slave relationship, whereby the master node acts as a control center for cluster management, configuration, update, and repair while compute nodes (or slaves) perform the actual parallel computations. Having a single point of control from which administrators can execute parallel commands and install parallel nodes across the entire cluster helps enterprises to reduce systems and network administration time—improving efficiency both in cluster maintenance and in management of third-party software packages. This single point of control also can be replicated for high availability; a methodology for supporting reconstruction of master nodes and their configuration is part of the Felix approach to cluster management.

Felix is designed to recognize clusters in a generic manner, independent of detailed hardware configurations in the cluster nodes (see Figure 2). For instance, Felix allows a mix-and-match of Red Hat Linux distributions on different nodes in the cluster. This facilitates support for multiple operating system versions, customized versions, or both on the same cluster.

Felix deploys and configures Dell OpenManage software, which comprises Dell OpenManage Server Administrator, Dell OpenManage IT Assistant, and Embedded Server Management packages. Dell OpenManage can provide hardware-level monitoring and control of multithousand-node terascale clusters. The combination of a centralized management console, compute-node hardware management agent, and embedded remote access provides low-level hardware management for each node. Felix provides cluster administrators with software-level monitoring and management that complements Dell OpenManage to deliver a complete, easy-to-use cluster management system.

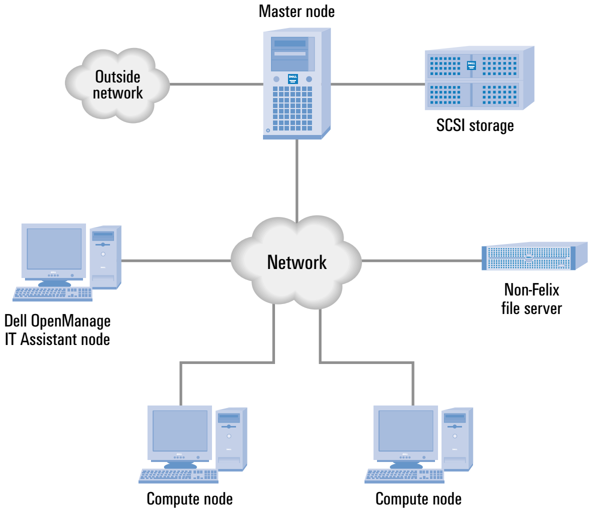


Figure 2. Felix provides administrators with a single point of control over the cluster

Having a single point of control from which administrators can execute parallel commands and install parallel nodes across the entire cluster helps enterprises to reduce systems and network administration time.

Felix uses industry-standard protocols and Intel®-standard Preboot Execution Environment (PXE) technology to achieve remote installation of the compute nodes. Through a node-naming and IP-addressing scheme, Felix facilitates the deployment of large clusters—first by building smaller and more manageable subclusters and then by linking the subclusters through hierarchical relationships. Different communication modes are supported between master and compute nodes, including secure and nonsecure modes.

Finally, troubleshooting and diagnostic mechanisms are an integral part of the Felix application. Felix can be configured to provide failover for the master node, with minimal or no perceptible interruption in cluster applications. In addition, Felix offers specific procedures to recover both compute nodes and master nodes from hard disk and network interface failures. Master node failover takes advantage of the unique Felix design. The Felix view of the cluster is summarized in a set of configuration files. These configuration files are backed up on the compute nodes through a specialized backup procedure. If the master node fails, one of the compute nodes assumes the role of the master node after the execution of an automated command. This failover capability allows enterprises running HPC clusters to reduce the downtime caused by compute node failures.

Enabling more effective use of present and future terascale clusters

Although many enterprises use clusters in the 8- to 256-node mainstream, the number of clusters comprising 1,000 nodes or more is steadily increasing. As clusters continue to scale out, IT organizations require tools that allow cluster performance and manageability to keep pace.

MSTI offers HPC middleware that enables high performance on a massive scale. Designed to overcome operating system resource limitations and enhance application speed, MSTI ChaMPIon/Pro middleware helps applications to run efficiently even on terascale clusters. Other tools, such as the MSTI Felix application, can help save IT organizations time and money by reducing administrative overhead and system downtime. Felix, a comprehensive cluster management tool, is designed to speed cluster deployment, centralize management, and enable efficient node recovery and system reconfiguration for high availability.

Practical parallel production computing is a reality, but clusters remain effective only if they perform well and remain cost-effective at any size. MSTI cluster tools help enable system administrators to take full advantage of high-performance, open source hardware such as Dell servers, and proprietary software such as Dell OpenManage, to build large-scale clusters that satisfy today's processing needs and scale efficiently to meet tomorrow's demands.

Anthony Skjellum (tony@mpi-software.com) received his Ph.D. from the California Institute of Technology (Caltech) in 1990. His work on parallel software and message-passing systems at the Lawrence Livermore National Laboratory at Caltech and Mississippi State University provided leadership in the standardization of the Message Passing Interface (MPI) from 1993 to 1997, while the founding of MPI Software Technology, Inc. created a commercial provider for such middleware. Anthony is one of the four original authors of the MPICH Argonne/Mississippi State model implementation of the MPI-1 standard, which is still widely used today.

MPI Software Technology, Inc. (www.mpi-softtech.com) is a maker of middleware, tools, and products for HPC clusters. MSTI is a Dell Professional Services (DPS) partner, and provides deployment, deployment-enabling software, and optional middleware to complement Red Hat Linux and freeware stacks. MSTI validates freeware stacks for DPS and Dell Inc. as part of cluster-offering validation, in conjunction with the Scalable Systems Group at Dell.

FOR MORE INFORMATION

ChaMPIon/Pro and Felix software: http://www.mpi-softtech.com

MPI-2 Message Passing Interface Forum: http://www.mpi-forum.org

MPQC: http://aros.ca.sandia.gov/~cljanss/mpqc

Leveraging the Dell 2161DS

Remote Console Switch in High-Performance Computing Clusters

The Dell™ 2161DS Remote Console Switch provides administrators with secure, real-time access to high-performance computing (HPC) clusters from any location. In addition, the 2161DS switch utilizes inexpensive Category 5 cables rather than bulky keyboard, video, mouse (KVM) cables, to fit more servers in a single rack.

BY STEPHEN M. HAHN AND ED KRACH

As high-performance computing (HPC) clusters grow in power and sophistication, data center administrators need more effective management tools to connect, organize, monitor, and troubleshoot servers in real time. The Dell™ 2161DS Remote Console Switch helps simplify the administration of HPC cluster environments by providing uncomplicated cabling and remote, secure access to any server in the cluster at any time.

Understanding Remote Console Switch components

The Dell 2161DS Remote Console Switch consists of a rack-mountable keyboard, video, mouse (KVM) switch incorporating Avocent™ KVM Over IP™ technology (see Figure 1).[1] The switch provides secure analog (local) and digital (remote) connectivity to all major operating systems, server platforms, and serial devices. Administrators can access the switch locally through the On-Screen Display (OSD) graphical user interface (GUI) and remotely through the Dell Remote Console Software (RCS) Java™-based client interface.[2]

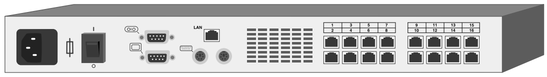


Figure 1. Dell 2161DS Remote Console Switch

The Dell RCS management application enables administrators to remotely control the 2161DS switch, thereby providing access to all attached servers in the HPC cluster. Remote administrators can have the same level of control over their servers through the network as they would have at the rack through the analog console. Up to 128 servers can be connected to a single switch. Each additional 2161DS can add one local and two remote users, providing seamless access to a nearly infinite number of servers.

[1] For more information on the 2161DS and KVM Over IP technology, see "KVM Over IP: A New Approach to Server Management" by Robert Bernstein and Aaron Jennings in *Dell Power Solutions,* August 2002.

[2] For more information on RCS, see "Managing Servers Using the Dell 2161DS Remote Console Switch" by Stephen M. Hahn in *Dell Power Solutions,* November 2002.

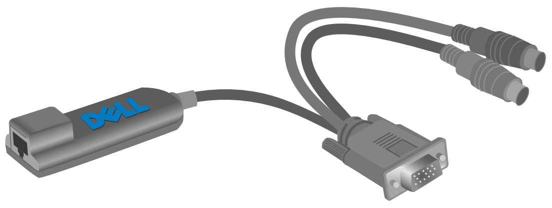


Figure 2. Dell Server Interface Pod

Features of the 2161DS switch include:

- KVM network device in a rack-mountable 1U chassis
- 16 Category 5 (Cat 5)/RJ-45 ports for server connections
- Connectivity through both a 10/100BaseT Ethernet port and a local analog KVM port
- Up to 1600 × 1200 video resolution of local analog connection
- Up to 1280 × 1024 video resolution of remote digital connection; maximum cable length of 32.8 feet (10 meters) between the 2161DS switch and the server
- Firmware upgradable through the network

The Dell Server Interface Pod (SIP) and Port Expansion Module (PEM) are additional accessories.

Server Interface Pod. The SIP, shown in Figure 2, replaces standard KVM cables by converting a server's KVM signals, then sending the converted signals through a single Cat 5 cable. The server provides power to the SIP through PS/2 connections to ensure that the server works with or without connectivity to the 2161DS switch. By using this keep-alive technology, the SIP provides continuous keyboard and mouse emulation.

The Dell 2161DS Remote Console Switch helps simplify the administration of HPC cluster environments by providing uncomplicated cabling and remote, secure access to any server in the cluster at any time.

Port Expansion Module. The PEM is an optional accessory that passively expands one of the 16 server ports on a 2161DS (see Figure 3). Up to 8 servers can be connected to a PEM, allowing for expansion to 128 servers for each 2161DS.

Benefiting from switch administration and control

The ability to securely administer and control the HPC environment from local and remote terminals is a critical benefit of the 2161DS switch. Security is provided through multilevel password protection and 128-bit Secure Sockets Layer (SSL) encryption.

Flexible access from local and remote locations effectively provides administrators with centralized control over multiple server sites. Using the 2161DS, administrators can control planned downtime better and manage unplanned downtime more efficiently. Enabling administrators to access a failed server through the switch expedites troubleshooting and reduces time to recovery.

Flexible access from local and remote locations effectively provides administrators with centralized control over multiple server sites.

Achieving consistency

For optimal control, administrators should connect, name, and maintain servers in a physically and logically consistent manner across all data centers. Software versions, databases, naming conventions, and cabling should all be consistent. Consistency enables administrators to perform the same fixes, changes, and upgrades the same way at every site, resulting in higher productivity, smoother transitions during upgrades, more reliable troubleshooting, and quicker recovery from unplanned downtime. The 2161DS switch facilitates planning for current use and future expansion by helping administrators achieve consistency from site to site—physically through the expanders and cabling of the connected servers and logically through the switch software.

Physical organization. The switch hardware consumes only 1U of rack space, and the Cat 5 cables are less bulky at the rack than thicker KVM cables. The switch also facilitates physical consistency by making duplication of site setups easier, so that administrators can more easily note and recognize required site-specific modifications when the systems are monitored.

Logical organization. Switch software enables administrators to control logical organization, making expansions, upgrades, and reconfigurations more efficient for both hardware and software. By implementing standard naming conventions, administrators can establish consistency among sites and clusters and expedite troubleshooting by applying necessary corrections at all sites from a virtual centralized console.

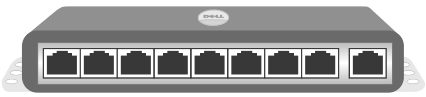


Figure 3. Dell Port Expansion Module

Realizing scalability, access, and compatibility

Physical and logical ease of organization provide the foundation for the scalability, access, and compatibility necessary to manage HPC clusters effectively. The 2161DS achieves these benefits in the following ways.

Scalability. Clusters that scale out well allow the addition of new servers as needed, and the 2161DS will connect up to 128 servers using PEMs. Beyond that, each port on the 2161DS can connect to an existing analog Dell switch that is attached to as many as 16 servers, enabling administrators to scale out a single 2161DS switch to connect a total of 256 servers.

Clusters are highly dynamic environments; moving or adding servers with traditional KVM switches requires administrators to use KVM menus and change the names and port locations of each affected server. With the 2161DS, the name of the server is stored in the SIP, so whenever a server is moved or a rack is re-cabled, the name of that server will automatically follow it to the new switch or port location.

Access. The 2161DS provides virtual centralized control by enabling multiple simultaneous sessions from local and remote locations, thereby helping administrators to monitor and support each site in a timely and productive manner.

Compatibility. A switch must be compatible with the software and hardware of all connected servers to facilitate smooth operation while changes are implemented across heterogeneous, multiplatform data centers. Because the 2161DS is not restricted to specific hardware or software, administrators have universal access to all components connected to the switch, thereby providing a stable base for the HPC cluster. No special software or drivers (such as Microsoft® Terminal Services or Symantec® pcAnywhere®) are needed on a server or host for remote connectivity.

Because the 2161DS is not restricted to specific hardware or software, administrators have universal access to all components connected to the switch, thereby providing a stable base for the HPC cluster.

Lowering costs

Company growth can be costly for IT organizations. As more people, processes, and data exact heavier loads on the HPC cluster environment, more servers are required, along with the supporting hardware and software. By scaling out to connect as many as 128 servers over a single IP connection, the 2161DS switch enables administrators to attach new servers for a fraction of what it would cost using an analog switching device. In addition, the 2161DS uses Cat 5 cables, which are less expensive than thicker KVM cables.

The remote access function in the 2161DS allows more productive use of IT staff. Even when located a continent away, an expert can log in remotely to assess the problem and make corrections—or more effectively direct the local efforts—thereby improving troubleshooting and time to recovery.

The 2161DS switch facilitates planning for current use and future expansion by helping administrators achieve consistency from site to site.

Managing HPC clusters effectively

As enterprise demands for computing power escalate, IT organizations must respond quickly with the ability to connect, organize, monitor, and troubleshoot servers in real time, wherever the administrators or operations centers are located. Data center managers achieve productivity goals through consistent configuration, effective monitoring capabilities, and simple, low-cost connections. The Dell 2161DS switch enables secure, effective administration and control of HPC clusters.

Stephen M. Hahn (steve.hahn@avocent.com) has been the Dell worldwide account manager at Avocent for the past three years. Previously, Steve was a technical sales engineer at Avocent. A Huntsville, Alabama–based company, Avocent helps data center operators manage their ever-expanding server farms.

Ed Krach (ed.krach@avocent.com) is manager of OEM Systems Engineering for Avocent. He serves as a technical consultant for Avocent original equipment manufacturer (OEM) projects, leads technical and sales training, and takes an active role in presenting the full range of Avocent connectivity products for data center management.

FOR MORE INFORMATION

Avocent:
http://www.avocent.com

Dell 2161DS:
http://www.dell.com/us/en/biz/products/model_svrac_1_svrac_console.htm

Guidelines for Building a High-Performance Computing Cluster

This article provides guidelines for planning an efficient, cost-effective HPC cluster based on Dell™ PowerEdge™ servers, including recommendations for management and Message Passing Interface (MPI) software.

BY HÅKON BUGGE

High-performance computing (HPC) clusters have become increasingly popular in a wide range of industries. IT organizations considering implementing a Dell™ server–based HPC cluster can benefit from understanding the steps involved in this task. This article describes these steps and explains how to determine the appropriate cluster configuration and management software to help maximize performance and lower total cost of ownership (TCO).

Considering physical requirements for building an HPC cluster

The first step in building an HPC cluster—before choosing the hardware and software—is planning the physical environment. An HPC environment must support the cluster, while leaving room to scale out as necessary. The space required varies with the number of processors. For example, rack-mounted systems can house a 128-CPU cluster in approximately the same space as that required for a large UNIX®-based server that typically accommodates 32 processors. Weight is also important; for example, a 500-node cluster requires a reinforced floor. Other considerations include cooling requirements, power sources, and backup.

Choosing the hardware architecture

The appropriate processor architecture for a cluster depends on performance requirements and budget. Traditional clustering software runs only on a homogeneous architecture. However, some communication and management middleware—such as Scali Manage™ and Scali MPI Connect™ software—can accommodate hardware heterogeneity. Using more flexible software allows administrators to design clusters with a mix of node architectures that best meet requirements.

Heterogeneity permits greater design flexibility and scalability. IT organizations can also realize cost savings from the ability to use legacy technology. A heterogeneous cluster can be designed to handle various tasks with optimized performance, such as pre- and post-processing. Additionally, administrators can replace failed nodes with the latest technology, yielding more power at or below the original cost. The ability to have heterogeneous architectures in a single cluster also allows joining two separate clusters into a single large unit for increased processing power.

To select an appropriate node architecture, consider the number of units required and how fast a single node should be able to execute its processing tasks. Choosing 64-bit rather than 32-bit architecture provides more processing power. This decision should be made considering the applications that will run on the cluster and the cluster's performance requirements and scale-out characteristics. For example, as administrators scale out an application, adding 32-bit servers may provide increasingly less incremental processing, so the more powerful 64-bit architecture could be a better choice.

Dell offers several core 32- and 64-bit HPC platforms. For more details and server configuration information, visit http://www.dell.com/hpcc.

Factory configuration streamlines cluster installation

Configuration is another hardware-related consideration. For cluster deployment, the boot order often must be changed to instruct the node to boot from the network rather than a CD or disk drive. Administrators can establish the BIOS setting in firmware or they can order a custom configuration from Dell to save setup time.

During system configuration, Dell also can set the BIOS option for enabling or disabling Intel Hyper-Threading Technology in the Dell PowerEdge™ 1750 and PowerEdge 650 to better meet application requirements. In Hyper-Threading, one physical CPU appears as two logical processors. Because this arrangement might require application licenses for twice the number of CPUs, and because Hyper-Threading seldom delivers a performance boost of more than about 30 percent, disabling Hyper-Threading can be more cost-effective if the application cannot take advantage of it.

Ultimately, hardware TCO will largely be determined by nontechnical issues such as service and support, so consider a vendor that provides a single point of contact for support and has proven competency in the HPC arena.

Determining which interconnect to use

Nodes use interconnects to function as a cluster. Available products range from legacy interconnects such as Gigabit Ethernet[1] to the more advanced interconnects such as Myricom® Myrinet®. Gigabit Ethernet is less expensive, but Myrinet offers higher performance, lower latency, and better scalability. Organizations often run multiple applications on a single cluster, making the choice of interconnects a complex one without a single right answer.

For example, a single-processor PowerEdge 650 is suitable for Gigabit Ethernet. A dual-processor PowerEdge 1750 offers more computational power, suggesting the choice of a faster interconnect. A cluster of PowerEdge 3250 servers may work best with a fast, low-latency interconnect to leverage the computing power of the Intel® Itanium® 2 processor and prevent the communication channel from becoming a bottleneck.

Another factor influencing interconnects is the Message Passing Interface (MPI) software (see "Resolving cluster communication software issues" later in this article). Scali MPI Connect can simplify cluster interconnect decisions because it handles both legacy and advanced protocols and standards: TCP/IP, Gigabit Ethernet with Direct Ethernet Transport (DET), Myrinet, InfiniBand™, and Scalable Coherent Interface (SCI).

Selecting the operating system

Open source software is increasingly available with professional support and maintenance. Therefore, more companies are using commercial versions of the Linux® operating system, such as Red Hat® Linux, which ships with Dell HPC hardware. Choosing a

SOFTWARE SIMPLIFIES MANAGING AND RUNNING DELL SERVER–BASED CLUSTERS

Dell provides turnkey HPC cluster configurations that incorporate powerful cluster management and interconnect software—such as products from Scali, a provider of high-performance clustering software.

Scali Manage. This software includes comprehensive tools for cluster installation, configuration, management, and monitoring, and supports the leading cluster interconnects and platforms. Scali Manage includes the following features:

- Central system for managing cross-enterprise cluster resources
- Platform-independent cluster management
- Advanced productivity tools for effective maintenance
- Ease and speed of cluster installation, configuration, and expansion
- Ability to manage both small and large systems
- Secure architecture

Scali MPI Connect. This software offers an interconnect-independent, integrated architecture that allows interaction with a single MPI implementation. Third-party applications need to be compiled only once to run on the leading interconnects. Binary programs linked with Scali MPI Connect can run on any of the supported interconnects without recompilation or relinking. Other features include:

- High-bandwidth, low-latency performance
- Advanced application debugging, tuning, and optimization
- Dynamic binding to operator-selected interconnects
- High reliability and system scalability

[1] This term indicates compliance with IEEE® standard 802.3ab for Gigabit Ethernet, and does not connote actual operating speed of 1 Gbps. For high-speed transmission, connection to a Gigabit Ethernet server and network infrastructure is required.

professionally developed and maintained operating system rather than freeware can help keep TCO low by sparing system administrators time-consuming manual maintenance and troubleshooting tasks.

Resolving cluster communication software issues

The message-passing middleware layer encapsulates the complexity of the underlying communication mechanism and shields the application from different methods of basic communication. Today, MPI has become the de facto standard for message passing. Although MPI is commonly used for parallel applications, developers face a significant challenge: virtually every brand of interconnect requires its own particular implementation of the MPI standard.

Furthermore, most applications are statically linked to the MPI library, which can create the following problems:

- **Conflicts:** If two applications run on the cluster and different versions of MPI link the applications, a conflict might occur. This inconsistency is solved by having one of the application vendors relink, test, and qualify its application for the other MPI version—a procedure that may be significantly time-consuming.
- **Disconnects:** Evolving demands from applications or errors detected and corrected in the MPI implementation can force one of the applications to use a newer version. In this case, a disconnect between MPI versions and application versions occurs, which again requires an application vendor to relink, test, and qualify its application for the other MPI version.
- **Inflexibility:** Administrators may wish to change MPIs. For example, with a Gigabit Ethernet interconnect, the TCP/IP stack may impose overhead that restricts application scalability. To switch to an MPI that can use leaner, more efficient protocols—such as Remote Direct Memory Access (RDMA)—again requires approaching the application vendors and asking for help with an upgrade or evaluation. This hurdle can deter IT organizations from implementing major improvements afforded by newer, more innovative communications software or interconnect hardware.

EXPLORING THE UNIVERSE WITH HPC

High-energy particle physics involves asking fundamental questions about the universe, such as how particles acquire their mass and why the amounts of matter and antimatter are not the same. Such questions help scientists explore the structure of the universe and how matter might be changed into different forms.

To speed up UK-based research on data from the Fermilab accelerator in the United States, a consortium—Oxford University, Liverpool University, Glasgow University, and University College London—chose a combination of Dell and Scali products that will help enable scientists to simplify complex analysis by providing the power of a traditional proprietary supercomputer at a fraction of the implementation and management cost.

The consortium chose an HPC cluster architecture consisting of Intel Xeon™ processor–based Dell servers running Scali MPI Connect and Scali Manage software. The clusters enable UK-based researchers to conduct multiple high-speed data analyses from the United States, requiring minimal systems maintenance and support. The HPC clusters of 10 compute nodes each offer 7.1 TB of usable disk capacity for simultaneous multiple analyses or for parallel processing of single large-scale analyses.

The HPC cluster provides the consortium with processing power to churn and analyze large amounts of data at an improved price/performance compared to closed-architecture supercomputers. Other key benefits include compatibility with open standards, flexibility, ease of upgrade, and scalability.

Administrators can avoid problems caused by static links to the MPI library by using software that offers dynamic binding between the application and the MPI middleware, and between the MPI middleware and device drivers for various types of interconnects. Offerings such as Scali MPI Connect can help enable administrators to develop the MPI implementation and the application independently of one another, because the application can take advantage of different interconnects or protocols without being changed or relinked.

Managing systems using the Scali Manage software package

Management software is the final layer in the cluster stack. IT organizations can benefit from a full-featured management package such as Scali Manage, which includes comprehensive tools for systems installation, configuration, management, and monitoring. In addition, compatibility with leading cluster interconnects and platforms can enable a single management application to support the broad range of systems architectures, nodes, and interconnects that may be dispersed throughout the enterprise. Management software that is designed to be centralized across clusters can help administrators create an integrated computing environment that can reduce costs, increase efficiencies, and improve overall price/performance.

Scali Manage facilitates Dell HPC cluster administration

For administrators managing Dell server–based HPC clusters, Scali Manage scales to support both small and large systems. Scali Manage streamlines the installation process with rapid installation of cluster

IMPROVING OIL EXPLORATION USING HPC CLUSTERS

Dell and Scali worked together to install a 910-node HPC cluster for Saudi Aramco—the world's leading oil producer and exporter and a top producer of natural gas—to help accelerate the company's oil and gas exploration program.

The supercomputing cluster comprises 910 Dell PowerEdge servers, adding 1,820 Intel Pentium® III processors to the Saudi Aramco compute engine. Dell implemented Scali Manage to help ensure that all 910 servers could be managed from a central point with maximum efficiency and reliability. Processing the seismic data allows geophysicists and geologists to more accurately assess where new reserves of oil and gas may be situated. The cluster is a critical tool for exploration.

A large oil strike can require up to 30 crews on standby at a high daily cost. Implementing the Dell cluster resulted in a positive oil strike in only two weeks, compared to four weeks using the legacy proprietary system—providing an immediate return on investment (ROI) for Saudi Aramco.

Using standards-based systems from suppliers such as Dell enabled Saudi Aramco to build its fastest, least expensive Prestack Time Migration (PSTM) system. The ability to rapidly and inexpensively add computing power to its cluster is one major benefit that Saudi Aramco derived from the Intel-based platform.

Dell deployed and commissioned the supercomputing cluster with the help of Dell Saudi Arabian distributor Al-Alamiah and software partner Scali. The Scali Manage architecture simplified integration of the Dell HPC cluster with the company's existing computing infrastructure, providing Saudi Aramco with an increase in computing capacity.

The Dell Applications Solutions Centre (ASC) in Limerick, Ireland, helped provide proof of concept on Saudi Aramco's initial 32-node installation.

software, including the operating system and middleware. It offers controlled installation of third-party applications and reinstallation of nodes—for example, when a node has been replaced.

Scali Manage works with the Dell Embedded Remote Access option (ERA/O), which enables remote management of critical servers. With ERA installed on the compute nodes, Scali Manage allows administrators to perform the following functions:

- Power on, power off, or toggle power on single or aggregated compute nodes
- Monitor and connect to the console ports of the nodes individually or collectively
- Replicate commands to several consoles by using a special broadcast window, avoiding the time-consuming and error-prone task of typing the same command several times

Scali Manage is also integrated with the Dell OpenManage™ software suite, which allows administrators to manage, monitor, and control the health of Dell PowerEdge servers from a central or a remote location. Incorporating information from Dell OpenManage into Scali Manage provides both a detailed view of the individual server nodes and a global view of the health and performance of the cluster as a whole.

Operationally, Scali Manage streamlines ongoing activities by providing a single management system for mixed nodes and interconnects; it is a single point of management for one or many clusters. Its easy-to-use, flexible interface enables advanced network and user administration as well as preventive management such as hardware monitoring for early fault detection. Finally, Scali Manage offers out-of-band management and disaster recovery.

Simplifying cluster management and achieving lower TCO

Computing clusters built with products from Dell and Scali have shown real-world success in compute-intensive environments. HPC clusters provide a university consortium with the processing power to analyze large amounts of data cost-effectively, and enable an oil exploration company to achieve an oil strike in half the usual time. For details, see "Exploring the universe with HPC" and "Improving oil exploration using HPC clusters."

Software components such as message-passing interfaces and management systems are important considerations when building an HPC cluster from the ground up. For today's IT environments, software like that from Scali simplifies cluster management and operation and helps save time, effort, and money—making administrators more productive and helping to achieve business goals by lowering TCO.

Håkon Bugge (hakon.bugge@scali.com) is a founder and vice president of product development at Scali (http://www.scali.com), a provider of management software and MPI middleware for high-performance clustering. Håkon has focused on developing clustering software, interconnect technology, and advanced CPU architectures. He has an M.S. from the University of Oslo, Norway, and has also served as a part-time lecturer in computer architecture at the University Graduate Center (UniK), Kjeller, Norway.

Networking Considerations for Clustered Servers

Reliable, low-cost switches can enhance the network performance of Intel® processor–based Dell™ PowerEdge™ servers in high-availability failover cluster and high-performance computing (HPC) cluster environments. This article provides guidelines that help IT managers choose the most effective network switches for their cluster interconnects.

BY EXTREME NETWORKS

Given tight IT budgets, data center managers around the world are rapidly implementing low-cost, high-performance clusters of standard computing and networking components to process compute-intensive tasks. By deploying clusters of relatively inexpensive Intel® processor–based Dell™ PowerEdge™ servers to execute mathematically complex models and simulations, enterprises can eliminate expensive symmetric multiprocessing (SMP) systems in all but the most complex applications.

A cluster can be defined loosely as any two or more servers working together in a coordinated fashion to deliver a common service or set of services. Although specific cluster computing and software elements may vary, the network is one critical component all clusters share. This article presents networking considerations that pertain to high-availability failover clusters and scalable high-performance computing (HPC) clusters. Scalable clusters are further categorized to address loosely coupled HPC applications, which use distributed memory, and tightly coupled HPC applications, which use shared memory.

High-availability clusters: Providing failover protection

Considering the demands of Internet-based applications, IT organizations must keep the enterprise infrastructure constantly working at a consistent, responsive level of performance, especially during reasonable peaks in user load. High-availability clustering was once the exclusive domain of mainframe and minicomputer systems. Now high-availability clusters are an option for most mission-critical business applications running in Microsoft® Windows® and Linux® operating system environments—including e-mail and messaging, databases, customer relationship management (CRM), enterprise resource planning (ERP), and shared file systems.

High-availability failover clusters are designed to improve system reliability and manageability. IT organizations deploy failover clusters to keep end-user applications operational during both planned system events (such as backups, maintenance, and peak user loads) and unplanned events (such as hardware failures and network outages). As part of the fault-tolerant design, each critical system component sends a short message, or *heartbeat*, over the network at regular intervals to indicate that it is still online.

Should a computing or networking component in the cluster fail to send a heartbeat within the predetermined time, the operating system will direct other servers in the cluster to perform tests that evaluate the health of the system in question. If a critical component has failed, cluster members will initiate emergency actions to assume the load and resources of the failed node. System administrators typically define the specific policy to be implemented if emergency failover action becomes necessary in the operating system software that manages the clustered applications.

Administrators can base failover policies on one of two system readiness levels:

- **Hot/Warm:** All nodes in the cluster are either in active service or in standby mode. Nodes in standby mode have access to all resources necessary to assume the load of a failed system with no perceptible downtime in end-user applications.
- **Hot/Cold:** Backup systems require some startup time before assuming the load of the failed system. The hot/cold failover policy potentially can lead to an interruption in end-user service.

High-availability failover clusters are generally deployed in server pairs (see Figure 1). For many applications, system administrators can establish straightforward networking requirements that leverage standard TCP/IP transport protocols, basic network interface cards (NICs), and enterprise-grade network switching equipment. Usually administrators connect each server to two different Ethernet switches. This *dual-home* configuration helps ensure that the failure of one Ethernet switch or Ethernet link will not cause server failover. The dual-home configuration also helps ensure that the failure of one server will not eliminate failover protection for the high-availability configuration. However, should a switch, link, or server fail, only one Ethernet link to each server would be active at a time.

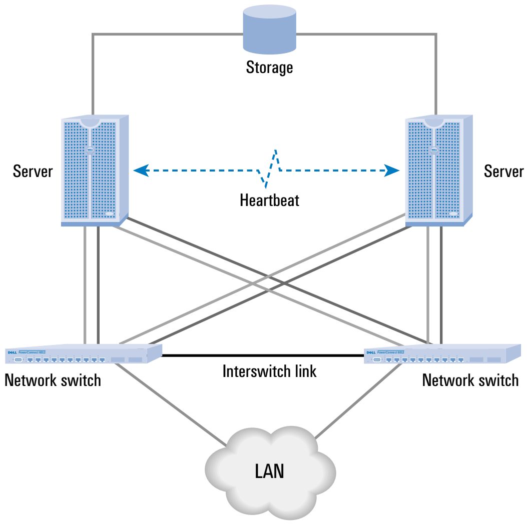


Figure 1. High-availability failover cluster architecture

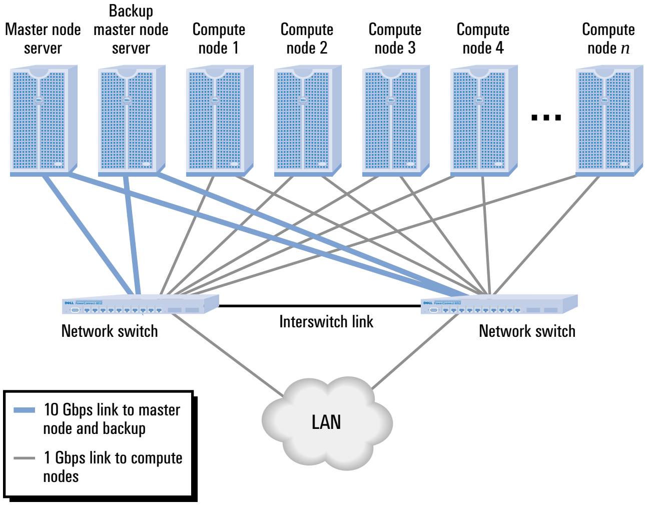


Figure 2. Loosely coupled HPC cluster architecture

In addition, multiple Ethernet links can be connected between each Ethernet switch and each server. Using the link aggregation technique, administrators can configure separate physical data channels into a logical channel that performs as one higher bandwidth link. Link aggregation enables administrators to provide another level of failure protection by balancing the load to and from each server across multiple Ethernet connections. In this scenario, the server will retain a network connection even if one link fails, without requiring administrators to reconfigure the network. Link aggregation also increases the total throughput of the active server.

Loosely coupled HPC clusters: Performing massively parallel computing tasks

IT administrators can configure stand-alone servers, desktop PCs, and rack-mounted blade server nodes in scalable, loosely coupled HPC clusters to process common, compute-intensive applications. Loosely coupled applications use distributed memory and require minimal communication among cluster nodes; generally, synchronization among nodes is not performance-critical. Any large computing task that can be divided into clean, self-contained segments—that is, a massively parallel computing task—works well in a loosely coupled HPC computing cluster (see Figure 2). Examples of loosely coupled applications running in Windows- and Linux-based environments include chip design simulation, financial modeling, scientific research, and computer graphics.

Whenever possible, developers design applications to take advantage of loosely coupled processing. For example, each frame of an animated movie can be represented by a 3-D model that includes colors, textures, and lighting descriptions. A *master node* server, which manages communication and job scheduling in the HPC cluster, distributes each frame to an individual *compute node* for processing. When the processing is completed (or in 3-D terms, "rendered"), the compute node returns the completed frame to the master node and awaits further instructions. Because massively parallel computing tasks can be executed across individual compute nodes with little or no interprocess communication, the number of cluster members that can help process distributed tasks is virtually unlimited. As a result, HPC clusters rank among the most powerful computers in the world, offering combined processing power measured in hundreds of gigaflops.[1] For examples of such clusters, visit the TOP500 list of the world's most powerful computer systems (http://www.top500.org).

Well-engineered Ethernet switches and NICs can handle most communication tasks in loosely coupled clusters. For large data sets such as those produced by 3-D computer graphics processing, administrators can deploy Gigabit Ethernet[2] and 10 Gigabit Ethernet switches. A 10 Gigabit Ethernet switch is usually configured with the master node, where aggregate responses from all the compute nodes could create a system bottleneck. To ease the load on servers during the transfer of large data sets, administrators should consider using Jumbo Ethernet frames. This option improves throughput and network efficiency by increasing the maximum transmission unit (MTU) of the switch to 9018 bytes per packet, compared to 1518 bytes in 10/100/1000 Mbps Ethernet frames.

As administrators scale out the cluster configuration, additional servers increase the demand on Ethernet switch ports. Network switches must offer wire-speed connectivity without blocking traffic at higher aggregate throughputs. Also, clusters require high-speed trunk connections between switches to ensure that performance does not degrade if traffic passes through multiple switches. To handle the aggregate loads, high-speed trunks require higher capacity connections than the server connections. For example, if servers are connected to Fast Ethernet, the trunks should be Gigabit Ethernet; if servers are connected to Gigabit Ethernet, the trunks should be 10 Gigabit Ethernet. Because this arrangement can be expensive, IT organizations should identify network switches that provide the best possible price/performance.

Web content delivery is particularly well suited to loosely coupled HPC clusters. For Web servers, the equivalent of the master node is a server load-balancing (SLB) network switch (see Figure 3). Coordinated by protocol-aware intelligence in the SLB switch, clustered Web servers process application-level requests from the SLB switch independently of one another, even when serving the same content.

Web content delivery is particularly well suited to loosely coupled HPC clusters. For Web servers, the equivalent of the master node is a server load-balancing (SLB) network switch.

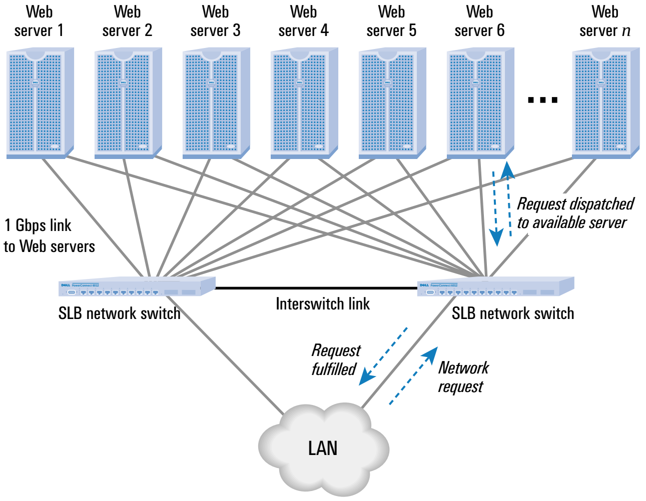


Figure 3. Loosely coupled SLB cluster architecture

Other applications that exhibit characteristics similar to Web content delivery and can be similarly load balanced include Domain Name System (DNS) servers, Simple Mail Transport Protocol (SMTP) servers, firewalls, intrusion detection devices, virtual private network (VPN) devices, and Wireless Application Protocol (WAP) gateways. New environments that can benefit from server load balancing are being discovered regularly.

Generally, SLB-clustered networks do not place great demands on individual cluster members or their network interfaces. However, deploying an additional layer of devices to accomplish the SLB function adds a layer of complexity to the network topology; switches that support SLB configurations enable administrators to manage and scale the network more easily. In particular, SLB-capable switches must be able to perform the following functions:

- Route addresses at both Layer 2 (physical hardware) and Layer 3 (network IP) of the Open Systems Interconnection (OSI) reference model
- Support higher-level network protocols
- Support best-practice SLB algorithms for distributing loads, such as TCP, FTP, HTTP, Secure Sockets Layer (SSL), SMTP, DNS, and many others

[1] One gigaflop equals 1 billion floating-point operations per second.

[2] This term indicates compliance with IEEE® standard 802.3ab for Gigabit Ethernet, and does not connote actual operating speed of 1 Gbps. For high-speed transmission, connection to a Gigabit Ethernet server and network infrastructure is required.

NETWORK CHECKLIST: CONFIGURING HPC CLUSTERS

IT organizations planning to purchase network switches for HPC cluster environments should look for the following features:

- **Non-blocking switch fabric:** The Ethernet switch should not slow down or block packets on their way through the network.
- **Jumbo frame support:** Increasing the maximum transmission unit (MTU) from 1518 to 9018 bytes improves network throughput and efficiency when transferring large data sets.
- **Modular scalability:** Network switches should expand easily as clusters grow, either by allowing room for additional internal port blades or by providing stackable external expansion modules.
- **Quality of Service:** QoS provisions enable administrators to prioritize delivery of the most time-critical communications within the cluster.
- **Link aggregation:** Using link aggregation techniques, administrators can increase network bandwidth and balance the processing load across multiple Ethernet connections.
- **Fast failover:** Support for hot/warm failover provides ready access to all the resources that cluster members need to assume the load of a failed node with no perceptible interruption in service.
- **10 Gigabit Ethernet support:** Configured with the master node, a 10 Gigabit Ethernet switch can help prevent system bottlenecks caused by aggregate responses from all the compute nodes.

- Assess server health and load by monitoring critical parameters such as response time, CPU utilization, and available memory

Tightly coupled HPC clusters: Processing complex mathematical models

Some applications—including advanced weather simulations, econometric modeling, and other types of highly complex mathematical programs—are difficult or impossible to decompose, and therefore must remain tightly coupled, requiring shared memory. The overall performance of such applications is highly sensitive to the latency and throughput of communications among the compute nodes. In the past, such applications were deployed either on expensive, proprietary SMP platforms or on lower-cost Intel processor–based compute nodes interconnected using expensive, proprietary networking equipment. Now, by using 10 Gbps Ethernet switches and network switch–enabled Quality of Service (QoS) techniques, administrators can eliminate proprietary network interconnects.

Key QoS provisions for tightly coupled HPC networks perform the following functions:

- Discriminate between communications that have critical latency requirements and less time-critical traffic
- Implement policies to control the granularity of network latency and throughput for critical traffic

For instance, the "sideways" traffic among CPUs in a tightly coupled system, often represented by remote procedure calls (RPCs), can be prioritized over management traffic, external requests, and bulk transfers of administrative reporting data. Emerging network technology and standards will help enterprises to build tightly coupled HPC environments as economically as they can build loosely coupled HPC clusters today. Soon, standards such as Remote Direct Memory Access (RDMA), combined with Gigabit Ethernet and 10 Gigabit Ethernet switches and sophisticated QoS features, will allow administrators to configure inexpensive compute nodes for most tightly coupled HPC cluster applications, thereby avoiding the high cost of proprietary networking gear.

Improved price/performance: Lowering the cost of enterprise computing

Intel processor–based cluster components can dramatically reduce the costs of deploying both mission-critical business applications and compute-intensive scientific models and simulations. Network switches, essential for high-availability failover clusters and HPC clusters, can help administrators achieve faster throughput, greater reliability, and better scalability.

In concert with Dell PowerEdge servers, Extreme Networks® scalable Gigabit Ethernet switches can help provide the reliability, features, and price/performance that enterprise environments require—whether deployed in a cluster of high-performance computing servers, load-balancing Web servers, or high-availability applications servers.

Extreme Networks (http://www.extremenetworks.com) delivers effective applications and services infrastructures by creating networks that are faster, simpler, and more cost-effective. Headquartered in Santa Clara, California, Extreme Networks offers its network switching equipment in more than 50 countries.

FOR MORE INFORMATION

Dell PowerEdge servers: http://www.dell.com/poweredge

Extreme Networks: http://www.extremenetworks.com

Network Installation of IA-64 Nodes in an HPC Cluster

Remote network booting and installation reduce administrative overhead for high-performance computing (HPC) clusters. Network installations are based on the Preboot Execution Environment (PXE) protocol. Dell™ PowerEdge™ 3250 servers are built with IA-64 architecture, which uses the Intel® eXtensible Firmware Interface (EFI) specification. In these servers, PXE is implemented as a core EFI function, which helps improve the network installation process.

BY JASMINA JANCIC AND AMINA SAIFY

Network booting is a method of booting or installing servers in which one or more of the programs loaded during a boot sequence are obtained across the network from a remote server. Network booting is not a new concept. The process, which is based on the Preboot Execution Environment (PXE) protocol, originated in the mid-1980s but was largely limited to booting and configuring diskless clients—machines that do not store any programs on a local hard drive, but download all components, including the operating system (OS), from the remote server. However, the network-booting concept can also be applied to remote network installations. Administrators can remotely install (or reinstall) and configure a machine just by plugging it into a network and booting through a network card, without any user interaction or prior OS installation.

Remote, automated installation methods are indispensable in high-performance computing (HPC) clusters. HPC clusters utilize standards-based hardware components to build parallel computational systems that execute parallel and distributed code. The application code runs across multiple compute nodes, which provide computational horsepower. Depending on the application requirements, an HPC cluster may consist of several thousand compute nodes.

Network installation offers advantages in HPC cluster environments

Manually installing a node in a cluster requires direct access to the node through a keyboard and monitor attached to the node. If the installation is performed from a CD, the administrator must either type in all configuration information or enter the location of the configuration file (also known as a kickstart file), which specifies installation choices. An entirely manual installation from a CD can take up to 30 minutes, so a manual installation of a large cluster could take hours to complete. Conversely, administrators may complete network installation of Dell™ clusters in minutes. Remote, automated installation can greatly reduce administrative overhead for HPC clusters by decreasing installation time and effort, and minimizing the chance of human error.

Remote, automated installation methods are indispensable in HPC clusters.

Remote network installations are also highly useful for HPC cluster maintenance. Because HPC cluster applications use compute nodes that are essentially replicas of one another, each node does not store data locally and individual nodes are not customized. Therefore, network installation enables administrators to easily rebuild failed nodes; the failed node can be reimaged over the network by reinstalling an exact copy of the OS that was running before the critical failure. Using network installation in an HPC cluster environment also helps simplify upgrades. Upgrade packages and configurations need be applied only to the central server, thus enabling administrators to maintain just one copy of the configuration files and images.

Remote, automated installation can greatly reduce administrative overhead for HPC clusters by decreasing installation time and effort, and minimizing the chance of human error.

eXtensible Firmware Interface improves network installation for clusters with IA-64 architecture

Recently, Dell introduced the Dell PowerEdge™ 3250, a high-performance server that supports up to two Intel® Itanium® 2 processors, which use IA-64 Instruction Set Architecture (ISA). The PowerEdge 3250 is an excellent compute node for an HPC cluster, because its IA-64 architecture overcomes some limitations of IA-32 architecture to accommodate applications that have large memory footprints.

Dell offers several turnkey HPC cluster solutions that bundle the PowerEdge 3250 server with software and support for clusters of 8, 16, 32, 64, and 128 compute nodes. The cluster uses eXtensible Firmware Interface (EFI) network installation principles, so clients must be IA-64 although the installation server can be either IA-32 or IA-64. EFI, a new specification by Intel, defines the interface between the OS and platform firmware, providing a standard environment for booting an OS and running preboot applications. In practice, the interface serves as a pre-OS execution environment.

Traditionally, the BIOS requires an OS loader to have knowledge about the workings of some hardware components in the server. Because EFI specifies interfaces to platform capabilities, it relieves the OS loader of the need for extensive knowledge about the hardware components. EFI defines a building block between the OS loader and the firmware so that both the OS loader and firmware can be developed independently of each other, provided that they limit their interactions to the EFI interface (see Figure 1).

Also, EFI does not require a network interface card (NIC) to be PXE-enabled. The client side of the PXE protocol implements a bootstrap loader, necessary to perform a PXE boot. For a non-EFI network boot, the loader is implemented in the NIC and BIOS of the server, tying this PXE functionality to the NIC vendor. In such a scenario, administrators must ensure that a NIC is PXE-enabled if they want to add a new NIC to a system and perform a PXE boot from it.

The bootstrap loader required for network boot and installation is now built into EFI, where it is implemented as a core protocol called PXE 32/64 BaseCode. This protocol is common to all network-interface hardware, so vendors are no longer responsible for implementing this portion of PXE functionality.[1] The choice of a NIC card is no longer based on PXE compatibility, which potentially reduces the cost of the NICs, improves compatibility, and simplifies administration.

PXE functionality enables remote booting and installation

At the core of remote booting and installation is PXE functionality, which IA-64 platforms implement in EFI. PXE is defined on a foundation of industry-standard Internet protocols and services: namely TCP/IP, Dynamic Host Configuration Protocol (DHCP), and Trivial FTP (TFTP). Along with the DHCP standards, these specifications standardize the interactions between clients and servers. PXE is an extended protocol that works without disrupting the operations of a standard DHCP server. Clients and servers that are aware of PXE extensions recognize and use this information, and those that do not recognize the extensions ignore them.[2]

Remote booting and installation requirements and procedure

The Intel EFI-enabled PXE boot sequence divides network booting into a bootstrap loader, which is responsible for locating and obtaining the bootstrap program, and a network bootstrap program (NBP). The NBP locates, downloads, and executes the OS.

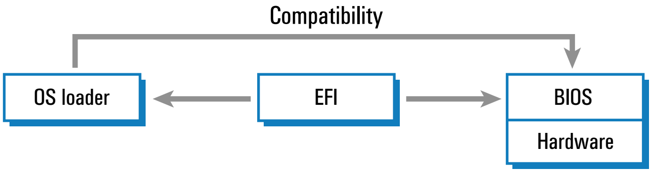


Figure 1. EFI overview

[1] For more information, see "PXE Manageability Technology for EFI," *Intel Developer Update Magazine*, http://intel.com/update/departments/initech/it10004.pdf.

[2] *Preboot Execution Environment (PXE) Specification Version 2.1*, Intel Corporation, ftp://download.intel.com/labs/manage/wfm/download/pxespec.pdf.

The requirements for network installation in a Linux environment include:

- A bootstrap loader (implemented in EFI in IA-64 architecture)
- A DHCP or bootp server to return an IP address and other information when it receives a Media Access Control (MAC) address
- A TFTP server to send the kernel images and configuration files required in the boot process
- A Linux kernel
- A RAM disk contained in the loaded image
- A Network File System (NFS) or combined NFS and FTP server for providing the OS distribution to be installed

EFI network installation steps are as follows:

1. Once the client is booted, EFI broadcasts a DHCP request to the local network, requesting an IP address based on the MAC address that it receives (see Figure 2).
2. The DHCP server responds with the IP address for the client and the IP address for the boot server.
3. Using this information, the client sends a request to the boot server for the NBP file.
4. The boot server sends the NBP filename and the TFTP configuration.
5. The client sends a request to the TFTP server to download the NBP file specified in the previous step.
6. The TFTP server sends the NBP file requested by the client.
7. The NBP searches the TFTP server for a configuration file that indicates which kernel and RAM disk to download. If the kernel image is a boot kernel, the client will start installing from the network, using a kickstart file specified in this configuration file.
8. The client requests the kernel and the RAM disk specified in the configuration file.
9. Once the client receives the kernel, the kernel takes control of the client for the installation; it downloads the kickstart file from the NFS/FTP server for the installation.
10. After downloading the kickstart file, the kernel uses the kickstart file to start the OS installation.

Because EFI specifies interfaces to platform capabilities, it relieves the OS loader of the need for extensive knowledge about the hardware components.

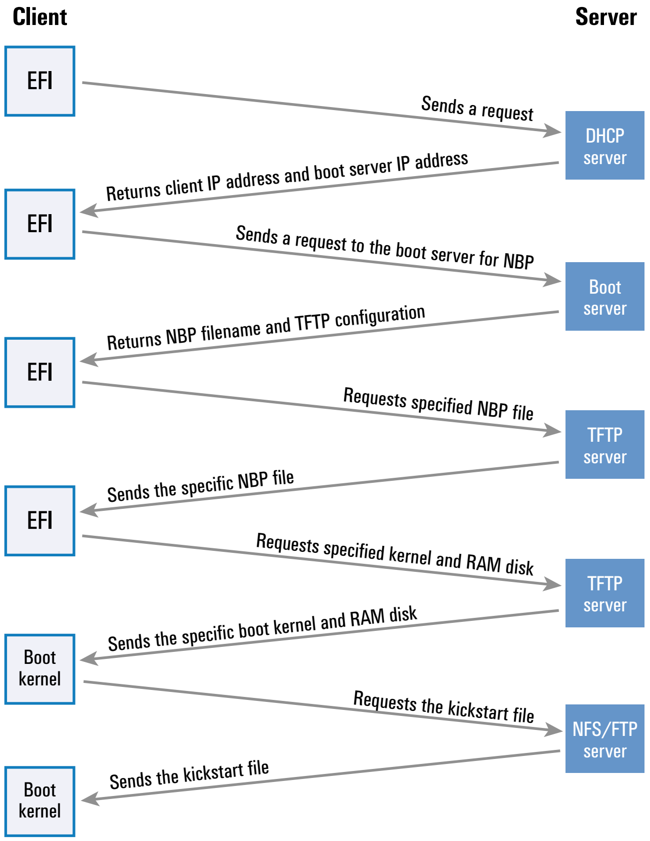

Client
Server
EFI
Sends a request
DHCP server
Returns client IP address and boot server IP address
EFI
Sends a request to the boot server for NBP
Boot server
Returns NBP filename and TFTP configuration
EFI
Requests specified NBP file
TFTP server
Sends the specific NBP file
EFI
Requests specified kernel and RAM disk
TFTP server
Sends the specific boot kernel and RAM disk
Boot kernel
Requests the kickstart file
NFS/FTP server
Sends the kickstart file
Boot kernel

Figure 2. EFI network installation steps

Remote installation can reduce TCO

Remote network booting and installation offer many important advantages for large server farms and HPC cluster installations. Network booting and installations are based on the PXE protocol.

Today, with the introduction of the Intel EFI specification in servers that use IA-64 processors, PXE can be implemented as a core EFI function. Embedding PXE functionality helps simplify network installations and reduce compatibility issues by divorcing the PXE protocol from the NIC. Servers that use IA-64 processors, such as the Dell PowerEdge 3250, can help streamline total deployment time from hours to minutes, reduce the possibility of errors during the configuration, and simplify updates and node replacements—all of which may help organizations lower total cost of ownership.

Jasmina Jancic (jasmina_jancic@dell.com) is a member of the Scalable Systems Group at Dell. She received her master's degree in Computer Science from the Georgia Institute of Technology, specializing in cluster and systems monitoring and high-performance computing.

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

An Introduction to Windows-based Clusters and the Computational Clustering Technical Preview Kit

This article provides an overview of the integrated Microsoft® Computational Clustering Technical Preview (CCTP) kit, 2003 Edition. The CCTP kit contains software packaged by Microsoft that assists in the design, implementation, configuration, and management of Microsoft Windows®–based high-performance computing (HPC) clusters.

BY MUNIRA HUSSAIN; JEFFREY MAYERSON; JENWEI HSIEH, PH.D.; AND KEVIN NOREEN

High-performance computing (HPC) clusters, particularly Beowulf-type clusters, have been widely adopted. Constructed from standards-based computing components, HPC clusters offer competitive price/performance. These clusters are typically used in universities and national laboratories for research as well as in various industries to solve technical engineering problems. As HPC cluster technology continues to mature and gain acceptance, more applications are being developed and written for parallel programming environments. For example, financial market applications include risk management, financial engineering, and stochastic analysis.

HPC clusters are operating system–agnostic and can be constructed from standards-based components determined by performance needs such as processor speed, I/O bandwidth, and memory. Decades ago, UNIX®-based systems dominated parallel computing. As a result, many organizations transitioning to HPC clusters today consider a migration from UNIX to Linux® or BSD operating systems the easiest path, involving the least amount of code porting and investment. However, HPC clusters running the Microsoft® Windows® operating system offer an alternative, innovative form of computing.

The Computational Clustering Technical Preview (CCTP) kit is a collection of software packaged by Microsoft that provides the tools and utilities needed for building Windows-based computational clusters. The CCTP kit, 2003 Edition, consists of an integrated bundle containing evaluation versions of operating systems, Microsoft programming tools, third-party software applications, and computational development libraries. In addition, the CCTP kit includes information sharing and best practices in the form of white papers. The CCTP package is periodically updated and revised to include the latest technologies.

Exploring CCTP kit software components

The CCTP kit includes the software building blocks needed to set up a Beowulf-type cluster on a Windows operating system. The package consists of cluster message-passing libraries, monitoring and management tools, compilers for code optimization, debugging tools, and parallel implementation software. Some of these components take advantage of the Microsoft .NET framework and architecture. The .NET framework is required on nodes to implement the Microsoft common language runtime (CLR).

Code can be compiled to Microsoft intermediate language (MSIL) and executed by any node running the CLR.

Operating systems. The CCTP kit comes with Windows 2000 Advanced Server and Windows XP Professional, and each operating system includes a 120-day free trial license. A Windows 2000 Advanced Server node can be used to create users and set up domain services. User and resource management is centralized to the domain controller to simplify the cluster configuration. Compute nodes also utilize Dynamic Host Configuration Protocol (DHCP) and Domain Name System (DNS) services running on the master node to help centralize network management; the DNS server provides secure dynamic updates to registered compute nodes within the cluster. Windows XP is a client version of the operating system, with stripped-down services and utilities.

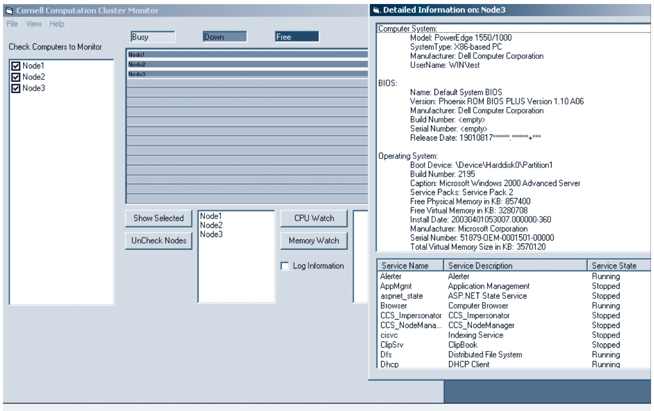
Message Passing Interface. Both commercial and open source Message Passing Interface (MPI) libraries are included in the CCTP package. The MPI Software Technology MPI/Pro® library is a commercial implementation of the MPI standard; MPICH.NT is an open source, high-performance version of the standard ported by Argonne National Laboratories. MPI/Pro depends on the .NET framework runtime environment, through which it provides remote process startup and management services. These functions offer secure login to users on compute nodes.

Compilers. The CCTP package comes with evaluation versions of compilers from Intel and Microsoft. Microsoft Visual Studio® .NET development studio is included along with Intel® Fortran and C++ compilers for Windows. The Intel compilers can be used in the Visual Studio C++ environment to provide object and source code compatibility. Both Intel and Microsoft compilers are optimized for Intel architecture and Windows operating system platforms.

Programming libraries. Intel Math Kernel Libraries (MKL) are used by developers in scientific and research fields to perform linear algebra, vector math functions, and Fast Fourier Transform (FFT) calculations. MKL are multithreaded and employ OpenMP™ threading technology to make use of all available processors. MKL also contain BLAS subroutines. The PLAPACK and LAPACK routines and libraries included in the CCTP kit are used in parallel implementations and in solving linear algebra algorithms and applications.

Debuggers and programming tools. Another added feature of the CCTP kit is the Intel VTune™ Performance Analyzer tool, which is used to probe a function, module, or application. VTune displays performance data and helps to identify bottlenecks and contingencies within software. This tool provides advanced debugging and can enhance the execution of software benchmarks and applications.

Cluster monitoring tool. The ability to manage a large number of nodes in an HPC cluster is critical. Therefore, CCTP includes the Computation Cluster Monitor (CCM) tool built by the Cornell Theory Center. It consists of a graphical user interface (GUI) and a daemon called cmond. The daemon must be installed on all compute nodes in the cluster. CCM lets administrators know the status of the nodes: free, down, or busy. It can display a graph of CPU and memory information for multiple nodes collectively, logging and storing this data as history. The tool also captures system-level information—such as BIOS, memory, hardware, and types of services—for each node (see Figure 1).



```
Computer System:
        Model: PowerEdge 1550/1000
        SystemType: X86-based PC
        Manufacturer: Dell Computer Corporation
        UserName: WIN\test

BIOS:
        Name: Default System BIOS
        Version: Phoenix ROM BIOS PLUS Version 1.10 A06
        Manufacturer: Dell Computer Corporation
        Build Number: <empty>
        Serial Number: <empty>
        Release Date: 19010817******.******+***

Operating System:
        Boot Device: \Device\Harddisk0\Partition1
        Build Number: 2195
        Caption: Microsoft Windows 2000 Advanced Server
        Service Packs: Service Pack 2
        Free Physical Memory in KB: 857400
        Free Virtual Memory in KB: 3280708
        Install Date: 20030401053007.000000-360
        Manufacturer: Microsoft Corporation
        Serial Number: 51879-OEM-0001501-00000
        Total Virtual Memory Size in KB: 3570120
```

Service Name	Service Description	Service State
Alerter	Alerter	Running
AppMgmt	Application Management	Stopped
aspnet_state	ASP.NET State Service	Stopped
Browser	Computer Browser	Running
CCS_Impersonator	CCS_Impersonator	Stopped
CCS_NodeMana...	CCS_NodeManager	Stopped
cisvc	Indexing Service	Stopped
ClipSrv	ClipBook	Stopped
Dfs	Distributed File System	Running
Dhcp	DHCP Client	Running

Figure 1. CCM GUI displaying the status of three nodes, including system configuration and services information

Along with the CCM application included in the CCTP kit, Microsoft Windows Advanced Server also includes a built-in tool called Perfmon (an abbreviation of "performance monitor"). This tool can monitor a large number of parameters—for instance, a complete range of counters for system usage and resources such as TCP/IP connections, CPU, and memory. The Perfmon tool also enables logging and sends alerts when certain thresholds are set. Because Perfmon is tied to Windows 2000 security, certain user groups and settings can be configured to limit the monitoring activity.

Job scheduler. Resource monitoring and workload management is important when multiple users share a cluster. This type of management may even be important for a single user who sends multiple computation requests to a cluster. The CCTP package contains the ClusterController™ job scheduler, a commercial application ported to Windows by MPI Software Technology. Jobs can be entered on ClusterController using either a batch file or an interactive format. Once jobs are queued, ClusterController monitors the state of the cluster, starts and stops jobs, and delivers the output.

In addition to the basic components described in this section, the CCTP kit includes various tools to enable application porting from UNIX to Windows. These tools are continually updated as more applications and benchmarks become available.

Deploying compute nodes on a Windows cluster

One of the challenges in setting up any cluster is the deployment and installation of compute nodes. Microsoft Windows has a built-in tool known as Remote Installation Services (RIS), which can be used to deploy the Windows operating system over a network to a large number of nodes simultaneously. RIS was originally available with Microsoft Windows 2000 server products for the deployment of client operating systems. The Windows Server 2003 server family includes an advanced version of RIS capable of deploying server operating systems. The distribution is copied to the RIS server and custom files can be created to taper the node installation and automate the process.

Configuring the RIS server

The RIS server can be any server running Windows 2003, but it must be connected to either a Fast Ethernet or Gigabit Ethernet[1] switch. The RIS server should be configured as a DHCP server and have DNS with Microsoft Active Directory® directory service enabled. (*Note:* The DHCP and RIS functions can be performed by separate systems or combined onto the same server. The following example assumes that they are combined.)

Using the Windows Server 2003 CD, RIS can be installed on the server as a Windows component. Executing risetup.exe on the RIS server will copy the necessary distribution files to the server and then set up services. This executable program starts copying all the necessary files to a directory specified by the user. Once this activity is complete, the RIS service should be set to Enabled.

The RIS server must be authorized in Active Directory. This authorization enables the RIS server to interact with the compute nodes on the network; if authorization is not set, the compute nodes cannot contact the RIS server.

User settings and permissions can be set to verify that a user can create accounts in the domain before deploying a compute node operating system. This function also is handled in Active Directory.

Setting up compute nodes

Each compute node should possess the minimum hardware required for the operating system deployment. Compute nodes must support the Preboot Execution Environment (PXE) mechanism or be a supported network adapter with a RIS startup disk, and PXE must be enabled in the BIOS as the first booting sequence. All compute nodes should be connected to the same network as the RIS server.

When the compute node PXE boots, it sends a DHCP request to the RIS server. After this contact is complete, an IP address is assigned by the RIS server—which, in this example, is also the DNS and DHCP server—to the compute node through the following process:

1. Compute node broadcasts a request for an IP address across the subnet
2. RIS server responds with a DHCP offer message (DHCPOFFER)
3. Client receives offer and responds
4. DHCP server sends an acknowledgment (DHCPACK)
5. Compute node sends a request for a boot server
6. RIS sends a boot server request offer
7. Compute node and RIS server complete the packet transfer process

Looking toward the future of Windows-based clustering

Windows high-performance clustering has been evolving over the past few years and will continue to develop with other technologies. Many applications and tools are available only in the Windows environment, and these may become some of the first applications and tools used for Windows-based clustering. Early adopters believe that Windows-based HPC clustering holds great potential and may capture a broad range of market segments in the future.

Acknowledgments

The authors would like to thank Greg Rankich from Microsoft for reviewing and providing feedback for this article.

Munira Hussain (munira_hussain@dell.com) is a systems engineer in the Scalable Systems Group at Dell. She has a B.S. in Electrical Engineering and a minor in Computer Science from the University of Illinois at Urbana-Champaign. Her current research interests include prototyping commercial and software solution stacks. She has also worked on productizing HPC clusters on IA-32 and IA-64 platforms.

Jeffrey Mayerson (jeffrey_mayerson@dell.com) is a systems engineer in the Scalable Systems Group at Dell. He completed his degree in Computer Information Systems and Management from the University of Wisconsin-Madison.

Jenwei Hsieh, Ph.D. (jenwei_hsieh@dell.com) is an engineering manager of the Scalable Systems Group at Dell. He has a Ph.D. in Computer Science from the University of Minnesota and a B.E. from Tamkang University in Taiwan.

Kevin Noreen (kevin_noreen@dell.com) is a product marketing strategist for clustering at Dell and has been with the Dell Enterprise Systems Group since 1998. Kevin obtained his B.A. in Management Information Systems from the University of Iowa.

FOR MORE INFORMATION

HPC clusters: http://www.microsoft.com/HPC

Microsoft Knowledge Base: Article number 325862. http://support.microsoft.com/?scid=fh;[ln];kbhowto

[1] This term indicates compliance with IEEE® standard 802.3ab for Gigabit Ethernet, and does not connote actual operating speed of 1 Gbps. For high-speed transmission, connection to a Gigabit Ethernet server and network infrastructure is required.

Optimizing the Virtual Data Center

The ideal virtual data center dynamically balances workloads across a computing cluster and redistributes hardware resources among clusters in response to changing needs. The challenge is to implement these load-balancing and resource-balancing features so they are transparent to client applications.

BY J. CRAIG LOWERY, PH.D.

Administrators have greatly improved the enterprise infrastructure by using high-availability and high-performance computing (HPC) cluster systems that provide fault tolerance and load balancing. However, these improvements in robustness generally increase costs, so data center managers are now beginning to focus on reducing those costs. Because clustered systems are sized to accommodate peak rather than average demand, data center managers have decided that underutilized nodes in lightly loaded clusters are wasteful and suspect these may be the key to improving the bottom line.

Typically, several application clusters exist in a data center. Instead of permanently sizing clusters for peak loads, managers should be able to move servers and storage among application clusters in response to the current demand for each application. Because all applications are unlikely to experience peak demand simultaneously, managers can save money by reducing the total number of units deployed, moving idle units automatically from cluster to cluster as demand dictates. If total demand exceeds available resources, applications can be prioritized to ensure that critical systems do not starve. If increased demand persists, administrators can boost capacity simply by connecting new units to the existing infrastructure. These concepts form the basis for the *virtual data center*.[1]

The statistical multiplexing of hardware across application clusters is at the core of the virtual data center concept. In the virtual data center, administrators physically configure hardware only once; software creates logical associations among hardware components as needed. For example, virtual LANs (VLANs) can be configured through software that resides on a network switch. By managing hardware as pools of

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

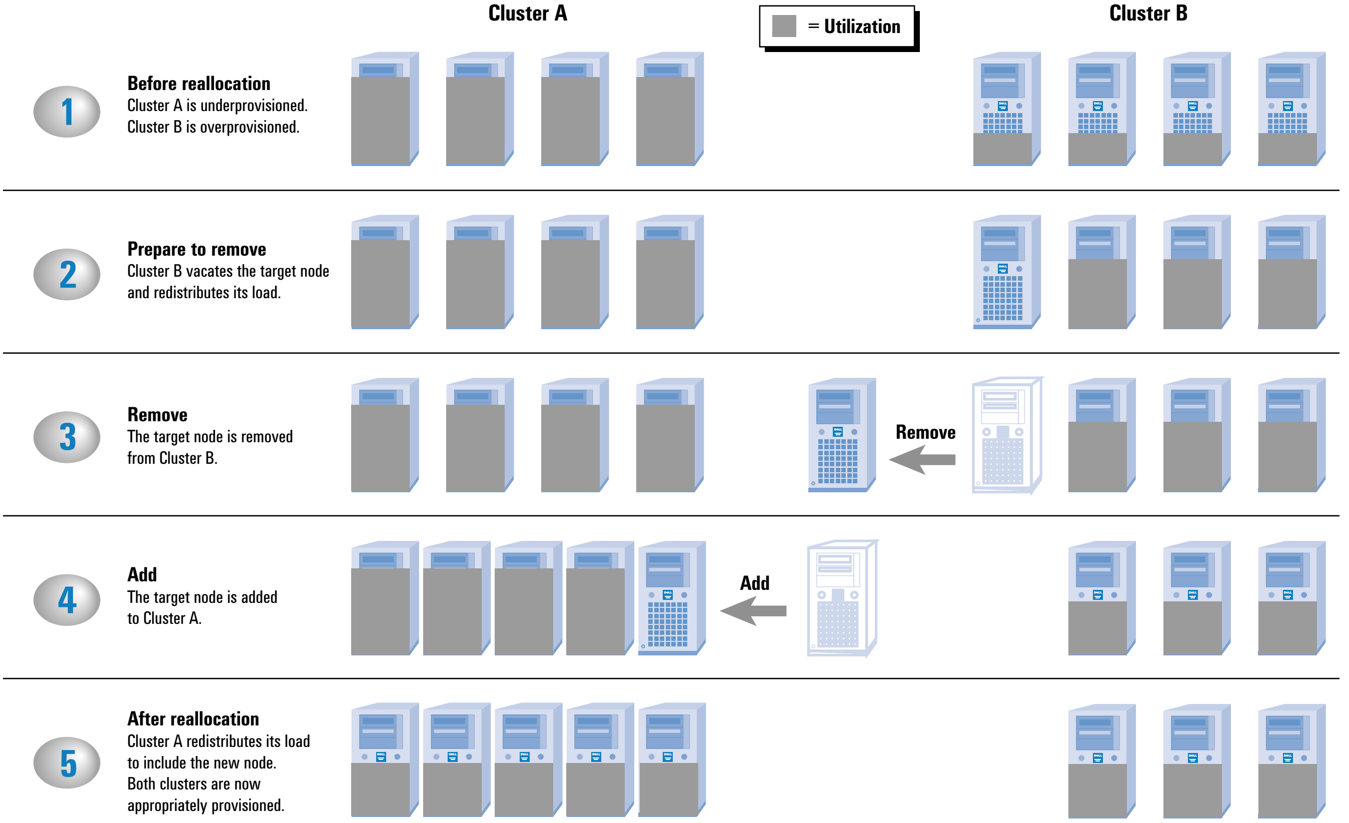

Cluster A
= Utilization
Cluster B
1
Before reallocation
Cluster A is underprovisioned.
Cluster B is overprovisioned.
2
Prepare to remove
Cluster B vacates the target node and redistributes its load.
3
Remove
The target node is removed from Cluster B.
Remove
4
Add
The target node is added to Cluster A.
Add
5
After reallocation
Cluster A redistributes its load to include the new node.
Both clusters are now appropriately provisioned.

Figure 1. Load balancing and resource reallocation

similar, easily "relocated" components, the virtual data center can optimize real-time, dynamic resource allocation purely through software.

Balancing loads and resources

Two mechanisms work in tandem to achieve maximum performance and optimal resource utilization in the virtual data center:

- **Load balancing:** The ability to redistribute work across the nodes in a cluster, commensurate with each node's processing capacity
- **Resource balancing:** The ability to move nodes among clusters to increase and decrease cluster sizes and thus cluster processing capacities

Load balancing occurs within clusters; resource balancing occurs between clusters. Synergy in both workload distribution and resource utilization is the main goal of the virtual data center. Although simple in concept, the implementation is not trivial.

A centrally acting controller that orchestrates operations across the entire data center could ideally make use of three load-balancing and resource-balancing mechanisms: *redistribute* work, *remove* node, and *add* node.[2] For example, Figure 1 shows how the virtual data center could balance loads and resources between two clusters running different cluster software, as follows:

1. **Before reallocation.** Cluster A is experiencing heavy demand and nearing the saturation point; it is *underprovisioned* because it requires additional hardware resources. Cluster B has spare capacity; it is *overprovisioned* because it has an abundance of hardware resources.
2. **Prepare to remove.** An identification algorithm targets a node in Cluster B for transfer to Cluster A. System administrators can program the logic in the identification algorithm to align with business objectives, such as minimizing the impact on clients or minimizing the time until the cluster makes the transferred node available.

[2] For more information about virtual data center management software and the role of the global engine, see "Managing the Virtual Data Center" by J. Craig Lowery, Ph.D., in *Dell Power Solutions*, August 2003.

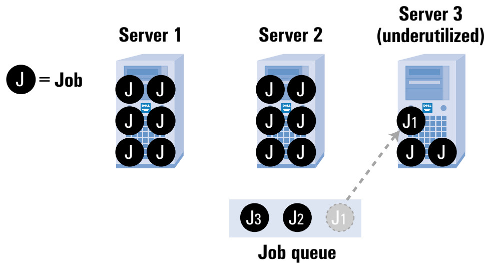


Figure 2. Job scheduling workload redistribution

Cluster B uses workload redistribution methods (discussed in the next section) to vacate this node.

3. **Remove.** The vacant target node is removed from Cluster B.
4. **Add.** The target node is added to Cluster A, which begins to redistribute its work across the new cluster member.
5. **After reallocation.** A steady-state workload exists in Cluster A; both Clusters A and B are provisioned appropriately.

As shown in Figure 1, the *redistribute*, *add*, and *remove* operations enable the reallocation of hardware resources according to changes in demand. The *add* operation is easy to implement because it does not require an immediate reaction from the affected cluster; new nodes are integrated in a nondisruptive fashion during the *redistribute* operation. The *remove* operation also is simple because work can be redistributed to vacate a target node before removing that node. Clearly, *redistribute* is the most critical of the three operations.

Redistributing workloads

Whereas load-balancing mechanisms determine the appropriate allocation of work across nodes, workload redistribution is the means of achieving that allocation. That is, workload redistribution techniques move a thread of execution—a job, a process, or a session—among nodes.

Assigning new jobs to nodes, or *job scheduling*, is a primary function of cluster operating systems. Usually the scheduler chooses the least utilized node for a new job, as shown in Figure 2, but it can employ other criteria as well. Once a job starts on a node, it runs to completion on that node.

Assigning new jobs to nodes, or job scheduling, is a primary function of cluster operating systems.

A *job* is usually defined as a process group spanning a time from creation to completion. Alternatively, a job may be defined as the unit of work performed by an always-resident server process in response to a particular client request; in this instance, the request represents the job. Generally, for workloads characterized by uniformly short job lengths, utilization is nearly equivalent across the nodes in steady state. However, when job lengths are unknown or highly variable, job completion times are impossible for the job scheduler to predict. Consequently, utilization across nodes in the cluster becomes skewed over time as the scheduler makes inefficient assignments.

Three operations implement the load-balancing and resource-balancing mechanisms: redistribute work, remove node, and add node.

In addition, unpredictable job completion times can diminish the value of the *remove* operation. When the cluster must vacate a node, the job scheduler excludes that node from new assignments. Because the completion times for existing jobs on the node are unpredictable, it is impossible to determine when the node will be ready.

Managing process migration

The most difficult workload redistribution method to implement is *process migration*. In this approach, the job scheduler initially assigns a process (that is, a job) to one node. The cluster management software subsequently suspends the process, moves it to a different node, and resumes execution (see Figure 3).

To appreciate the difficulties of process migration, consider the types and sizes of state information that general processes own. This state information must be copied from one system to another to effect a migration. When there is no overhead, process migration provides the greatest flexibility and the fastest response to configuration changes for jobs that possess very little state information. Sometimes migrating processes with sizable memory structures—such as arrays, large local files, temporary working files, network connections, and user interface I/O paths—is more time-consuming than simply letting the processes finish on the node to which they were initially assigned.

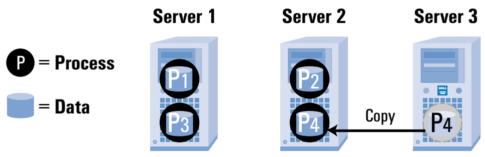


Figure 3. Process migration workload redistribution

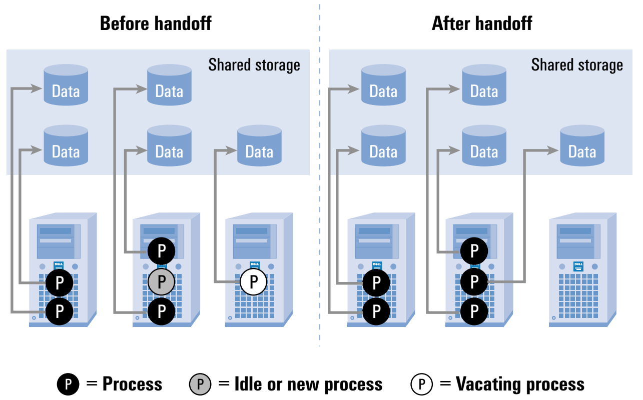


Figure 4. Session migration workload redistribution

If migration overhead were uniformly low, process migration would be an excellent workload redistribution mechanism. A refinement of the process migration concept is *session* or *transaction migration*, whereby a vacating process on one node hands off an in-progress transaction to a new or idle process on another node (see Figure 4). A combination of shared storage and network communication typically facilitates the handoff. Several options exist for implementing the session migration mechanism. For example, processes could keep state information in local storage and copy it to shared storage only when they must vacate the node. However, this approach offers little improvement over process migration.

Conversely, processes could work directly from shared storage all the time, allowing a process to vacate almost immediately if necessary. In this scenario, another process on a different node could pick up where the vacating process left off simply by accessing the share. Although this approach has been possible for a long time in shared-memory multiprocessor systems, it was considered impractical for cooperating servers on a LAN. However, relatively inexpensive, high-performance interconnects such as InfiniBand™ and Gigabit Ethernet[3] fabrics are enabling bandwidth-intensive cooperative activities, such as working directly from a storage area network (SAN) and remote direct memory access (RDMA).

Today's environments in which low-cost, standards-based hardware components can communicate at performance levels approaching that of the system bus enable systems designers to consider workload redistribution methods such as session migration for mass deployment. Because workload redistribution is the key to load and resource balancing, high-performance interconnects are critical to the success of the virtual data center.

An interesting variation of process migration that also can leverage new interconnect technologies involves virtual machine–hosted operating systems. Products such as VMware™ GSX Server™ and ESX Server™ software and Microsoft® Virtual Server run several guest operating systems concurrently on a host operating system by simulating a reference implementation of node hardware. All state information (the image) for the guest operating system resides in a large file, normally located in the local file system of the host.

This method makes it possible to suspend the execution of the guest operating system with its current state fixed in the image file, move the file to another host system, and resume execution. The transfer time for the image file introduces problems similar to those of process migration. However, by keeping the image file in shared storage and enabling multiple host systems to access the image file across a high-performance interconnect, nearly instant migration of an entire execution environment is attainable.

Because workload redistribution is the key to load and resource balancing, high-performance interconnects are critical to the success of the virtual data center.

The drawbacks to this method include the overhead for supporting the virtual machine and the large-size granularity at the operating system level rather than at the process level. Administrators could assign one process per virtual operating system to achieve finer granularity, but the virtualization overhead would become prohibitive. Even so, the combination of virtual operating systems, shared storage, and high-performance interconnects can be considered a step forward in achieving the goals of the virtual data center.

Achieving the primary goal of transparency

For nearly 20 years, the concepts of load balancing, resource balancing, and workload redistribution have appeared in academic literature describing distributed operating systems.[4] Much research into creating such systems has led to many of the advances now incorporated into the virtual data center concept.

[3] This term indicates compliance with IEEE® standard 802.3ab for Gigabit Ethernet, and does not connote actual operating speed of 1 Gbps. For high-speed transmission, connection to a Gigabit Ethernet server and network infrastructure is required.

[4] "Distributed Operating Systems" by Andrew S. Tanenbaum and Robbert Van Renesse in Association of Computing Machinery *Computing Surveys*, vol. 17, no. 4, December 1985.

To be truly successful, the virtual data center must be able to host applications that are indifferent to underlying hardware details and location, and are unaware that they may be subject to migration.

One key differentiating characteristic of distributed operating systems, as opposed to traditional operating systems, is *transparency*. That is, users of the system neither know, nor need to know, what network components are cooperating to service their requests and how the components accomplish those tasks.

Because the virtual data center is a form of the distributed operating system, transparency is a primary goal. To be truly successful, the virtual data center must be able to host applications that are indifferent to underlying hardware details and location, and are unaware that they may be subject to migration. Application developers should not have to worry about synchronizing with reconfiguration events. Ideally, the virtual data center presents applications with a virtual machine that provides continuity of execution at all times, making no special demands on applications to accommodate reconfiguration below the virtualization layer.

Some data center management software uses network boot capabilities to make a node execute entirely from shared storage—similar to diskless workstations. This method allows a node to change "personalities" by rebooting to a different image under the direction of some controlling agent, thereby providing location transparency. Although certainly useful, this method has limited benefit in the virtual data center because it is essentially coarse-grained job scheduling; administrators must wait for all executing jobs to complete before rebooting the node to a new image.

One way to achieve finer time granularity (that is, move applications around the data center more quickly) without sacrificing transparency is to construct environments specifically designed to relieve applications of the burdens associated with load balancing and workload redistribution. For example, many Oracle® products—most notably Oracle9*i*™ Real Application Clusters databases—include this capability, which Oracle refers to as *grid computing*.[5] Application programming interfaces (APIs) exist for location-transparent data access and processing, while the Oracle software stack provides a consistent virtualization layer above the hardware and operating system. Session migration is a natural extension of the grid features currently available in Oracle products.

Moving closer to the ideal virtual data center

Achieving cluster reconfiguration and load balancing that is both predictable and transparent is extremely difficult and presents several trade-offs. Job scheduling is transparent because no migration occurs, but it is not predictable. Process migration also is transparent but not predictable, because a potentially large amount of state information must be transferred. Session migration is more predictable when processes work directly from shared storage across high-speed network interconnects, but this predictability often is achieved at the expense of transparency. Products coming to market implement each of these approaches, some in ways that may eventually overcome their traditional shortcomings.

Technologies described years ago in academic journals are finally taking shape as tangible products.

Despite these challenges, much progress has been made in moving toward the ideal virtual data center, and the pace of development is quickening. Technologies described years ago in academic journals are finally taking shape as tangible products. High-performance interconnects, virtualization-ready execution environments, and standard components and protocols are paving the way to the eventual realization of this long-pursued computing model.

J. Craig Lowery, Ph.D. (craig_lowery@dell.com) is chief security architect and a software architect and strategist in the Dell™ Product Group–Software Engineering. Craig has an M.S. and a Ph.D. in Computer Science from Vanderbilt University and a B.S. in Computing Science and Mathematics from Mississippi College. His primary areas of interest include computer networking, security, and performance modeling.

FOR MORE INFORMATION

Microsoft Virtual Server:
http://www.microsoft.com/windowsserver2003/ evaluation/trial/virtualserver.mspx

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

VMware:
http://www.vmware.com

[5] The Oracle definition of grid computing differs somewhat from other connotations of grid computing. For more information about grid computing, visit http://www.globus.org.

Toward Transparency:
Setting Standards for Security

The computing industry is setting security standards through consensus, which helps promote transparent security implementations. By offering factory-installed operating systems preconfigured with security settings based on Center for Internet Security (CIS) benchmarks, Dell demonstrates its commitment to consensus-driven security standards.

BY J. CRAIG LOWERY, PH.D., AND CRAIG PHELPS

Technology generally tends to mature in phases. Initial proofs of concept are followed by explosive innovation, then a cooling-off period in which society assesses the full effects of a technology, identifies the advantages and disadvantages, and then integrates the technology into mainstream processes. Computer technology has been no different. Although innovation continues, many agree that new hardware and software features are less important than solving outstanding problems. Security is one such problem.

Computer users demand products that are secure by default and employ transparent security measures that do not make a product more difficult to use. Dell incorporates such customer requests into its product development cycles, but it must respond without increasing cost or complexity. For guidance, Dell turns to consensus-driven security standards.

Consensus-driven standards lead to transparency

When a broad set of stakeholders can contribute to a standard, they capture and consistently document best practices and solutions; everyone involved knows what to expect. A dearth of security standards leads to confusion and complicated, proprietary security implementations—the antithesis of transparency. Agreement on how best to achieve secure computer systems is just beginning. Best practices as communicated through nonprofit organizations such as the SANS (SysAdmin, Audit, Network, Security) Institute are prime examples of industry efforts to consolidate opinion on security topics.

Perhaps the most notable example of successful security through consensus is the work performed by the Center for Internet Security (CIS). CIS is a consortium of security specialists from nearly all sectors, including government, higher education, finance, and health care. Although many organizations have stepped forward with recommended security settings, CIS approaches this problem by first achieving a consensus among security professionals on proper configuration settings, and then providing a tool that measures how closely a system comes to meeting those settings.

CIS benchmarks reflect security recommendations

CIS provides a library of configuration standards known as benchmarks. Volunteer CIS members identify a system for which no consensus benchmark currently exists, gather recommendations concerning the best way to configure such a system so that it is more secure, and write a document that, after intense review and scrutiny by the members, achieves CIS benchmark status.

CIS benchmarks fall into two categories: Level I and Level II. Systems that meet Level I benchmarks achieve what CIS calls a "prudent level of minimum due care." Most security professionals would agree that these settings achieve a baseline for a reasonably secure system. Level I settings and actions require no special expertise to apply and usually will not diminish system usability. Many of the settings disable services that are not needed in most environments. Other settings disable features that are convenient but generally unnecessary. Level II benchmarks are much more complex; they further increase system security, but require more expertise in their application because they can affect usability.

Each benchmark has an associated scoring tool to help assess system compliance. When run on a target system, the tool produces a numeric score from 0 to 10—0 meaning no compliance and 10 indicating full compliance. Attaching a defined metric to security compliance enables system administrators to track improvements in overall security preparedness.

CIS initially focused its benchmark evaluation on operating systems but has expanded its scope to include firewalls, routers, and applications, among other targets. Because CIS is an open forum, anyone in the industry has access to the benchmarks and scoring tools at no charge through the CIS Web site at http://www.cisecurity.org.

Dell makes sense of the consensus

Dell believes in disseminating information such as best practices by promoting standards. The CIS benchmark model, shown in the top half of Figure 1, exemplifies this philosophy in the security arena—being user-driven, it is particularly suited to the Dell™ direct model, which places high importance on quickly integrating customer feedback into products.

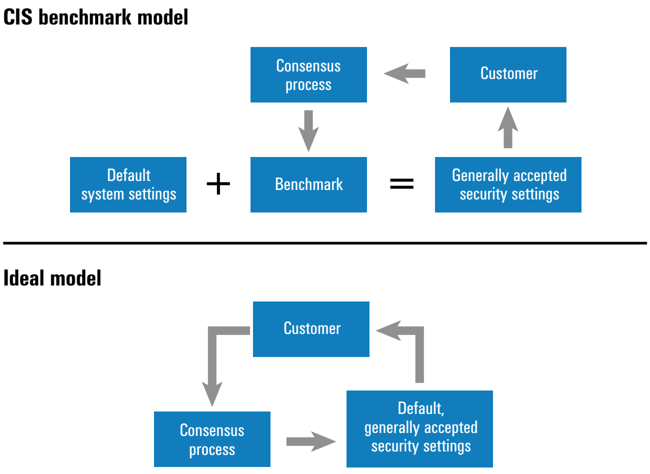


Figure 1. Achieving consensus: two models

Products such as Dell systems with CIS benchmark settings can help bring widely accepted security improvements to market now, while computer users wait for software releases that already incorporate these benefits.

Through its factory build-to-order and custom factory integration services, Dell offers Microsoft® Windows® 2000 operating systems preconfigured with CIS Level I benchmark settings. Although this offering resulted from suggestions by government customers, anyone can request it on Dell OptiPlex™, Dell Precision™, and Latitude™ desktops and laptops. Customer acceptance and further demand may prompt offerings for other operating systems and hardware platforms in the future.

By delivering systems preset to CIS benchmarks, Dell illustrates the viability of the consensus mechanism. Ideally, the consensus settings will be incorporated into software products at the source, as shown in the lower half of Figure 1, rather than being achieved through post-installation configuration changes. Dell believes this is beginning to happen, and offers as evidence the new "secure by default" philosophy that influenced the default configuration settings for the Microsoft Windows Server 2003 operating system. Products such as Dell systems with CIS benchmark settings can help bring widely accepted security improvements to market now, while computer users wait for software releases that already incorporate these benefits.

J. Craig Lowery, Ph.D. (craig_lowery@dell.com) is chief security architect and a software architect and strategist in the Dell Product Group–Software Engineering. Craig has an M.S. and a Ph.D. in Computer Science from Vanderbilt University and a B.S. in Computing Science and Mathematics from Mississippi College. His primary areas of interest include computer networking, security, and performance modeling.

Craig Phelps (craig_phelps@dell.com) is security brand manager in the Dell Public Sector Marketing Group. Craig has an M.B.A. from the Marriott School of Management at Brigham Young University (BYU) as well as a B.S. in Psychology and B.A. in English from BYU. His primary focus is identifying security measures for the public sector and then implementing them across Dell product lines.

FOR MORE INFORMATION

Center for Internet Security: http://www.cisecurity.org

SANS Institute: http://www.sans.org

Recommendations and Techniques for

Scaling Microsoft SQL Server

To support many more users, a database must easily scale out as well as up. This article describes techniques and strategies for scaling out the Microsoft® SQL Server relational database management system (RDBMS) and provides scenarios illustrating scale-out deployments.

BY DON JONES

Most enterprise applications today run on a Microsoft® Windows®, UNIX®, or Linux® operating system–based relational database management system (RDBMS), such as Microsoft SQL Server 2000. Scalability has become a critical factor in the success of these applications as the number of users relying on them has grown.

The Internet also has profoundly affected the need for scalability. Once exposed to just a few thousand users, the data in many corporate databases now must be accessed by tens of thousands of concurrent users through e-commerce sites, Web services, and other Internet-based applications. Scaling databases to support these users is a major concern for both database software developers and database administrators.

Differences between scaling up and scaling out

When database performance worsens, administrators typically address the problem first by scaling up—that is, by trying to optimize performance in the current environment. Because many database applications have inefficient designs or become inefficient as their usage patterns change, finding and improving the areas of inefficiency can yield significant performance benefits. Fine-tuning the database server can help perform more queries, handle more users, and run more efficiently.

SQL Server scales up fairly well—to a point. In one real-world scenario, for example, a company's database required a nine-table join to look up a single customer address. Selectively denormalizing the tables and applying strategic indexes allowed SQL Server to execute address queries much faster. Because address lookups were a common task for this company, even a minor per-query improvement significantly enhanced overall server performance.

When database performance worsens, administrators typically address the problem first by scaling up.

Unfortunately, scaling up is limited in how much it can improve an application's performance and ability to support more users. For example, take a database whose sole function is to perform a single, simple query—no joins, no need for indexes. A high-performance SQL Server computer—for example, a quad-processor server with 4 GB of RAM and several fast hard drives—could probably support tens of thousands of users who must concurrently execute that one query. However, this server might not be able to support a million users. In this situation, scaling up—fine-tuning—would be insufficient, because such a simple query leaves little room for improvement. To begin supporting many more users, scaling out is a better solution.

Scale-out strategies redistribute workloads

Scaling out SQL Server, a more complicated process than scaling up, requires splitting a database into various pieces, then moving the pieces to different, independent SQL Server computers. The grocery-store checkout line presents a good analogy for comparing the two processes. In a busy grocery store with only one checkout lane open, a long line of unhappy customers would quickly materialize.

A scale-up approach—installing faster barcode scanners, requiring everyone to use a credit card instead of writing a check, or hiring a faster cashier—can make the checkout process itself more efficient. These measures might improve the situation, but not solve the problem; customers would move through the line more quickly, but they still would have only one checkout lane.

A better solution would be to scale out—in this analogy, by opening additional checkout lanes. Customers could now be processed in parallel by completely independent lanes. To make the analogy closer to a database scale-out scenario, the grocery store could have specialized lanes: one that expedites processing (customers purchasing 15 items or fewer), and another that focuses on produce, which often takes longer because it must be weighed and not simply scanned.

An ideal, if unrealistic, solution might be to retain a single lane for each customer, but to divide each customer's purchases into categories to be handled by specialists: produce, meat, boxed items, and so forth. Specialized cashiers could minimize their interactions with each other, keeping the process moving speedily along. Although unworkable in a real grocery store, this solution illustrates a real-world model for scaling out databases.

General strategies for scaling out databases

Database managers can consider two basic scale-out strategies for distributing the workload of a database across multiple servers. Most major RDBMS platforms, including SQL Server, provide the means to make these strategies possible.

SQL Server farms replicate the database

The first approach simply adds more servers. Consider a scenario in which a company has an office in New York and one in Los Angeles. Both offices have several thousand users who frequently query data from a corporate application, such as an order-processing database. Users rarely change data in the system, but they frequently add new data. In this scenario, users in both offices are overloading the database. Even if the database is a well-written multitier application, processing all the information on only one database server at the back end can create a bottleneck.

Scaling out SQL Server, a more complicated process than scaling up, requires splitting a database into various pieces, then moving the pieces to different, independent SQL Server computers.

Figure 1 illustrates one way to address the problem: a SQL Server farm. In this technique, two database servers each contain a complete copy of the database. Each office houses one server, and the users in each office connect only to their local server. Changes and new records are replicated between the servers by using SQL Server replication. To avoid conflicts when adding new records, each office might, for example, be assigned a unique range of order ID numbers, ensuring that new records created in either office can be uniquely identified across both copies of the database.

This strategy is perhaps the simplest means of scaling out SQL Server. Although replication is not easy to set up and maintain on SQL Server, neither is it extremely difficult. The strategy works well even with many servers and copies of the database.

However, the data replication strategy does incur some drawbacks, especially latency. Neither copy of the database will ever match the other exactly. As new records are added to each copy, time elapses before replication begins. With only two servers in the company, each server might be as much as an hour out of sync with the other, depending upon how administrators set up replication.

Adding more servers, however, involves difficult replication decisions. For another scenario, consider the six-office setup depicted in Figure 2. Each of the six offices has its own independent SQL Server system—an excellent design for scalability. However, latency could be very high. If each SQL Server replicates with its partners just once every hour, then total system latency could be three hours or more. A change made in the Los Angeles office would replicate

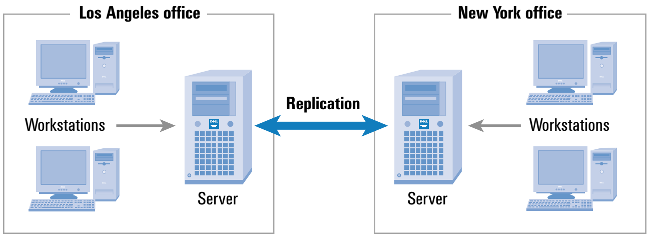


Figure 1. SQL Server farm

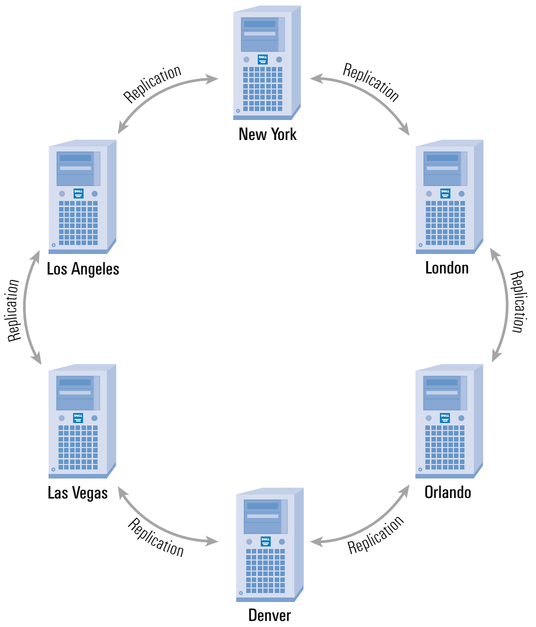


Figure 2. Six-server farm

to New York and Las Vegas in about an hour. An hour later, the change would reach London and Denver. An hour later, it would arrive in Orlando. Given such high latency, the entire system would probably never be synchronized completely.

Administrators can reduce latency, but at a performance cost. If each of the six servers replicated with each of the other five servers, the system could *converge,* or be universally in sync, about once an hour (assuming again that replication occurred every hour). Figure 3 shows such a *fully enmeshed* design.

In this fully enmeshed design, each server must maintain replication agreements with five other servers, and must replicate with each server every hour. This much replication, particularly in a busy database application, would likely slow response so much that the performance gain achieved by creating a server farm would be lost. Each office might require two servers just to maintain replication and meet users' needs. Although fairly easy to implement, the server farm technique has a point of diminishing returns.

Distributed partitioned databases move tasks to different servers

A more sophisticated strategy—but one that is also more difficult to implement—involves partitioning the database and moving the pieces to different servers. Unlike the simplified order-processing database example previously discussed in "SQL Server farms replicate the database," most real-world database applications tend to rely on an equal mix of data reading and data writing. For example, an order-processing application might include a product catalog that is largely read only, a customer-order database that is write heavy, and tables containing supplier information that are equally read-write. These three closely related database segments—catalog, orders, and supplier tables—are fairly task-independent: diverse users within the organization tend to use each database differently. Merchandisers might write to the catalog but do little else. Customer service representatives might read the catalog and write to the orders tables but never access the supplier tables. The warehouse staff might read the catalog and read from and write to the supplier tables. This division of labor indicates where the database itself can be split, as Figure 4 illustrates.

Scaling out SQL Server can offer benefits not only in improved application performance, but also in greater redundancy and availability.

Administrators can use two basic approaches to implementing the distributed partitioned database strategy. The first is to modify the client application so that it understands the division of the database across multiple servers. Straightforward yet somewhat time-consuming, this solution does not work well for the long term. Future changes to the application could result in additional divisions, which would in turn require additional reprogramming.

A better approach is to program the client application to use stored procedures, views, and other server-side objects—an ordinary best practice for a client-server application—so that the client application need not be aware of the physical location of the data. SQL Server offers different techniques, such as distributed partitioned views, to handle this setup.

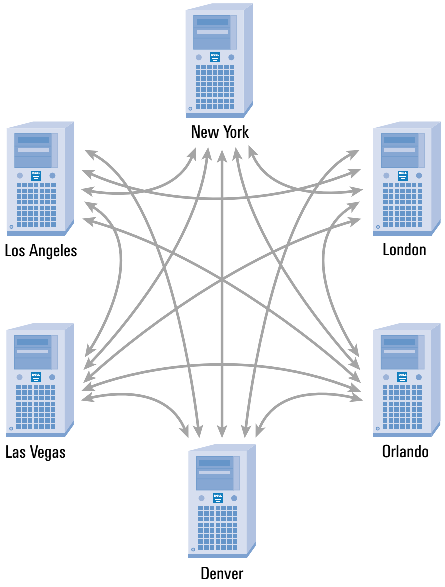


Figure 3. Fully enmeshed six-server farm

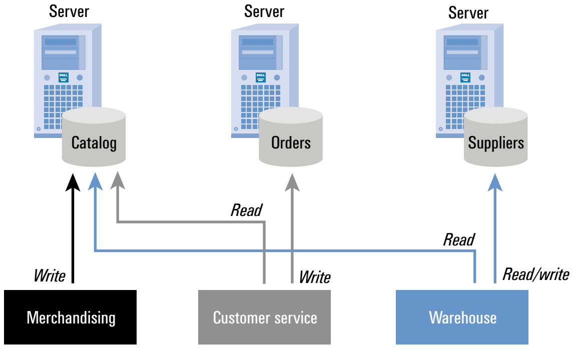


Figure 4. Identifying task-based divisions in the database design

Scale-out techniques using SQL Server and Windows

SQL Server and Windows offer several techniques to enable scaling out, including SQL Server–specific features such as distributed databases and views and Windows-specific functions such as Windows Clustering.

Distributed partitioned views help create virtual tables

SQL Server distributed partitioned views allow developers to create views that combine tables from multiple SQL Server computers into a single virtual table. This method logically divides a database across multiple SQL Server computers. Rather than reprogramming client applications to understand the division of the databases, developers can create distributed views that present a virtualized version of them. These tables appear to client applications as if they were on a single server. Meanwhile, SQL Server combines the tables, which are spread across multiple servers, into a single view.

Distributed views are a powerful tool in scaling out. They allow developers to redistribute databases transparently to the end users and their business applications. As long as client applications are designed to use the views rather than the direct tables, the tables themselves can be rearranged and scaled out as necessary without the client application being aware of any change.

The workload required to create and present the view to client computers is shared by all servers participating in the view—or by all servers in the federation. SQL Server 2000 is the first version of SQL Server to make a significant improvement to this approach, because the data within the views can be updated by client applications as if the data were in a regular table. The updates are cascaded back to the necessary participant servers.

> SQL Server and Windows offer several techniques to enable scaling out, including SQL Server–specific features such as distributed databases and views and Windows-specific functions such as clustering.

Replication of distributed partitioned databases reduces latency

Another scale-out approach involves partitioning a database across multiple servers and then replicating the database copies. Like the six-server order-processing farm described earlier, each server contains a complete database. In this method, each server is responsible for a different set of rows. SQL Server replication is used to keep each copy of the database updated. This method allows each server to immediately access its own rows and provides reasonably low latency for access to rows created on other servers. Client applications often must be modified to understand this structure. In many partitioned database schemes, data rows may be modified only on the server that owns them, with the changes then being moved to the other servers through replication. Client applications must know how to determine which server owns a row before making modifications.

Windows Clustering facilitates high availability and scalability

Besides improving performance, Windows Clustering can help avoid the risk of server failure when scaling out. For example, a two-node active/active cluster has two independent SQL Server servers. These nodes can be configured as a server farm, in which each server contains a complete copy of the database and users are distributed between them. An alternative is a distributed database architecture, in which each server contains one logical half of the entire database. In either architecture, a failure of one server is not catastrophic because Windows Clustering enables the other server to transparently take over and act as two servers.

Over-engineering is the key to a successful active/active cluster. Each node should be designed to operate at a maximum of 60 percent capacity. If one node fails, the other node can begin running at 100 percent capacity, incurring only about a 20 percent loss of efficiency. Still, performance is generally well within an acceptable range considering that, after failover, applications must run on half as much hardware.

Setting up clusters can be extremely complex. In Windows Clustering, the software is not difficult to use, but the underlying hardware must be absolutely compatible with Windows Clustering—and most hardware vendors have exacting requirements for cluster setups. Purchasing preconfigured clusters from a major server vendor, such as Dell, can help simplify cluster setup. The cluster is designed to be ready to run on delivery, and both the vendor and Microsoft can provide cluster-specific technical support if necessary.

High-performance storage to boost SQL Server response

High-performance storage is an often-overlooked performance benefit for SQL Server—particularly external storage area networks

UNDERSTANDING DATABASE INEFFICIENCY

Databases can be inefficient for several reasons:

- **Poor design:** Many application developers do not excel at database design. Some, for example, have been taught to fully normalize the database at all costs, which can lead to significantly degraded performance. Sometimes project schedules do not permit enough design iterations before the database must be locked down and software development begins. In some cases, the application itself is not designed well, resulting in an incomplete database design that must be patched and expanded as the application is created.
- **Change:** An application used in a way unintended by its designers can reduce efficiency. The application may have expanded and begun suffering from "scope creep"—the growth or change of project requirements. In this case, redesigning the application from the beginning to meet current business needs may be the best solution to database inefficiency.
- **Growth:** Databases are designed for a specific data volume; once that volume is exceeded, queries may not work as they were intended. Indexes might need to be redesigned or at least rebuilt. Queries that were intended to return a few dozen rows may now return thousands, affecting the underlying design of the application and the way data is handled.

These problems are difficult to address in a live, production application. Scaling up tends to have a limited effect. Although developers may agree that the application's design is inefficient, companies are reluctant to destroy a serviceable application and start over without serious consideration.

Scaling out can offer a less drastic solution. Although scaling out requires considerable work on the server side, it may not require much more than minor revisions to client-side code, making the project approachable without completely re-architecting the application. Scaling out might not be the most elegant or efficient way to improve performance, but it does help alleviate many database and application design flaws. It also can allow companies to grow their database applications without needing to redesign them from the beginning.

(SANs) that rely on Fibre Channel technology rather than traditional SCSI disk subsystems. Because high-performance storage enables an existing server to handle a greater workload, it constitutes an example of scaling up rather than out.

SQL Server is a highly disk-intensive application. Although SQL Server includes effective memory-based caching techniques to reduce disk reads and writes, database operations require significant data traffic between a server's disks and its memory. The more quickly the disk subsystem can move data, the faster SQL Server will perform. Some industry estimates suggest that 75 percent of idle time in SQL Server results from waiting for the disk subsystem to deliver data. Improving the speed of the disk subsystem can markedly improve overall SQL Server performance.

Moving to additional RAID-5 arrays on traditional copper SCSI connections is a simple way to improve disk space. However, high-speed Fibre Channel SANs offer the best speed, as well as myriad innovative recovery and redundancy options—making them a safer place to store enterprise data.

Scale-out strategy for improving SQL Server performance, redundancy, and availability

As applications grow to support tens and hundreds of thousands of users, scaling is becoming a mission-critical activity. Scaling up—improving efficiency by fine-tuning queries, indexes, and so forth—helps IT organizations do more with less. However, scaling up can require high administrative overhead and may have limited effect. Administrators might spend two weeks to achieve a 1 percent performance gain, an improvement that cannot compare to the much higher gains promised by a well-planned scale-out design.

Although seldom considered as a target for scaling out, SQL Server is well suited to this strategy, in both server farm and more sophisticated distributed database approaches. Scaling out SQL Server can offer benefits not only in improved application performance, but also in greater redundancy and availability.

Don Jones is a founding partner of BrainCore.Net, and has more than a decade of experience in the IT industry. Don's current focus is on high-end enterprise planning, including data availability and security design.

FOR MORE INFORMATION

This article is based on an excerpt from the free eBook *The Definitive Guide to Scaling Out SQL Server* (Realtimepublishers.com) by Don Jones, available at http://www.dell.com/sql/ebook

Dell | Microsoft SQL Server 2000: http://www.dell.com/us/en/esg/topics/products_software_pedge_001_database.htm

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

Building Out the Enterprise Using Automated Deployment Services

Automated Deployment Services (ADS), a component of the Microsoft® Windows® Server 2003 operating system, includes a new set of imaging tools and a secure infrastructure for rapid, large-scale remote deployment. ADS also offers a secure, reliable script execution framework for performing script-based administration on hundreds of servers as easily as upon a single server.

BY MICROSOFT CORPORATION

Enterprise data center administrators face significant challenges as they scale out their deployments of Microsoft® Windows® operating system (OS)–based servers. Automated Deployment Services (ADS), a new component in Windows Server 2003, Enterprise Edition, enables IT managers to efficiently deploy the Windows operating system onto bare-metal servers using its imaging tools, and to administer a large number of Windows-based servers through script-based remote execution. ADS helps IT organizations address the challenges involved in scaling out by offering the following features:

- **Secure, scalable remote deployment:** Automated deployment using integrated ADS services can facilitate secure, auditable, large-scale OS and application installation onto bare-metal servers.
- **Automated task sequences:** ADS can automatically issue reusable *task sequences,* or sets of commands, achieving more reliable and consistent operations.
- **Flexible imaging tools:** New imaging tools can quickly create gold-standard system image libraries and allow an individual image to be updated and edited without first being deployed to a server.
- **Multiple user interfaces:** Administrators can manage ADS or integrate ADS with other tools in their deployment process using command-line operations, Microsoft Management Console (MMC) user interface (UI) snap-ins, or a complete Windows Management Instrumentation (WMI) interface.
- **Reliable remote-execution framework:** Administrators can manage hundreds of servers remotely by leveraging their existing script investment.
- **Virtual Floppy:** By downloading an MS-DOS–based Virtual Floppy image that executes in memory, administrators can perform automated hardware configuration tasks.
- **Centralized data store:** Auditing is designed to be more reliable because the centralized data store maintains a complete history of all administrative tasks performed using the ADS infrastructure.

Understanding how ADS enhances deployment and administration

As the growth rate of Windows-based servers increases, managing systems deployment and administration throughout the enterprise has significantly increased total cost of ownership (TCO).

Automated deployment of operating systems and applications usually relies on scripts or on imaging and deployment tools from third-party vendors. Using scripts to manage a large number of Windows-based servers traditionally has been very time-consuming. In a UNIX® OS environment, administrators can use tools such as `rsh`, `ssh`, and `rdist` to remotely manage groups of servers whereas administrators in a Windows environment have had to manage each server individually. In addition, although script-based installation works well for many different hardware configurations, scripts tend to be very slow and generally are not standardized. Although imaging software is much faster, it can be inflexible; and keeping an image collection updated requires considerable effort.

Automated Deployment Services enables IT managers to efficiently deploy the Windows operating system through imaging and to administer a large number of Windows-based servers through script-based remote execution.

With ADS, Microsoft has extended the Windows Server 2003 platform to facilitate rapid, flexible deployment and smooth, script-based administration for large numbers of Windows-based servers. ADS consists of an integrated set of services—including a Controller Service, Network Boot Service (NBS), and Image Distribution Service (IDS)—that run on a controller, as well as volume imaging tools and a set of agents for deployment and administration (see Figure 1). Working together, these features can greatly enhance the ability of administrators to deploy and manage a large number of Windows servers.

Controller Service orchestrates ADS activity

The Controller Service is the operational heart of ADS, orchestrating all ADS activity. The service sends configuration information to other ADS services; provides several interfaces for administrator input; and maintains the master records for each device known to the ADS system, including details such as the actions associated with those devices.

The Controller Service performs the following functions:

- **Coordination and sequencing of tasks:** The Controller Service coordinates deployment and administrative activity. For example, it provides the appropriate boot commands for each device to the Preboot Execution Environment (PXE). During deployment, the Controller Service coordinates the task sequence that configures the server and installs the images locally on the system. When the system reboots after the initial image deployment, devices check with the Controller Service for any waiting boot commands—enabling administrators to completely repurpose a device on its next boot, as though it were a bare-metal server.
- **Secure communication with devices:** The Controller Service communicates securely with the ADS Deployment and Administration Agents that reside on devices and run task sequences.
- **Centralized recording of device data and administrative activity:** The Controller Service uses either Microsoft SQL Server 2000 Desktop Engine (MSDE) or Microsoft SQL Server 2000 to store all device and configuration data and to log information for all tasks performed on the devices. A complete audit trail for each task run at the device is available through the Controller Service.
- **Logical grouping of device assets:** Although the Controller Service can manage each device individually, it generally works with *sets*, which are groups of managed devices that can be addressed as a single entity. For example, to deploy an OS on a group of devices, administrators need to reference only the set name of the group. A set also can contain references to other sets, creating a hierarchy that allows administrators to execute everyday management commands on all devices with a single command. Administrators can use smaller sets to provide more detailed control over devices. A single command from the command line or MMC snap-in can initiate actions to hundreds of devices at one time.

NBS offers boot command capability

Network Boot Service (NBS) works with Dynamic Host Configuration Protocol (DHCP) to provide ADS with boot command capability.

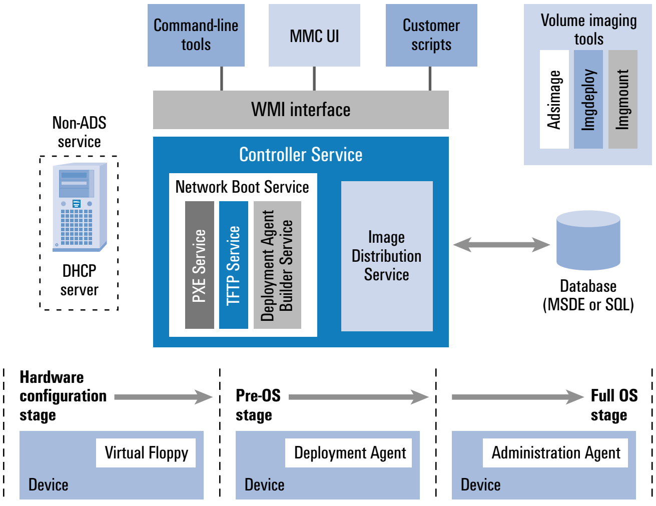


Figure 1. ADS architecture

To begin a communication session with ADS, a PXE-enabled device uses DHCP to discover the PXE Service. NBS includes ADS PXE Service, which sends PXE boot commands to devices. This service can instruct the destination device to perform the following functions:

- Download and boot an ADS Deployment Agent created for the device by the Deployment Agent Builder Service
- Download and boot a Virtual Floppy image
- Ignore the PXE boot request
- Abort PXE operations and boot to hard disk

NBS also includes ADS Deployment Agent Builder Service, which creates a device-specific agent at boot time. After the PXE boot occurs and the Deployment Agent option is chosen, the Builder Service can implement the following tasks:

- Perform a local hardware discovery on the device and send that data to the controller
- Build a customized Deployment Agent based on the device's hardware configuration
- Download the customized agent to the device, on which the agent executes in local memory

NBS provides a way to download an MS-DOS–based Virtual Floppy image that executes in memory. Utilities usually provided by server vendors can perform hardware configuration tasks (such as RAID drive configuration and BIOS flashing) in the Virtual Floppy environment.

NBS uses the Windows Server 2003 Trivial FTP (TFTP) Service to download both Deployment Agents and Virtual Floppy images. Either on the controller or elsewhere in the network, ADS requires that a DHCP server be present to provide IP addresses to configurable devices.

IDS simplifies management of device images

Another service within ADS is the Image Distribution Service (IDS), which provides storage and communications capabilities for managing device images. Imaging is a straightforward process that enables users to capture and rapidly deploy complete OS and application bundles to one or many devices simultaneously. IDS allows administrators to quickly and efficiently deploy device OS images created using the ADS imaging tools.

The imaging process comprises several discrete tasks:

- **Installing a reference device:** Using a hardware configuration similar to that of the devices to be deployed with the image, administrators install the base OS and any other applications and device drivers that the image will need in production, and add the Administration Agent (see the "ADS agents enhance pre- and post-deployment manageability" section later in this article).
- **Preparing the reference device for imaging:** Administrators use the Microsoft System Preparation (Sysprep) tool to prepare the system for imaging. Sysprep removes unique system information and prepares the system to run Windows Mini-Setup the next time it boots. This procedure allows every device created from this image to generate the information that makes it unique to the Windows environment.
- **Capturing the boot disk partition using image tools:** ADS allows administrators to manually capture a device image by running a command-line tool on the reference system. Alternatively, administrators can use ADS to remotely capture the image when the system reboots.
- **Deploying the image to new or existing devices:** ADS deploys the OS image to a new or existing device, which can then join the production environment. From a single image, hundreds of devices can be deployed easily, quickly, and in a completely reproducible manner.

After the images are captured, they are added to the IDS image store, where they are kept on the IDS server's file system. While the images are downloading to the device, the communications channel is encrypted to prevent eavesdropping on the image data. The Controller Service orchestrates the imaging operation, communicating with the device through a secure connection using the Secure Sockets Layer (SSL) protocol. IDS can deploy images to multiple devices over unicast and multicast networks. In addition, IDS enables users to adjust the amount of network bandwidth that is used during the multicast imaging transmissions.

Volume imaging tools offer powerful image handling features

Any file allocation table (FAT) or NT file system (NTFS) volume can be captured and deployed when used in conjunction with Sysprep, but administrators gain enhanced capture and editing benefits when working with NTFS-based file systems. ADS has a powerful, flexible set of imaging tools: imgdeploy, imgmount, and adsimage.

Imgdeploy. This image capture and restoration tool enables image compression, which reduces storage requirements; image encryption, which protects images for transport; and defragmented restoration, which captures images in file order so that files are

As the growth rate of Windows-based servers increases, managing systems deployment and administration throughout the enterprise has significantly increased TCO.

automatically defragmented when restored to a device.

Imgmount. This tool enables true editing of images captured with imgdeploy. Imgmount can mount and access captured images exactly as a file system does. After mounting, an image can be read from and written to by any standard Windows tool or application. The saving in administration time is considerable: the image can be maintained without modifying the reference system from which it was captured and then recapturing the image.

Adsimage. This tool is accessed from the command line or the MMC UI and lists images available for deployment. It adds images, deletes images, or updates image properties.

With ADS, Microsoft has extended the Windows Server 2003 platform to facilitate rapid, flexible deployment and smooth, script-based administration for large numbers of Windows-based servers.

ADS agents enhance pre- and post-deployment manageability

ADS provides two control agents, one to perform pre–OS deployment tasks and one to manage post-deployment operations:

- **Deployment Agent:** Used to handle deployment operations such as disk partitioning and image downloading, the Deployment Agent is a highly optimized, reduced-functionality version of Windows Server 2003 that loads into a RAM disk on the device. This agent is available only for environments using PXE boot control and is the key facilitator of a pure network-based OS deployment.
- **Administration Agent:** Used to handle post-deployment task execution, the Administration Agent allows administrators to initiate operations on the device that can perform any scriptable command and to run local applications or any application reachable on the network. By using the Administration Agent, administrators also can restart the entire deployment to completely repurpose a given device.

Task sequences automate commonly performed jobs

At boot time or after deployment, the Controller Service issues jobs to devices. Jobs consist of individual tasks sequenced together; these sequences can be run against one or more devices and are stored as XML files on the controller. A typical task sequence used to deploy a new device with PXE includes the following steps:

1. Execute a Virtual Floppy to update system BIOS or configure a RAID controller.
2. Request and boot the Deployment Agent.
3. Partition the hard disk.
4. Download an OS image to the hard disk.
5. Modify the Sysprep.inf file that was downloaded in the deployed image to give this device a unique host name, product identifier, and so on.
6. Configure the Administration Agent in the downloaded image to communicate with the controller when the OS first boots.
7. Reboot.
8. Instruct the device to boot to the local hard disk on the next boot.

When the device boots from the newly downloaded image for the first time, an abbreviated setup runs in an unattended mode to perform final configuration of the OS.

Task sequences can consist of the following commands and scripts:

- PXE boot commands
- Deployment Agent internal commands
- Any Windows application or script locally available to the destination device or on a network share accessible by the destination device
- A script available on the controller to be downloaded to the destination device and run on the destination device
- A script available on the controller to be run directly on the controller

When using ADS tasks, the Controller Service logs all issued commands and their output to the service's database, providing a complete audit log as well as an error log for debugging commands.

XML task sequences can be created with any XML authoring tool, but ADS also includes a simple Sequence Editor for administrators who prefer not to work directly with XML (see Figure 2). The ADS Sequence Editor provides a Windows-based UI for creating ADS sequences. It also provides templates for the most common administrative tasks, including deployment steps such as partitioning the disk, downloading an image, and personalizing the image. To create a new task sequence or to edit existing sequences, administrators simply open the Sequence Editor, manipulate parameters, and save the result as a new sequence.

Administration tools and programmatic interfaces enable tailoring ADS

ADS provides a series of command-line tools and an MMC snap-in to manage its services. Especially useful is the programmatic interface exposed through the WMI interface. All ADS features are available through this interface, enabling in-house developers to tailor and manage ADS for their particular environments.

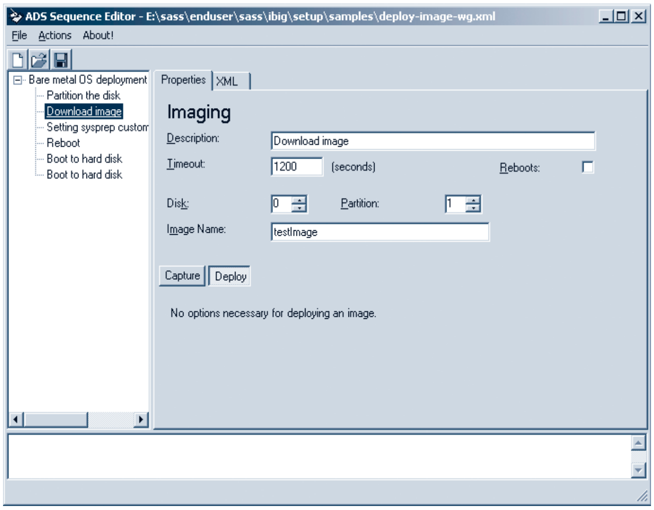
WMI *object areas* are those entities available for scripting or application development by using Microsoft Component Object Model (COM) communication. The primary object areas include:

- **Devices:** Physical devices within the data center
- **Sets:** Collections of devices; a given device may be in multiple sets
- **Job templates:** Definitions describing a simple job or a task sequence
- **Jobs:** Tasks in progress or already completed
- **Images:** Captured volumes available for deployment
- **Services:** Each of the ADS services

Some command-line tools available to manage and operate ADS include:

- **Adsarchive:** Removes previous jobs from the database and stores them in an XML file
- **Adsdevice:** Manages device records and allows administrators to list, add, delete, control, or modify device records
- **Adsimage:** Lists the templates available to execute as jobs; installs and deletes templates

Please note that the tools operating the Controller Service are built directly on the WMI interface; other command-line tools that provide imaging or certificate management neither depend on nor require WMI functionality to operate. Examples of these other command-line tools include imgmount.exe, which creates a volume from an image file and mounts it; and dskimage.exe, which captures the sectors to and from a floppy disk and creates a file for hardware configuration or BIOS flashing.



Imaging

Description: Download image

Timeout: 1200 (seconds) Reboots: □

Disk: 0 Partition: 1

Image Name: testImage

No options necessary for deploying an image.

Figure 2. ADS Sequence Editor Properties tab

All together, the three interface options offer administrators the flexibility of either using the GUI to be guided through tasks and procedures or scripting directly to the command-line interface.

Additionally, an MMC snap-in provides GUI access to the ADS services (see Figure 3). All together, the three interface options offer administrators the flexibility of either using the GUI to be guided through tasks and procedures or scripting directly to the command-line interface.

Meeting requirements of the ADS network environment

ADS requires a well-connected networking environment—10 MB/sec or higher connectivity among the controller, IDS, NBS, and managed devices. ADS makes no provisions for general retries when tasking devices or downloading images. That is, operations either complete successfully or fail—making ADS impractical for usage over low-speed wide area network (WAN) links.

In a subnetted or virtual local area network (VLAN) environment, the scope of DHCP and PXE broadcasts is usually constrained by the IP subnet or VLAN of the attached network. In these environments, administrators must configure DHCP forwarding at the network routing points to return all the devices to one DHCP and NBS server. In Cisco® networking environments, administrators accomplish this by configuring an IP Helper at the router or Layer 3 switch. In environments that do not support DHCP forwarding in the routers, Microsoft provides a DHCP forwarding service in the server operating systems to perform this function.

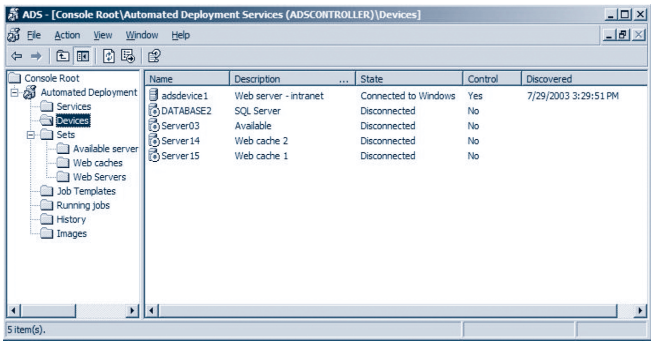
For large device deployments, ADS provides multicast support, which can reduce the overall bandwidth used when deploying disk images. In a switched or routed environment, equipment must be configured and capable of supporting the multicast feature.

Providing security in ADS

The ADS security model is divided into three distinct realms:

- **Initial deployment stages using DHCP and PXE:** The DHCP and PXE protocols are not designed to authenticate during the initial phases of bare-metal deployment. Consequently, Microsoft recommends that the deployment network interface reside on a dedicated management network.

- **Communication among ADS services; communication between Controller Service and device agents:** ADS uses a public key infrastructure (PKI) and SSL server authentication



Name	Description	State	Control	Discovered
adsdevice1	Web server - intranet	Connected to Windows	Yes	7/29/2003 3:29:51 PM
DATABASE2	SQL Server	Disconnected	No	
Server03	Available	Disconnected	No	
Server14	Web cache 2	Disconnected	No	
Server15	Web cache 1	Disconnected	No	

5 item(s).

Figure 3. MMC snap-in GUI for managing ADS devices

to create private out-of-band management communication channels. In addition, images transferred between the IDS and the Deployment Agent Builder Service during image capture and deployment are encrypted by default.

- **Security context for ADS and ADS agents:** The Controller Service operates in the context of the system service, the equivalent of local Administrator access level. This environment is necessary because of the configuration control delegated to the ADS Controller Service. On managed devices, however, the Administration Agent is installed by default using the local system's Administrator account. This account can be changed to that of any user who has the appropriate authority to run the applications and scripts used as ADS jobs. Also, whatever account is chosen will require access to network resources if users execute scripts or applications from network share points.

The ADS security model assumes that the data center is physically secure and free from network intrusion. More specifically, the network between the servers running ADS and the devices should be secure. Generally, access to the ADS Controller Service is restricted to users who are members of the local Administrator group on the controller. Users who are not members of the Administrator group cannot run jobs from the controller unless explicitly granted permission by an account at the Administrator level using role-based security functions. System administrators should limit physical access to the data center and to the Administrator group.

The Windows Server 2003 platform now includes ADS—tools and services that enable rapid, flexible, automated deployment and administration across hundreds of servers.

Besides securing access to the Controller Service, administrators should secure additional data used by ADS. On the controller, XML task sequence files and script files that are referenced through job templates are stored on the file system and should be protected by file system access control lists (ACLs) to ensure that non-administrators do not have modify rights to these files. Images ready for deployment are stored by IDS and also protected by the file system. Finally, the Administration Agent running on a device may reference scripts that should also be locked down by the file system.

By taking advantage of ADS, administrators can automate operations, increase security, and lower TCO of their Windows OS deployments.

Simplifying large-scale Windows deployments and lowering TCO

Using traditional means to scale out the Windows OS to large numbers of servers can be complex, time-consuming, and expensive. The Windows Server 2003 platform now includes ADS—tools and services that enable rapid, flexible, automated deployment and administration across hundreds of servers. By taking advantage of ADS, administrators can help to automate operations, increase security, and lower TCO of their Windows OS deployments.

Microsoft Corporation (www.microsoft.com) was founded in 1975, and is the worldwide leader in software, services, and Internet technologies for personal and business computing. This article was the collaborative work of the ADS Team within the Windows Server Group at Microsoft.

FOR MORE INFORMATION

ADS: http://www.microsoft.com/windowsserver2003/ads

Automated Deployment Services Technical Overview:
http://download.microsoft.com/download/1/f/2/1f2ba202-98f8-49b2-bd36-ef914e45636b/adsoverview.doc

Windows Server 2003:
http://www.microsoft.com/windowsserver2003/default.mspx

WMI: http://www.microsoft.com/whdc/hwdev/driver/wmi/default.mspx

Lightweight Disaster Recovery

for Windows- and Linux-based Environments

Organizations managing data centers separated within a metropolitan area have an alternative to wide area global clustering for disaster recovery. By building upon a Fibre Channel infrastructure, administrators can implement a lightweight disaster recovery strategy that maintains application and data availability during local data center failures and site disasters—providing an extra level of real-time availability without substantially increasing costs.

BY MICHELLE MOL AND JAMES GENTES

Many enterprises require a high availability and disaster recovery plan for their mission-critical applications and databases. Meeting such requirements under a constrained budget is possible if administrators can leverage existing hardware and infrastructure. Traditional approaches to disaster recovery range from standard, tape-based data recovery to synchronous, wide area data replication. Organizations without the infrastructure to support wide area replication can explore other ways to achieve rapid recovery of applications and data while protecting against site failure.

By using VERITAS Cluster Server™—either as a stand-alone product or in combination with other VERITAS® products such as VERITAS Volume Manager™—administrators can achieve high availability and disaster recovery in most data center environments. Three primary architectures form the foundation for application availability and disaster recovery in Microsoft® Windows® and Linux® operating system–based clusters: local area clusters, metropolitan area clusters, and wide area clusters.

Each of these approaches has specific advantages and disadvantages; ultimately, the choice is determined by a combination of infrastructure capabilities, affordability, acceptable data loss, and future planning. This article focuses on metropolitan area clustering, which provides high availability and lightweight disaster recovery within a radius of about 62 miles (100 kilometers).

The difference between local clustering and wide area disaster recovery is relatively clear: Local clustering ensures high availability by protecting against hardware and software failures in one data center location. In contrast, wide area disaster recovery is designed to protect against catastrophes that affect large areas by replicating data to distant locations. However, the challenge for IT organizations is how best to maintain budgets while accomplishing both high availability and disaster recovery in a deployment that utilizes the current infrastructure. By combining clustering with remote mirroring, these challenges can be tackled (see Figure 1).

Traditionally, disaster recovery sites have been located hundreds or even thousands of miles away from the data center to safeguard against far-reaching disasters. Now, more organizations are being challenged to deliver online application and data protection providing near-immediate recovery with reduced budgets.

Many organizations are evaluating cost-effective disaster recovery approaches that resemble local, high-availability clusters but span multiple sites within a range of approximately 62 miles. Such distances can provide disaster recovery protection against most physical threats, including fires, floods, and power outages.

For threats that reach beyond an entire metropolitan area, traditional tape backup and vaulting usually will suffice. To meet more stringent availability and recovery requirements over greater distances, organizations should deploy wide area clustering and replication over IP.

Local area clustering: High availability

Local high-availability clustering, based on shared disk architecture, should be the first method of deployment when server or application failure is the topmost concern. This method enables applications or databases to recover from failure using locally attached storage rather than relying on a remote copy. The standby host uses the same physical data set that was used by the primary host. As a best practice, failover always should be performed locally for application or server failure.

The local high-availability clustering environment consists of a redundant server, network, and storage architecture that facilitates application and data availability by linking multiple servers with shared storage. Nodes are linked with private heartbeats, usually Ethernet, that communicate the state status of the entire cluster. Each node in the cluster can access the storage of any other node, provided that all cluster components reside in the same data center location. The principal advantage of local area clusters is that applications and databases recover quickly by using data on shared storage, which helps to prevent data loss. The key drawback is that the data center is a single point of failure.

Metropolitan area clustering: Disaster recovery

Metropolitan area disaster recovery, or *campus clustering,* comprises a single cluster that stretches across two sites using Fibre Channel connectivity to facilitate data mirroring and cluster communication. Typically, organizations deploy this architecture to provide disaster recovery over short distances when a Fibre Channel storage area network (SAN) infrastructure is already in place.

Implementing a campus cluster will provide a lightweight form of disaster recovery for environments in which a traditional wide area disaster recovery approach using replication is either not suitable or too costly. Campus clusters eliminate both the hardware array and the physical building as a single point of cluster failure—effectively providing applications and data fault tolerance for nearly all failures except campus-wide disasters. Clustering and data mirroring components enable this type of configuration.

Using a product such as VERITAS Volume Manager, administrators can mirror data synchronously between two sites in a campus cluster configuration. Should the primary site fail, an exact copy of the data will reside at the secondary location.

Many VERITAS customers on Wall Street have deployed campus clusters using VERITAS Volume Manager to mirror data between Manhattan and Jersey City, New Jersey; VERITAS Cluster Server provides application and database availability. The metropolitan area cluster architecture allows these financial services organizations to meet government regulations and service level agreements (SLAs) for disaster recovery.

If a local disaster strikes, all the services—including applications, databases, and data—fail over from the affected site to a secondary building. An application or database will fail over locally at the primary site before failing over to the remote site. Consider an example in which three servers reside in building A and a fourth server resides in building B. If one server fails in building A, the application fails over to another server in building A. If building A is down, all running services in building A will fail over to building B. Remote mirroring helps ensure that a synchronous copy of the data is accessible at building B.

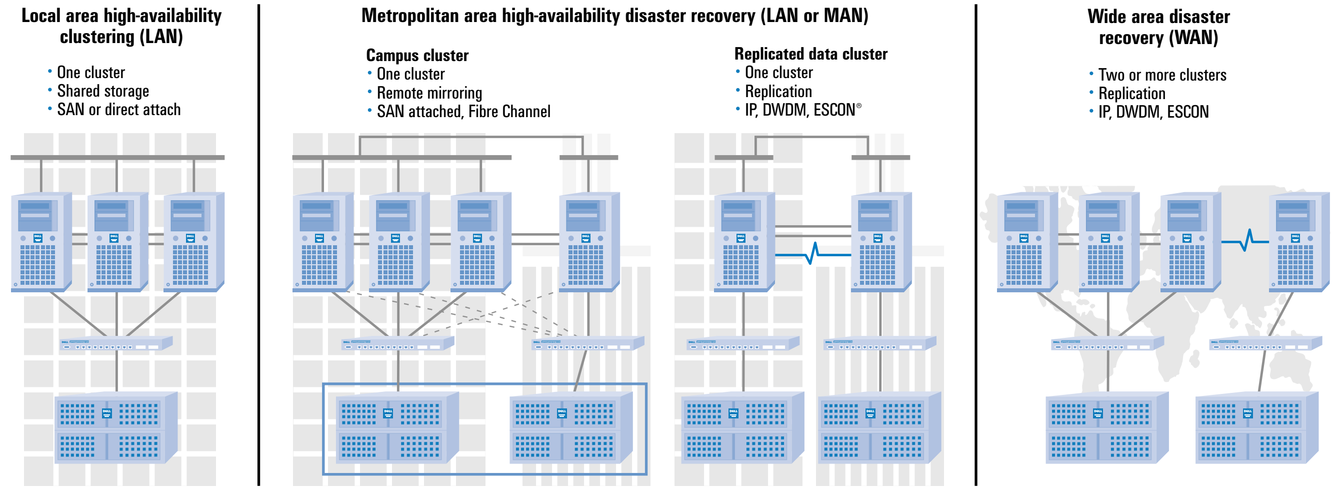


Figure 1. Cluster architectures for high availability and disaster recovery

Wide area clustering: Global coverage

Wide area disaster recovery provides the highest level of protection for data and applications, offering all the benefits of local and metropolitan area clustering while removing any distance limitations. Typically connecting sites using IP, wide area disaster recovery architecture requires two or more data centers. Should the primary site fail, all of its services and data are moved to a secondary hot site, which then becomes available to users.

Administrators must select secondary sites carefully. For example, the secondary site should not be located on the same fault line or power grid as the primary site, nor should it be near an airport or on the same power grid as an airport. In addition, the secondary site should be far enough away from the primary site to avoid weather patterns that could affect both locations. Several companies, including VERITAS, provide consulting services for deploying wide area disaster recovery.

Because the infrastructure, personnel, and hardware necessary to equip a remote site are costly, enterprises typically deploy wide area disaster recovery only when they are forced to comply with stringent government regulations or SLAs. As an option, organizations can leverage geographically dispersed office locations to implement wide area disaster recovery architecture. A remote office, already connected to the primary site over IP and equipped with servers and storage, can readily serve as the secondary hot site should the primary site fail.

Lightweight disaster recovery: Scalable protection

Best practices for lightweight disaster recovery start with tape backup and data redundancy using disk volume mirroring. Shared storage helps keep applications online when administrators deploy a scalable cluster infrastructure in conjunction with an automated disaster recovery plan that facilitates wide area data center migration. As business needs evolve, enterprises can add corresponding layers of availability, building upon a backup strategy that is already in place. By deploying products that enable lightweight disaster recovery, such as an integrated software suite from VERITAS, organizations can protect data and applications in their Windows- and Linux-based clusters at a minimal cost.

Michelle Mol (michelle.mol@veritas.com) is a product marketing manager for VERITAS Cluster Server. She has an M.S. in Technology Management from Pepperdine University.

James Gentes (james.gentes@veritas.com) is a senior product manager for VERITAS Availability Solutions on Windows and Linux platforms.

VERITAS Software Corporation (http://www.veritas.com) is the world's leading storage software company, providing data protection, application performance, storage management, high availability, and disaster recovery software.

LIGHTWEIGHT DISASTER RECOVERY ARCHITECTURE

A metropolitan area network environment includes the following components:

- One cluster stretched across multiple buildings, data centers, or sites
- Connections using a single subnet and Fibre Channel SAN
- Up to 32 nodes distributed freely among buildings, data centers, or sites
- Local storage mirrored between cluster nodes at each location
- Data switches using dense wavelength division multiplexing (DWDM) to enable distances up to 62 miles (100 kilometers)

Potential drawbacks include the relatively expensive cost of a Fibre Channel SAN infrastructure. In addition, latency of the storage networking infrastructure determines the distance between storage arrays. However, lightweight disaster recovery clusters offer the following important advantages:

- Local high availability within each site, as well as protection against site failure, using campus clustering
- Cost-effective, simple disaster recovery that does not require replication
- Remote mirroring to protect against data loss by copying data synchronously to both sites
- Minimal downtime for applications and databases using automatic or manual failover
- Existing Fibre Channel SAN infrastructures leveraged
- Look-and-feel of locally deployed cluster environment; no specialized configuration necessary
- Protection against the power grid as a single point of failure
- Data center expansion facilitated by creating one logical grouping of servers, storage, and applications across sites
- Failover across multiple networks allowed by providing Domain Name System (DNS) updates

FOR MORE INFORMATION

VERITAS: http://www.veritas.com

VERITAS Cluster Server: http://www.veritas.com/products/category/ProductDetail.jhtml?productId=clusterserver

VERITAS Volume Manager: http://www.veritas.com/van/products/volumemanagerwin.html

Exploring Windows Server 2003 Cluster Management

The Microsoft® Windows® Server 2003 operating system includes flexible, customizable, and powerful management tools. This article describes ways to automate cluster administration and improve cluster management.

BY ANANDA SANKARAN

The Microsoft® Windows® Server 2003 operating system provides powerful management capabilities that IT organizations can use to administer server clusters based on Microsoft Cluster Service (MSCS). By leveraging the tools and other utilities included with the Windows Server 2003 operating system and Resource Kit, administrators can formulate flexible management schemes customized for specific cluster administrator tasks. These tools are simple, but very powerful.

An enhanced command-line infrastructure is one key benefit of the Windows Server 2003 family. Windows Management Instrumentation (WMI), a component of the Windows operating system, provides a consistent means of accessing systems management information. The WMI Command-line (WMIC) utility allows administrators to perform WMI-related management tasks from the command line or through scripts and other management applications. Such capabilities give Windows Server 2003 the power and flexibility often associated with UNIX®-based systems—at a low total cost of ownership (TCO).

Generating SNMP traps for cluster-specific events

Simple Network Management Protocol (SNMP) traps are a standard mechanism for managed systems and devices to communicate system-specific events over the network to a management application. The management application can then take a suitable action, such as alerting an administrator, sending an e-mail, or paging a responsible party.

Windows Management Instrumentation, a component of the Windows operating system, provides a consistent means of accessing systems management information.

Administrators can configure the cluster-specific events logged into the Windows event log to generate SNMP traps, which are then forwarded to a management application. The evntwin command within Windows Server 2003 starts the Event to Trap Translator, a

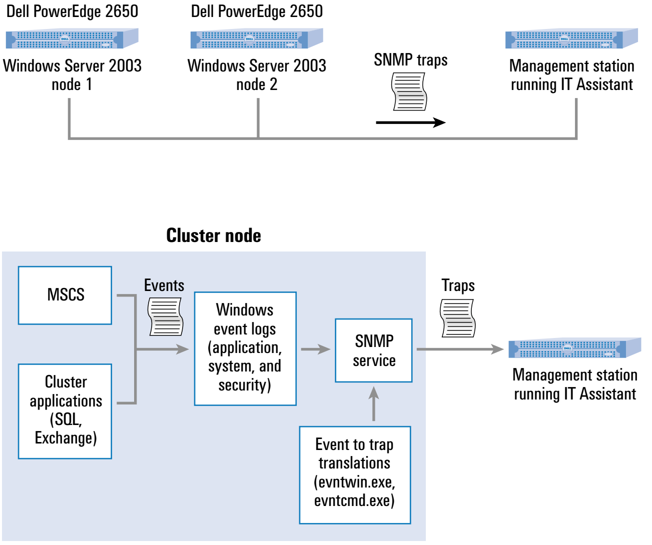

Dell PowerEdge 2650
Windows Server 2003 node 1
Dell PowerEdge 2650
Windows Server 2003 node 2
SNMP traps
Management station running IT Assistant
Cluster node
MSCS
Cluster applications (SQL, Exchange)
Events
Windows event logs (application, system, and security)
SNMP service
Event to trap translations (evntwin.exe, evntcmd.exe)
Traps
Management station running IT Assistant

Figure 1. Translation of cluster events to SNMP traps in a two-node cluster

graphical user interface– (GUI–) based utility that helps to translate Windows events to traps (see Figure 1). A command-line interface (CLI) version of this tool, called evntcmd, also exists. These utilities are installed as part of the SNMP service installation on Windows Server 2003. For more information, see Windows Server 2003 online help, available from the operating system GUI.

Administrators should use evntwin on all the cluster nodes to identify and register cluster events for translation into SNMP traps. Using evntwin on all the nodes is important because the cluster virtual server may reside on any node, which means that cluster events may be logged on different nodes at different times. Also, if cluster event replication is enabled (the default), every cluster event is logged into every node's event log—so all nodes generate SNMP traps for the same event. To prevent multiple SNMP traps from being generated, administrators should disable event replication.

Dell™ OpenManage™ IT Assistant, a management application with a rich feature set, can be configured to receive SNMP traps from the cluster. Administrators must also configure the SNMP service on each cluster node to send traps to the IT Assistant server station. To correctly interpret SNMP traps, IT Assistant requires further configuration; for more information, visit http://support.dell.com and download the Dell OpenManage IT Assistant User's Guide. Once configured, IT Assistant can send e-mail, pages, and alerts based on the SNMP traps sent to it.

Using WMI to manage clusters

WMI is an operating system component that manipulates and stores information about managed objects (see Figure 2). A *managed object* is a logical or physical system component—such as a hard drive, network router, database system, or cluster—that creates a data representation of an object. Management applications and scripts can use WMI to query and set management information on enterprise hardware and software components, providing a means to automate cluster administration tasks.

By using server cluster WMI provider classes, administrators can create scripts triggering specific cluster events to write information to an audit log in real time.

WMI derives from the Web-Based Enterprise Management (WBEM) initiative, a set of management and Internet standard technologies developed to unify the management of enterprise computing environments. The core component of the WBEM standard is a specification known as the Common Information Model (CIM).

CIM provides a uniform and consistent data description mechanism for any data provided by the managed objects. WMI includes a CIM object manager and a CIM object repository for manipulating and storing this data. A management application or script can obtain information regarding a managed object from the WMI infrastructure. For more information on WMI, see the Windows Server 2003 online help available from the operating system GUI or visit the Microsoft Developers Network (MSDN) at http://msdn.microsoft.com.

Exploring useful scripting resources

Windows Server 2003 contains a standard set of WMI classes and providers for handling managed objects of the Windows operating system. The Scripting API (application programming interface) for WMI facilitates developing quick, simple scripts for accessing these classes (see Figure 3). For example, administrators can write a

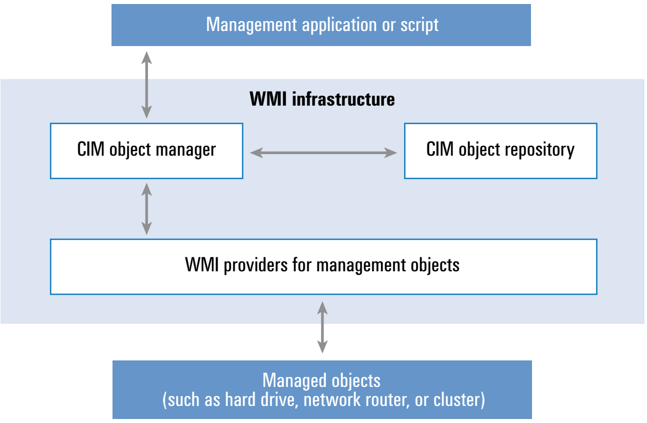

Management application or script
WMI infrastructure
CIM object manager
CIM object repository
WMI providers for management objects
Managed objects
(such as hard drive, network router, or cluster)

Figure 2. WMI architecture

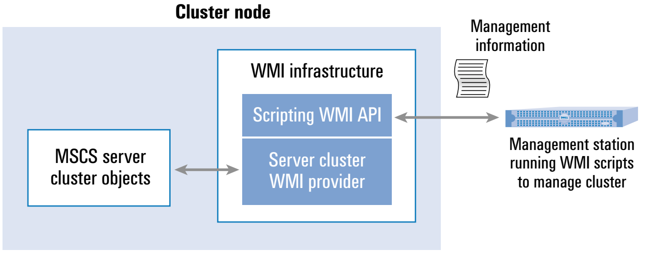

Cluster node
WMI infrastructure
Scripting WMI API
Server cluster WMI provider
MSCS server cluster objects
Management information
Management station running WMI scripts to manage cluster

Figure 3. WMI script management of MSCS clusters

script to manage a cluster using the server cluster WMI provider and classes.

Scripts written using Windows Script Host (WSH) or Microsoft Visual Basic® Scripting Edition (VBScript) can perform operations on file system objects, manipulate network printers, or change environment variables. Other useful scripting tools include wbemtest.exe, for checking the list of WMI classes on a system, and scriptomatic.exe, for developing WMI scripts. Scriptomatic.exe automatically recognizes WMI classes on a system and helps to include them in a script. Administrators also can use the wmic.exe utility to directly query WMI managed objects on a system. Both wmic.exe and wbemtest.exe are shipped with Windows Server 2003; for scriptomatic.exe, visit Microsoft Technet at http://www.microsoft.com/technet.

Writing an audit-logging or alert e-mailing script

Audit logs help to track all changes and problems associated with clustered components. These logs simplify discovering how changes occurred, understanding the changes, and monitoring or preventing the recurrence of those changes. By using server cluster WMI provider classes, administrators can create scripts triggering specific cluster events to write information to an audit log in real time. Administrators can determine the log format and the information to be written. Also, the same technique can be used for sending real-time e-mail alerts when crucial cluster events are generated. For examples, please see “Sample VBScript scripts for cluster management” in the online version of this article (http://www.dell.com/powersolutions).

Using the ClusterRecovery utility to recover a quorum disk

Cluster applications store data on shared disks accessible by all nodes in the cluster. The cluster configuration database itself and cluster checkpoint files are stored in a special shared disk called the quorum disk (see Figure 4). Administrators can recover the application data on shared disks by using standard data backup software, but recovering data on the quorum disk requires a different procedure.

The cluster checkpoint files (.cpt files) store the cluster resource configuration data in the registry of the cluster node owning the resource. The cluster service continually updates the checkpoint file from the corresponding registry data. During failover, these files are used to update the registry of an available node, so they must be consistent with registry data for proper operation of the cluster and applications.

A quorum disk recovered from standard backup may not contain the latest resource registry state in the checkpoint files, so the data in the registry could be overwritten by the recovered quorum data, leading to inconsistent resource states. After a failure, the checkpoint files should be re-created from the latest registry settings on the nodes, not from the quorum. However, the cluster database information and other data on the quorum can be recovered using standard backup procedures.

The ClusterRecovery utility in the Windows Server 2003 Resource Kit helps recover resource checkpoint files on the quorum and failed shared disks from the latest registry data that exists on the nodes (see Figure 5). Administrators also can use

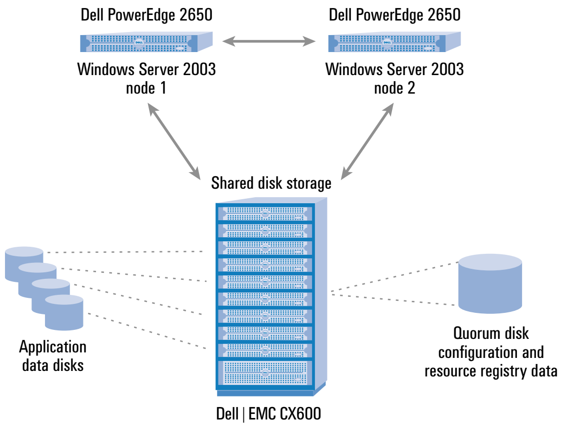

Dell PowerEdge 2650
Windows Server 2003 node 1
Dell PowerEdge 2650
Windows Server 2003 node 2
Shared disk storage
Application data disks
Quorum disk configuration and resource registry data
Dell | EMC CX600

Figure 4. Shared cluster disks in a two-node cluster

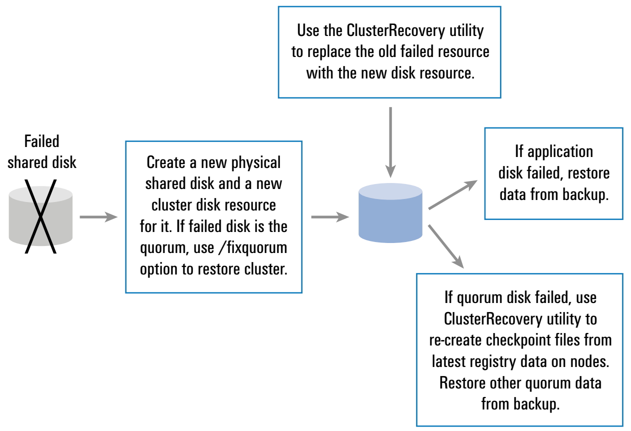

Use the ClusterRecovery utility to replace the old failed resource with the new disk resource.
Failed shared disk
Create a new physical shared disk and a new cluster disk resource for it. If failed disk is the quorum, use /fixquorum option to restore cluster.
If application disk failed, restore data from backup.
If quorum disk failed, use ClusterRecovery utility to re-create checkpoint files from latest registry data on nodes. Restore other quorum data from backup.

Figure 5. Cluster disk recovery procedure

this utility for re-creating a checkpoint file for a single resource, not just the whole cluster.

Following a disk failure in a cluster, administrators typically replace the failed physical disk and restore data files recovered from backup onto the new physical disk. However, before performing a restore on the new physical disk, administrators must create a logical cluster disk resource within MSCS that represents the new disk.

Using ClusterRecovery, administrators can include the new resource as the old failed resource and then delete the old resource representation of the failed disk. The utility automatically transfers the properties of the failed resource to the new resource. Any dependencies on the old resource are changed to point to the new resource. The new disk now has the same resource representation in the cluster that the failed disk did, and administrators can restore application data to the new disk from backup.

Taking advantage of native Windows Server 2003 tools

High-availability computer clusters running on standards-based hardware such as Dell PowerEdge™ servers using the Windows Server 2003 operating system offer many advantages to IT organizations, including cost-effective reliability, availability, and high performance. Managing these systems can be complex, but Windows Server 2003 provides several tools for simplifying cluster management. Using these tools, administrators can track and log system events, write scripts to generate alerts, and recover cluster disks safely. By taking advantage of native support in the Windows Server 2003 operating system, administrators can automate cluster administration tasks and gain more control over the clusters they manage.

Ananda Sankaran (ananda_sankaran@dell.com) is a software engineer in the High-Availability Cluster Development Group at Dell. His current interests related to high-availability clustering include cluster management, databases, and storage area networks (SANs). Ananda has a master's degree in Computer Science from Texas A&M University.

FOR MORE INFORMATION

Microsoft TechNet: http://www.microsoft.com/technet

MSDN: http://msdn.microsoft.com

WBEM: http://www.dmtf.org/standards/wbem

Windows Server 2003: http://www.microsoft.com/windowsserver2003

WMI FAQ: http://www.microsoft.com/windowsserver2003/community/centers/management/wmi_faq.mspx

Using Imaging to Deploy Microsoft Windows Server 2003

Imaging provides an efficient way to deploy Microsoft® Windows® Server 2003 on Dell™ PowerEdge™ servers. This article discusses Microsoft deployment tools and general concepts about creating image profiles, including techniques for preparing the source server as well as packaging, deploying, and testing the system image.

BY MATTHEW PAUL AND TRAVIS CAMPBELL

Drive imaging refers to the process of capturing the contents of a hard disk and saving those contents as a profile that can later be applied either to configure a different system or to restore the original system after a failure. Images enable organizations to avoid the time-consuming task of manually installing operating systems, custom applications, and drivers on each server. Administrators can create one or two master images that they apply to all servers through network transfer or CD installation. A uniform server profile helps ensure consistent system performance and simplify maintenance and subsequent upgrades.

Drive imaging provides several advantages, including the speed with which images can be applied and the ability to compress the contents of a drive to nearly half their actual size. Compression enables faster download times, because less data is sent over the network, and a smaller image takes up less valuable hard drive or CD space. Contents are uncompressed as the image is applied to target systems.

Planning a Windows Server 2003 image deployment

Several options exist for deploying Microsoft® Windows® Server 2003 operating systems on Dell™ PowerEdge™ servers. Administrators can specify the operating system (OS) to be preinstalled before delivery; they can install the OS manually on every server; or they can automate the installation process by deploying preconfigured images to servers (see Figure 1).

Whatever the deployment method, administrators must verify that the hardware of both source and target systems meets the requirements for the specific version of Windows Server 2003 they plan to install: Standard Edition, Enterprise Edition, Datacenter Edition, or Web Edition.[1] Software and drivers also must be compatible with Windows Server 2003. Managers can obtain PowerEdge BIOS, firmware, and drivers from the latest Dell OpenManage™ Server Administrator CD or online at the Dell support site (http://support.dell.com).

To successfully deploy preconfigured images of Windows Server 2003, administrators should also observe the following considerations:

- **Unattended mode:** Verify that the installation routines of the applications to be installed on target systems support an unattended mode. Both applications and drivers should be stored on a networked data share or on a CD to enable quick access when creating an image.

[1] For more information about specific hardware requirements for each version of Windows Server 2003, visit http://www.microsoft.com/windowsserver2003/evaluation/sysreqs/default.mspx.

- **File system:** Decide on the file system. The file allocation table (FAT), FAT32, and NT file system (NTFS) can be used with Windows Server 2003. NTFS is often the best choice because it supports larger disk drives, better recovery options for system crashes, local directory and file security, data encryption, and data compression.
- **Backup settings:** Save the computer settings and files of the target system onto another network server before deploying images.
- **Network load:** Determine whether the network can handle the increased traffic of downloading image profiles to the target systems. Imaging servers on the weekend or after hours may help minimize disruption.

Preparing the source server for imaging

The first step in the imaging process is to select and prepare an appropriate source server. The configuration of that server must be kept as simple as possible; most customization should occur on the target system, not the source. When deciding what components to install on the source server, administrators should take the following precautions:

- Confirm that the source server has similar hardware and a compatible hardware abstraction layer (HAL) to the target system.
- Note that the imaging of domain controllers is not supported. To deploy a domain controller, run dcpromo.exe on the target system after the image has been applied.
- Install any software that is dependent on Microsoft Active Directory® directory service or security ID (SID) on the target system, not on the source.
- Ensure that adequate licenses exist for all software being imaged, including (but not limited to) the OS and applications.

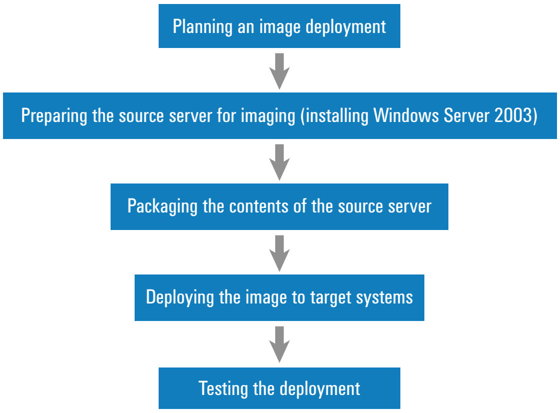

Planning an image deployment
Preparing the source server for imaging (installing Windows Server 2003)
Packaging the contents of the source server
Deploying the image to target systems
Testing the deployment

Figure 1. Imaging process overview

Images enable organizations to avoid the time-consuming task of manually installing operating systems, custom applications, and drivers on each server.

Installing Windows Server 2003

Administrators can install Windows Server 2003 on the source server in three different ways: manually; by booting from the Windows CD using an answer file; or by setting up a network share with the source server using an answer file. After installing the OS, administrators can prepare the source server for imaging. From this point on, any changes made to the source server will likely be captured and deployed to the target systems, so administrators should ensure that all drivers and application packages have been properly configured and stored on the source server before beginning the imaging process.

Creating images using factory mode

With the release of Windows XP, Microsoft introduced a new mode of image creation called factory mode, and Windows Server 2003 extends this functionality to servers. Factory mode uses the familiar Sysprep utility to create an image that can be deployed to vastly different target systems. An image created in factory mode is fully populated with all driver and application files needed for deployment. Once created, the image is copied onto the target system, where it runs through a factory-mode phase using winbom.ini to complete the installation of required applications and customizations. Afterward, the target system reverts to the familiar mini-setup sequence for the target system's first boot.

For images created in factory mode, winbom.ini contains all the configuration information needed to configure the OS—including the capability to add additional OS components; change various computer settings (such as display refresh, resolution, power management, and source path); specify the plug-and-play (PnP) path for drivers; manage application installs; and create new user accounts. Figure 2 shows a sample winbom.ini file.

Copying driver files

Administrators must copy all driver files onto the source server before they can install the drivers on the target system. Because the location of these files may vary, administrators should create a custom directory structure in which to copy all driver files. More than one folder can be specified in the PnP path; therefore, administrators may choose to create a directory structure that contains subfolders for each group of similarly configured target systems. For example, the folder C:\drivers\systemA could contain the drivers for systems of type A, while the folder C:\drivers\systemB could store the drivers for systems of type B.

```
[Version]
Signature="$Chicago$"
[Factory]
reseal=no
[ComputerSettings]
DisplayResolution=800x600x16
[PnPDriverUpdate]
TargetRoot=c:\
DevicePath=drivers
WaitForPnP=Yes
UpdateInstalledDrivers=Yes
[OEMRunOnce]
"Executing sysprep", "sysprep.standard", APP
[sysprep.standard]
InstallType = Standard
CmdLine = -quiet -reboot -nosidgen -reseal
SetupFile = sysprep.exe
SourcePath = c:\sysprep
```

Figure 2. Sample winbom.ini file

Integrating applications into the image

Once all the driver files are copied onto the source system, applications must be integrated into the image. Administrators can install an application in two primary ways. The first method involves copying files into a local directory on the source server and registering the application's unattended installation command line in winbom.ini. The application will be fully installed when the image is deployed to the target system, including specific information about the unique configuration of the target system. In the sample winbom.ini file in Figure 2, sysprep.exe is called to reseal the system. All application installations must be scheduled above this line in the `[OEMRunOnce]` section.

The second method is to install the applications directly onto the source server, a technique that works best when every target system receives the same applications. This approach reduces the time required to complete the setup on the target system and minimizes the complexity of managing the application installation in the winbom.ini file.

Realistically, most IT departments will use a combination of these two methods to integrate applications when creating an image. Often, certain applications will be installed on some target systems but not others. Each image contains only one winbom.ini file. To install different applications on a target system, the administrator can overwrite the existing winbom.ini file with one specific to a particular set of target systems. This technique allows administrators to install a wide variety of software configurations using only one image.

To install different applications on a target system, the administrator can overwrite the existing winbom.ini file with one specific to a particular set of target systems.

Enabling support for mass storage devices

After drivers and applications are integrated, administrators use the Sysprep utility to add support for mass storage devices in the image. Because the drivers for mass storage devices must be loaded before booting the target system, driver installation using standard PnP detection (described in the section "Copying driver files") will not work. Instead, administrators add support for mass storage devices by entering the command line `BuildMassStorageSection = Yes` in the `[Sysprep]` section of the sysprep.inf file. Doing so enables the system to support all natively supported mass storage drivers, and allows systems with supported mass storage devices to boot following image deployment.

Running the command line `sysprep.exe -factory -quiet -forceshutdown` causes the OS of the source server to automatically complete its preparation to enter factory mode. After finishing this process, the system powers off, marking the end of the preparation phase for the image build process.

Packaging the contents of the source server

At this stage, the source server should be fully populated with all drivers, applications, and OS customizations that the target systems require. The next task is packaging the contents of the source server's software configuration and storing that package in an accessible location before deployment. Many organizations package the software using a third-party software application designed to transform the software image that resides on the source server's hard drive into a compressed file set.[2]

Drives can be imaged in several ways. One method is simply to remove the hard drive from the source server and install it as a non-bootable disk on a secondary system. The imaging software can then run from the secondary system and generate the image from the source server hard drive. Other techniques include using a network boot disk or a bootable CD, or installing a new primary hard drive in the source server to boot the source server to a host OS. In any case, the result of the drive imaging process should be a software image, stored in one or more files, that can be deployed to target systems.

[2] For URLs to third-party imaging products, see "For more information" at the end of this article.

Deploying the image to target systems

The method of deployment is often closely related to the way the image files have been packaged. For example, if the image files were captured onto a network share, the easiest method of deployment might be over the network, using the network boot disk to copy the image files directly onto the hard drives of the target systems. By contrast, image files captured and stored on CD are installed most easily from that CD. If administrators have not yet installed hard drives in the target systems, they may opt to temporarily install the hard drive of each target system in a secondary system. Then they can perform the image installation from a hard drive image stored on the secondary system, provided that the OS boots for the first time on the target system. When a disk duplicator is available, the hard drive of the source server may be duplicated directly onto the hard drives of the target systems. Regardless of the deployment method chosen, the source software configuration will be copied directly onto the hard drives of the target systems.

Once the image has been fully deployed, a target system will boot back into factory mode when the end user first powers it on. If winbom.ini has been properly configured and placed into the C:\sysprep folder, this process will be completely automated: drivers and applications will install, and OS customizations will be applied. When the server completes this process, it will reboot and enter into a mini-setup sequence. As with earlier Microsoft Windows server operating systems, administrators can customize the sysprep.inf file to bypass some of the data collection screens. Once the end user completes the mini-setup sequence, the target system should be fully configured and customized.

Testing the deployment

Before rolling out the image on all target systems, best practices demand a pilot test to verify the image. The pilot should include most types of hardware from the production environment. Boot devices constitute a major area to authenticate, because the image must boot on every target system. Once a target system boots, all the individual hardware devices can be detected and applications installed with factory mode.

Administrators must verify that all of the new drivers are installed and check the driver properties for digital signatures. Versions for updated drivers that replaced native drivers can be confirmed in the Properties dialog box associated with the driver. Checking Device Manager—a graphical user interface (GUI) for viewing and changing the properties of devices installed in a computer—can confirm that every hardware device has a corresponding driver. Any bangs (that is, exclamation points) associated with corresponding drivers in the Device Manager indicate that those drivers are probably missing from the image.

The method of deployment often is closely related to the way the image files have been packaged.

By helping to reduce administrative overhead and increase deployment speed, drive imaging provides one of the most efficient, cost-effective ways to install Microsoft Windows Server 2003.

A pilot should include application verification. Whether the application was part of the original image or installed using factory mode, administrators should ensure that the application installed correctly and is working properly.

Providing an efficient, cost-effective deployment method

By helping to reduce administrative overhead and increase deployment speed, drive imaging provides one of the most efficient, cost-effective ways to install Microsoft Windows Server 2003. Deploying this operating system on dependable Dell PowerEdge servers can facilitate reliable operation and consistent results. Dell and Microsoft work closely together to help ensure that all drivers are digitally signed and each server is tested thoroughly to verify compatibility.

Matthew Paul (matthew_paul@dell.com) is an engineer for Factory Install Development in the Dell Enterprise Systems Group. He has a B.S. in Computer Information Systems from California State Polytechnic University at Pomona and is a Microsoft Certified Systems Engineer (MCSE).

Travis Campbell (travis_campbell@dell.com) is an engineer for Factory Install Development in the Dell Enterprise Systems Group. He has a B.A. from The University of Texas at Austin and is an MCSE.

FOR MORE INFORMATION

PowerQuest® Drive Image® imaging products:
http://www.powerquest.com/driveimage

Symantec® Norton Ghost™ imaging product:
http://www.ghost.com

Microsoft Windows Server 2003 deployment kit:
http://www.microsoft.com/windowsserver2003/techinfo/reskit/deploykit.mspx

Deploying a Scalable, Manageable Enterprise Web Site

Through efficient planning and the use of a Dell | EMC storage area network (SAN), Dell updated the intranet for its 40,000 employees worldwide with a highly scalable architecture. Dell enhanced the cost-effectiveness and manageability of its new intranet by automating both production Web server configuration and content publishing using the Microsoft® Application Center Web application deployment and management tool.

BY MARC MULZER

Just a few years ago, managing a corporate Web site was simple. Web masters developed content on their desktops and published directly to a single Web server, making content updates immediately available to users. Today, the scale and complexity of enterprise Web sites have increased drastically. Administrators must accommodate greater and more variable demands in load, rendering enterprise Web site planning and management a complex task. Web site production is now an interdependent team effort that involves graphic designers, content area managers, content authors, administrators, and content deployers.

Today, enterprise Web site architecture typically consists of one or more development servers, a staging environment, and multiple production Web servers. Staging servers function as a gateway through which code must pass before it enters production. A staging server contains a code repository database, from which content authors check out files using a system such as Microsoft® Visual SourceSafe® (VSS) version control software. Development work is performed on a development server, and once files are updated, content authors check in their updated files to the staging server.

Access to production Web servers is usually restricted to a few administrators. These individuals transfer updated files from the staging server to production Web servers at scheduled intervals (see Figure 1). Once files are posted to production Web servers, end users can view the new content.

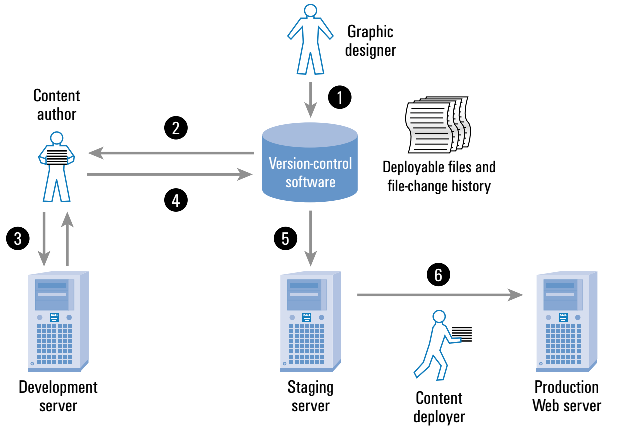
Web site production is now an interdependent team effort that involves graphic designers, content area managers, content authors, administrators, and content deployers.

Production Web servers comprise groups of identically configured servers that sit behind traffic load balancers. These Web server farms provide better reliability and scalability than a single server, and scaling out—the process of adding more servers to the farm—enables businesses to compensate for increases in intranet site traffic. However, the management challenge is threefold: how to provision additional servers quickly, maintain additional servers inexpensively, and ensure that server configurations are identical—all without raising IT costs.

Providing scalable storage with a SAN

To reduce administrative overhead and move toward a more centralized Web management paradigm, Dell began planning a new enterprise intranet in late 2002 to serve approximately 40,000 employees worldwide. The project—a 30 GB site consisting of about 400,000 documents—was completed in 2003. Dell enhanced both the scalability and cost-effectiveness of its intranet through an efficient Web site architecture that included a Dell | EMC storage area network (SAN).

Enterprise Web sites require storage systems for Web content that can grow easily as demand increases. While direct attach storage can offer better performance and reliability than large internal disk drives, the shared storage provided by a SAN can improve performance even more. In addition, SAN storage is designed to offer reduced management and provisioning costs compared to direct attach storage. Backups over a SAN are generally faster than backups involving individual drives, and SAN backups do not add to traffic on the LAN. SANs also increase scalability. Scaling out on a single SAN is easier than either increasing space on direct attach storage hardware or upgrading internal drives.



1. Create Web page templates
2. Check out Web pages
3. Manipulate Web pages
4. Check completed Web pages back in
5. Copy Web pages for approval and deployment
6. Publish code for production

Figure 1. Content development workflow

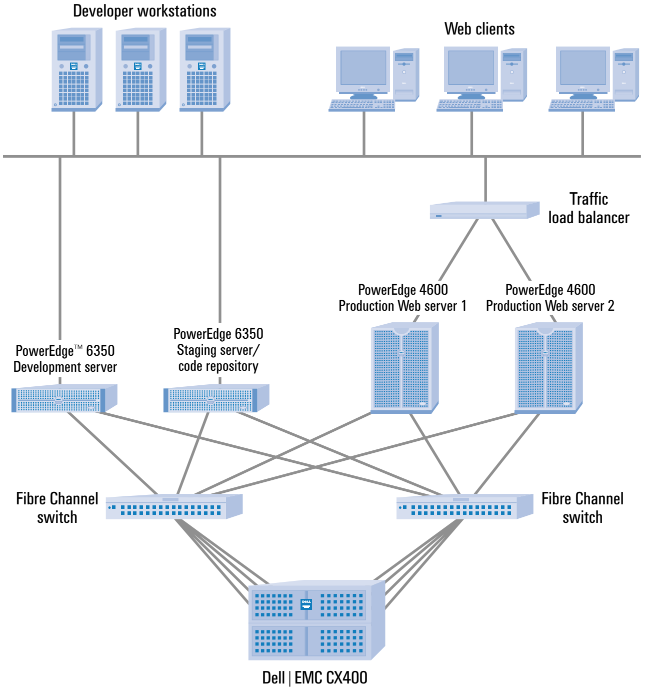

Developer workstations
Web clients
Traffic load balancer
PowerEdge 4600 Production Web server 1
PowerEdge 4600 Production Web server 2
PowerEdge™ 6350 Development server
PowerEdge 6350 Staging server/ code repository
Fibre Channel switch
Fibre Channel switch
Dell | EMC CX400

Figure 2. Dell intranet architecture

Dell upgraded its intranet using a Dell | EMC CX400 storage array for the SAN storage (see Figure 2). To efficiently configure and manage SAN storage, all servers on the SAN were provisioned with a 76 GB logical drive through EMC® Navisphere® storage management software.

The Dell™ intranet site comprised one development server, one staging server/code repository, and two production Web servers attached to the Dell | EMC storage array using redundant Fibre Channel switches. The production Web servers and staging server in the SAN were deployed as a high-availability Web cluster to enable centralized deployment of content through the staging server.

Automating configuration management

To simplify management and administration, servers in production Web server farms must be configured identically. However, the configuration of Web servers and the files on them tend to become inconsistent, or *drift,* over time, leading to inconsistent server configurations and files. Several factors contribute to server drift. For example,

Today, enterprise Web site architecture typically consists of one or more development servers, a staging environment, and multiple production Web servers.

CREATING A MANAGEABLE ENTERPRISE WEB SITE

Proper planning and organization before deployment can help enterprises create manageable Web sites. Web site architecture and workflow should be designed to reflect the structure of the content that will be developed and posted on the Web site. Content area managers, who are responsible for planning and conceptualizing enterprise Web sites, must consider content structure, content area access, and content development policies.

Content structure. Organizing content into manageable sections, often referred to as *content areas,* simplifies the deployment authorization and publishing process by granting a particular group of developers full access to a certain space within the structure of the Web site. The content area manager will create a code repository database on the staging server and assign corresponding content authors write access as needed.

Content area access. Once the structure of the content areas has been designed, the audience for each area must be defined, assembled into user groups, and granted access rights to the corresponding file structure. Every content area will have a Content Author group that has the right to make changes to existing documents and create new content. Each time these authors check in code, they have the option to deploy a copy of the file into an area where either content deployers or automated publishing processes can pick up the file and transfer the content to the live production site.

Content development policies. When several authors are working on the same site, organizations must plan and implement content development policies. Checking code in and out of a secured database helps ensure controlled access to the content. A version control system like Microsoft VSS helps ensure that no two content authors alter any given file at the same time. Such software also keeps an audit trail of content revisions.

secure environments have complex access-control structures that make it difficult for administrators to ensure manually that all copies of a particular file are identical on every Web server. Web site size also can exacerbate server synchronization issues. Enterprise Web sites may implement 5, 10, or even 20 servers to sustain the intense load created by thousands of daily hits. When very large content areas—20 GB or greater—and complex Web server configurations must be synchronized, human error also becomes a risk factor. Implementing reliable automated processes helps keep servers synchronized.

Setting up a production Web server farm as a cluster facilitates automation. Administrators can manage a complex cluster as a single virtual machine: software and configuration changes to a single server can be replicated automatically across all nodes in the cluster. Configuration management software handles synchronization across cluster nodes, thereby eliminating repetitive manual processes and significantly reducing both server maintenance time and the potential for human error.

Dell IT selected the Microsoft Application Center configuration management and content deployment tool to automate Web server synchronization. Application Center is installed on top of Microsoft Internet Information Services (IIS) on each server in the Web farm. The tool allows administrators to create a Web cluster that consists of one cluster controller and multiple cluster members. Instead of managing all servers in the cluster, administrators and content deployers work only with the cluster controller, whose configuration represents the desired state of the entire Web farm.

Application Center automatically detects changes on the cluster controller, either in the IIS configuration or in the file system, and propagates these changes to all cluster members. Despite this process, server drift may still occur if a cluster member is unavailable during synchronization because of network connectivity or hardware problems. In addition, server drift may develop if no changes are made to the cluster controller over an extended period, because no synchronization is initiated across the cluster members.

Dell enhanced both the scalability and cost-effectiveness of its intranet through an efficient Web site architecture that included a Dell | EMC SAN.

To apply necessary changes that may have been missed during synchronization, Application Center runs a replication process on the cluster controller at intervals scheduled by administrators. Together, the synchronization and replication processes help ensure that all Web servers in the cluster are configured identically.

Automating content publishing

Content publishing is another task that can be automated to help enterprise Web sites run more efficiently and keep management costs low. Previously, Dell required a team of engineers to manually initiate and manage the process of updating files from the staging server to production Web servers. These engineers scheduled update times to accommodate business needs. For example, one Dell group required updates at 15-minute intervals so that they could share content globally. However, any approach involving scheduled postings for a large number of content areas can incur substantial limitations:

- **Timeliness:** If scheduled updates occur hours apart, the enterprise cannot quickly provide up-to-date content.
- **Manageability:** If updates are scheduled too frequently, they quickly can become too numerous to monitor and manage effectively with available staff.
- **Quality control:** If incorrect content is published accidentally, content authors must wait until the next replication before they can overwrite their errors.
- **Efficiency:** If incorrect content must be corrected immediately, content authors must create an urgent help desk ticket or manually request that administrators or content deployers make changes manually. Both of these options add to administrative overhead and increase the risk of human error.

Enabling content authors to trigger the publishing process for their own files, rather than waiting for administrator-scheduled updates, can help resolve these issues. Content authors determine when a file is properly updated and ready for publication; therefore, the Web site infrastructure must link content authors to the content deployment process. If the staging server is set up to be the Application Center cluster controller, content authors can trigger deployments themselves, without the aid of administrators.

At Dell, content authors check out files from the staging server, using VSS, and update file content on the development server. When done, content authors check code back into the staging server database. As files are checked back in, the VSS client provides an option to simultaneously copy the file to a shadow directory listed in the content area's .ini file. If the shadow directory is configured as a root directory for the enterprise Web site on the staging server, Application Center detects changes to each file and immediately pushes the new content out to all the cluster members.

After new content is deployed to the Application Center cluster controller by content authors, it becomes available on each cluster member momentarily–delivering Web site updates in near–real time.

Scaling enterprise Web sites flexibly and cost-effectively

By implementing Application Center, Dell IT was able to retire hundreds of scheduled update jobs and eliminate practically all manual content deployments for its intranet. After new content is deployed to the Application Center cluster controller by content authors, it becomes available on each cluster member momentarily—delivering Web site updates in near–real time. Automated configuration and provisioning using Application Center also has helped increase the efficiency of the IT organization. Dell estimates that Application Center has helped to reduce server configuration time substantially, enabling fast response to variable load demands and enhanced consistency in configurations across cluster members.

By providing a fast and simple way to handle additional traffic, Dell | EMC SAN storage has helped improve the flexibility of the Dell intranet. Through effective planning, manageable storage, and the automation of configuration and content publishing, Dell has cost-effectively built a highly manageable intranet that can scale on demand to meet the dynamic business needs created by today's fast-changing markets.

Marc Mulzer (marc_mulzer@dell.com) is a systems engineer in Server and Storage Systems Engineering–Web Technologies at Dell. Marc works with Web technologies to architect scalable, highly available, and cost-effective Web application solutions. He has a B.S. in Computer Science from the College of Advanced Vocational Studies in Mannheim, Germany, and is a Microsoft Certified Systems Engineer (MCSE).

FOR MORE INFORMATION

Dell: http://www.dell.com

Microsoft: http://www.microsoft.com

Building a Highly Scalable and Available Data Environment for Oracle9*i* RAC

To provide a highly scalable and available database environment for Oracle9*i*™ Real Application Clusters (RAC), administrators must establish a highly available storage infrastructure. A storage area network (SAN) can provide redundant paths to storage, and running EMC® PowerPath® can leverage this redundancy by providing a mechanism for path failover in an Oracle9*i* RAC infrastructure.

BY ZAFAR MAHMOOD, PAUL RAD, AND ROBERT NADON

In today's business environment, high availability is required for mission-critical applications. The Oracle9*i*™ relational database management system (RDBMS) is highly available and scalable. The Oracle9*i* Real Application Clusters (RAC) option enables the Oracle9*i* RDBMS to be configured in a cluster database architecture where multiple nodes share the same storage. Oracle9*i* RAC provides high availability; if one node fails, the others take over and provide uninterrupted access to the database. However, if only one I/O path exists from each node to the shared storage in such an environment, this I/O path potentially becomes a single point of failure.

A typical Oracle9*i* configuration includes a storage area network (SAN), which can help provide a highly available data infrastructure by using redundant components to ensure that no component becomes a single point of failure. A Fibre Channel–based SAN fabric supports multipath routing between SAN switches. In a typical topology, a node has multiple Fibre Channel host bus adapters (HBAs), each of which are connected to the same SAN, resulting in multiple paths to the same device (see Figure 1). SAN storage devices can also accept multiple Fibre Channel connections.

Redundant paths in a SAN provide failover capability when any component in the data path fails. Multiple paths also enhance efficiency, allowing administrators to load balance SAN traffic by considering I/O across all available paths. In doing so, the SAN takes advantage of the additional bandwidth provided by each physical connection.

Although redundant I/O paths are beneficial for load balancing and link failover, they can create complications. Because each device on a SAN appears as a SCSI ID on each HBA that is connected to the SAN, a system with multiple HBAs connected to the SAN will behave as if each device on each path is a separate SCSI device. Thus, the operating system behaves as though multiple storage resources exist when in fact there are only multiple paths to the same resource.

Pointing to the same storage device along different paths could potentially cause data corruption and system crashes. To prevent such problems, administrators can install path management software, such as EMC® PowerPath®, on each node in the SAN. PowerPath enables multiple I/O paths to the shared storage by masking the paths, and presents the operating system with the appearance of a single SCSI connection. This masking ensures that the node receives a single view of the storage devices across multiple HBAs. PowerPath also automatically detects available paths, restores failed paths, and load balances I/O across all paths. When integrated in an Oracle9*i* RAC environment, PowerPath provides highly available, scalable, and fault-tolerant shared storage.

Using PowerPath for path failure detection

EMC PowerPath helps provide high availability by automatically detecting and restoring failed paths while storage arrays, nodes, and applications remain available. Should an HBA, a storage processor, or a cable fail, PowerPath completes an I/O request through another available channel, helping prevent the interruption of data to an application. PowerPath also provides automatic online path recovery after the path is repaired, which can reduce planned outages to restore services.

PowerPath virtual devices map paths to storage

PowerPath resides on a node as a software component between Oracle9*i* RAC, the Oracle® cluster file system (OCFS), and the HBA device driver layer (see Figure 2). PowerPath operates independently of applications, the RDBMS, management utilities, and file systems, allowing administrators to install and configure PowerPath without modifying existing software.

The PowerPath driver resides on the node, above the HBA driver. The node has multiple HBAs so that it can provide path failover through the PowerPath driver. The PowerPath driver enables virtual devices, which provide failure-resistant and load-balanced paths to the Dell | EMC storage array. An application references a PowerPath virtual device; in turn, the PowerPath driver manages path allocation to the storage array.

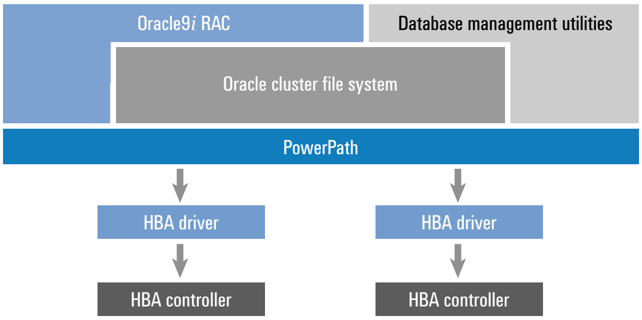


Figure 2. PowerPath in Oracle9*i* RAC software layer

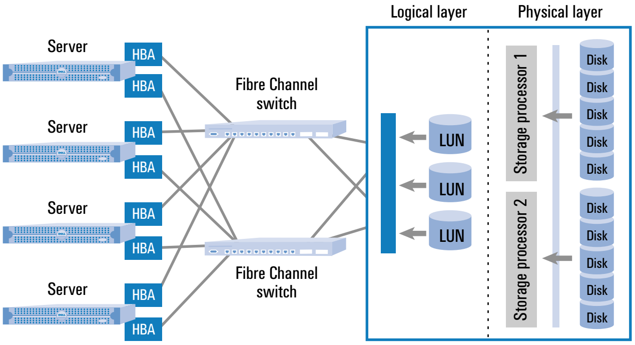


Figure 1. A highly available storage infrastructure

In the following example, four logical unit numbers (LUNs) are configured on a Dell | EMC storage array, which uses two storage processors to connect—through two separate Fibre Channel switches—to a node containing two HBAs. The resulting paths and their mappings to the Power Path virtual devices provide a total of 16 paths from a node to storage.

Having 16 paths to storage would ordinarily result in 16 logical devices being visible to the node. However, the PowerPath driver creates four PowerPath virtual devices, and each of these virtual devices maps four paths to a logical device on the storage array, as indicated in Figure 3.

Powermt management utility facilitates path management

For automatic failure detection and recovery, PowerPath provides an administration utility called Powermt, which provides a command

PowerPath virtual devices (4)	Paths (16)
/dev/emcpowera	/dev/sdb /dev/sdf /dev/sdj /dev/sdn
/dev/emcpowerb	/dev/sdc /dev/sdg /dev/sdk /dev/sdo
/dev/emcpowerc	/dev/sdd /dev/sdh /dev/sdl /dev/sdp
/dev/emcpowerd	/dev/sde /dev/sdi /dev/sdm /dev/sdq

Figure 3. Paths to logical devices for PowerPath virtual devices

line interface to the PowerPath environment. PowerPath periodically tests the paths for failure detection according to a built-in algorithm. Using the Powermt utility, administrators can set up a host node to perform autorecovery on failed paths by using the following command-line instruction to the PowerPath driver:

```
powermt set periodic_autorestore=on|off
```

For convenient management of a UNIX® or Linux® node configured with PowerPath, the Powermt utility provides several features, including:

- **Checking a PowerPath configuration:** The `powermt check` command checks the specified paths and, if desired, removes from the PowerPath configuration any paths marked dead.
- **Configuring paths to logical devices:** The `powermt config` command configures all detected logical devices as PowerPath devices and adds these devices to the PowerPath configuration, creating devices as required.
- **Removing paths from PowerPath management:** The `powermt remove` command deletes the specified path (or paths) from PowerPath's list of configured paths. It does not delete the logical device to which the paths refer.

Integrating PowerPath in an Oracle9*i* RAC environment

To perform I/O, Oracle9*i* RAC uses PowerPath virtual devices (such as /dev/emcpowera and /dev/emcpowerb, shown in Figure 3) instead of LUNs, or logical devices (such as /dev/sdb and /dev/sdc). If an actual device path fails, the PowerPath driver routes I/O to an alternative path without causing any interruption to the RDBMS functionality.

Integration of PowerPath with Oracle9*i* RAC is a simple process. PowerPath may be deployed in either a new or existing Oracle9*i* RAC implementation, as explained in the following sections.

Integrating PowerPath with a new Oracle9*i* RAC implementation

Oracle9*i* RAC uses shared storage on the SAN for its database, redo log, and control files. The RDBMS engine from each node in the cluster must have direct access to this storage to create or update any of these files. Without PowerPath, the RDBMS engine would directly access the LUNs on the shared storage to create the required files. With PowerPath integrated, the procedure is the same, with one important exception. The partitions and OCFS are created on the PowerPath virtual devices rather than on the partitioned LUNs:

SANs provide centralized data storage, and EMC PowerPath complements SAN architecture by helping manage redundant SAN paths to provide a high-availability environment for Oracle9*i* RAC.

1. Partition the PowerPath virtual devices on the shared storage array according to the database sizing requirements:
```
fdisk /dev/emcpowera
fdisk /dev/emcpowerb
fdisk /dev/emcpowerc
fdisk /dev/emcpowerd
```
2. Create the OCFS on the new partitions and mount the file system:
```
mkfs.ocfs -F -b 128 -L u01 -m /u01 -u 200 -g 300
   -p 0775 /dev/emcpowera1
mkfs.ocfs -F -b 128 -L u02 -m /u02 -u 200 -g 300
   -p 0775 /dev/emcpowerb1
mkfs.ocfs -F -b 128 -L u03 -m /u03 -u 200 -g 300
   -p 0775 /dev/emcpowerc1
mkfs.ocfs -F -b 128 -L u04 -m /u04 -u 200 -g 300
   -p 0775 dev/emcpowerd1
```
3. Set up the clusterware on the OCFS, and create the Oracle9*i* RAC database on the shared SAN storage.

The integration of PowerPath is transparent to the Oracle9*i* RAC database engine, as is the fact that each PowerPath virtual device points to multiple physical I/O paths.

Integrating PowerPath with an existing Oracle9*i* RAC implementation

In an existing Oracle9*i* RAC database that is set up without PowerPath, the database engine uses the LUNs on the shared SAN storage to create and update the database files. Integrating PowerPath is a straightforward process using the following steps:

1. Shut down all Oracle services in the cluster, including database listeners, Oracle Intelligent Agents, and Oracle Cluster Manager.
2. Shut down the database on all cluster nodes.
3. Use the `umount` command to unmount all OCFS volumes.
4. Configure the Dell|EMC storage array to support PowerPath, and install the PowerPath software on all cluster nodes. On each node, the PowerPath software will automatically create PowerPath devices that point to the existing LUNs.
5. Modify the /etc/fstab configuration file and replace the device names of each LUN (for example, /dev/sdb, /dev/sdf, dev/sdj, /dev/sdn) on the storage array with the PowerPath virtual device name (for example, /dev/emcpowera, /dev/emcpowerb, /dev/emcpowerc, /dev/emcpowerd).

6. Restart the OCFS and use the `mount` command to mount the file system.
7. Start up Oracle services, clusterware, and Oracle9*i* RAC database instances on all cluster nodes.

Using PowerPath to maintain availability

As reliable access to information becomes a critical mission for today's data centers, IT architects and administrators must use centralized, scalable storage to help create a highly available data infrastructure across the enterprise. SANs provide centralized data storage, and EMC PowerPath complements SAN architecture by helping manage redundant SAN paths to provide a high-availability environment for Oracle9*i* RAC.

Factors such as I/O load, effect of downtime, and availability of administrator maintenance time can help determine whether an organization's networked Dell | EMC storage would benefit from PowerPath. PowerPath offers fault tolerance to help eliminate downtime, load balancing to enable more efficient I/O traffic, and automatic detection and repair of paths that have failed, thereby helping to reduce administrative overhead. By integrating PowerPath with Oracle9*i* RAC, administrators can improve SAN uptime and provide a stable infrastructure for mission-critical data.

Zafar Mahmood (zafar_mahmood@dell.com) is a software engineer in the Dell Database and Application Engineering Deployment Department of the Dell Product Group. He has been involved in database performance optimization, database systems, and database clustering for more than six years. He is currently working on Oracle9*i* RAC implementations at Dell. Zafar has an M.S. in Electrical Engineering with a specialization in Computer Communications from the City University of New York.

Paul Rad (paul_rad@dell.com) is a senior software engineer in the Dell Database and Application Engineering Department of the Dell Product Group. Paul has master's degrees in both Computer Science and Computer Engineering from the University of Texas at San Antonio.

Robert Nadon (robert_nadon@dell.com) is a software engineering consultant in the Factory Installation Development Group of the Dell Product Group. He focuses on factory installation for UNIX-based (Linux and BSDi) operating systems and for the Oracle9*i* database. Another area of interest is Dell software RAID offerings. Robert has a B.S. in Computer Engineering from Texas A&M University. He is a Red Hat® Certified Engineer (RHCE) for Red Hat Linux 6.2 and 7.0.

FOR MORE INFORMATION

Dell | EMC storage: http://www.dell.com/emc

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

Oracle database: http://www.oracle.com/oracle9i

Using Red Hat Enterprise Linux AS to Achieve Highly Available, Load-Balanced Clusters

The Red Hat® Enterprise Linux® AS operating system integrates Cluster Manager and IP Load Balancing, features that improve cluster functionality. This article describes how these features can be combined with Dell™ PowerEdge™ servers and other components to achieve high availability and high performance in clusters.

BY DANNY TRINH

In response to growing demands for scalability, reliability, and serviceability, system administrators frequently must provide high-availability, high-performance clustering solutions for their existing networks. The Red Hat® Enterprise Linux® AS (formerly Red Hat Linux Advanced Server) operating system includes two types of integrated clustering functionality: Cluster Manager for high-availability clusters and IP Load Balancing for redundant, load-balancing clusters.[1] Administrators can combine these components with Dell™ PowerEdge™ servers, Dell PowerVault™ storage, Dell PowerConnect™ switches, Dell Fibre Channel switches and host bus adapters (HBAs), and high-speed network interface cards (NICs) to build clusters that meet the scalability and availability requirements of high-end, enterprise-class applications.

Using Cluster Manager to implement high-availability clusters

The minimum hardware requirements for a high-availability cluster of Dell PowerEdge servers running on Red Hat Enterprise Linux AS include:

- Two PowerEdge servers (acting as virtual servers), each with one NIC and one Fibre Channel HBA
- One Dell Fibre Channel array and one Fibre Channel switch
- One PowerConnect switch

Red Hat Enterprise Linux AS includes Cluster Manager and all other software and utilities required to implement high-availability clusters, such as the two-node failover cluster shown in Figure 1. High-availability

[1] IP Load Balancing is often known by its project name, Piranha.

clusters primarily use shared storage; therefore, both nodes connect to a Fibre Channel storage array. Fibre Channel offers superior performance and reliability for high-availability computing. To achieve failover, administrators can configure nodes as active/active, in which both nodes host application services, or as active/passive, in which one node hosts application services and the other is a backup waiting for failover.

Configuration of active/passive and active/active failover modes

In active/passive mode, the sole function of the backup node is waiting to take over when the primary services node fails. Although not necessary for cluster operation, ideally all nodes will have identical hardware so that clients cannot detect a difference after failover.

In active/active mode, both nodes provide services to clients, but not the same services. For example, given two network services such as Network File System (NFS) and Server Message Block (SMB) file sharing, node 1 might provide NFS and node 2 might provide SMB. If node 1 fails, node 2 takes over NFS file sharing while continuing to provide SMB. In active/active mode, services experience a performance drop when one node fails because the other node must perform twice the work.

Prevention of false failover

Red Hat Cluster Manager uses quorum partitions on shared storage as the primary mechanism for recording the state of cluster nodes.

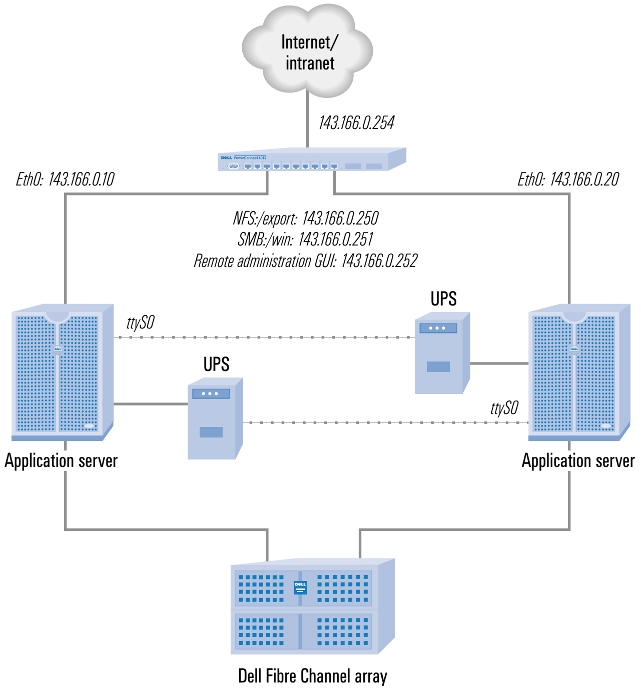


Figure 1. Two-node high-availability computing cluster architecture

> Red Hat Cluster Manager uses quorum partitions on shared storage as the primary mechanism for recording the state of cluster nodes.

Each node periodically sends update information to quorum partitions that tells the other node it is still up and running.

Additionally, each node sends a *heartbeat*—a periodic signal that indicates it is still running. Cluster Manager uses the heartbeat as a delay mechanism to prevent false failover. For example, if a cluster node experiences kernel panic, the node cannot send update information or heartbeat signals—a legitimate case for failover. However, if a cluster node is so busy with I/O-intensive tasks that it cannot send update information as normal, Cluster Manager uses the heartbeat signal—which the busy node can still send—to determine whether failover should wait for a few more seconds. These few seconds can make the difference between false failover and continuous service.

Cluster Manager provides two additional methods for preventing false failover and safeguarding data integrity: Shoot The Other Node In The Head (STONITH) and the watchdog timer. Although very unlikely, a failed node may continue to write data even though it no longer sends a heartbeat signal. Only one node can mount a local file system read-write at a given time; therefore, the working node must shut down the failed node to prevent any data corruption. STONITH allows a node to reset the power on a malfunctioning node, forcing it to reboot by turning off the errant node's uninterruptible power supply (UPS). Turning off the UPS prevents the dead node from sending update information or a heartbeat signal.

The watchdog kernel module (softdog.o) can be loaded when configuring Cluster Manager. With this module loaded, the system attempts to write to /dev/watchdog at certain polling intervals. If the system fails to do so, the kernel sends a power cycle message signal to the BIOS. The server then reboots itself when an administrator presses the reset power button.

Using IP Load Balancing to implement redundant, load-balancing clusters

The minimum hardware requirements for a load-balancing cluster of Dell PowerEdge servers running on the Red Hat Enterprise Linux AS operating system include:

- Two PowerEdge servers, each with two NICs, to act as active and backup routers
- Three PowerEdge servers, each with one NIC, to act as application servers
- Two PowerConnect switches

To implement IP load-balancing clusters, no additional software is needed; Red Hat Enterprise Linux AS has all necessary software and utilities, such as IP Load Balancing. This utility is an enhancement of the Linux Virtual Server (LVS), which uses the Network Address Translation (NAT) routing mechanism to deliver high availability and scalability for applications and services. Direct routing and IP encapsulation (tunneling) are alternative routing mechanisms for LVS, but Red Hat Linux does not support them.

A cluster configured for IP Load Balancing has two layers (see Figure 2). The first layer, which provides high availability, contains an active and a backup router. These routers are connected to two separate networks: public and private. The active router serves as a NAT router. All network traffic to and from both Internet and intranet users must pass through the active router on its way to and from the application servers (also known as real servers). Clients contact the IP Load Balancing cluster using the public virtual IP address (in Figure 2, 143.166.0.1) to request network services. At any one time, only one of the load-balancing routers is active, and it has both virtual IP addresses assigned to it. When the active router fails, the backup router detects this, takes over the virtual IP addresses, automatically promotes itself to NAT router, and becomes the active router.

Administrators can specify a scheduling algorithm for the load-balancing router to divide the network load among application servers. By default, IP Load Balancing uses a weighted least-connection scheduling algorithm. However, administrators can choose among many scheduling algorithms for one that best fits their specific network requirements. Both load-balancing routers also have unique IP addresses assigned to their public and private interfaces so they can monitor each other's health.

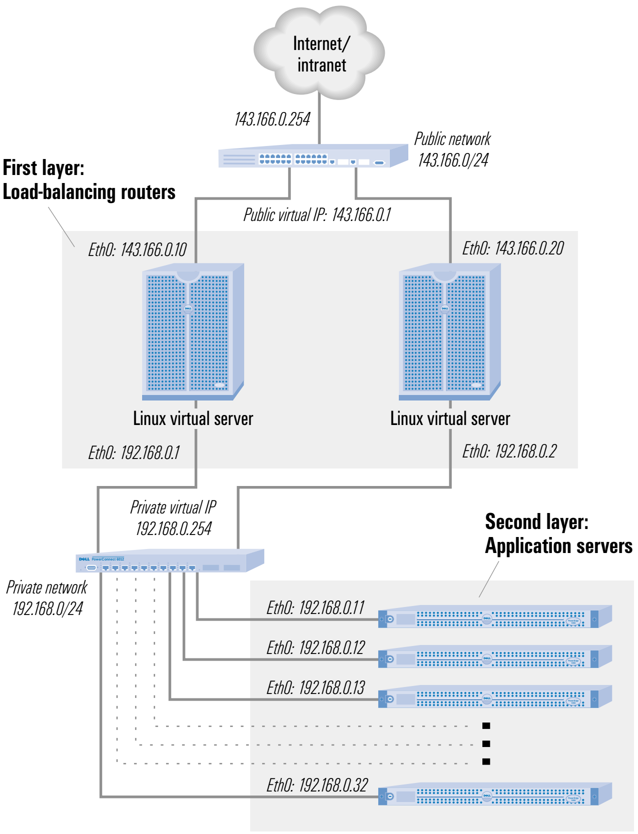


Figure 2. IP Load Balancing cluster architecture

A cluster running on Red Hat Enterprise Linux AS that combines Cluster Manager and IP Load Balancing functionality can provide excellent network services to clients.

The second layer in this configuration contains from 2 to 32 application servers connected to a private network. These servers provide such functions as Web and FTP serving. The application servers use one private virtual IP address—192.168.0.254 in the example shown in Figure 2—as their default route to send responses back to the clients; thus, the cluster appears as one server. Each application server must have a unique private IP address.

Utilities to set up IP Load Balancing

The Piranha Configuration Tool and the `iptables` command are the two main utilities needed to set up IP Load Balancing. Key software components of the Piranha Configuration Tool include:

- **pulse:** This controlling process runs on the load-balancing routers, monitors the network heartbeat, and controls health monitoring and router failover.
- **lvs:** This daemon runs on the active router and calls the ipvsadm service to keep the LVS redirection table up-to-date.
- **nanny:** This monitoring daemon runs on the active router to monitor health for each application server and virtual service.
- **piranha-gui:** This service starts a Web-based graphical user interface (GUI) for generating LVS configurations and monitoring the state of the IP Load Balancing cluster.

The `iptables` command sets firewall marks on packets for any of the multiports that are required by network services. For example, HTTP and HTTPS must be the same on the application server to which a client connects.

IP Load Balancing can be used for many network services that do not change data frequently, such as static Web sites, read-only

FTP servers, or video-streaming servers. Because each application server has its own storage, all data on one application server must be the same as that on the other application servers. Like Linux, IP Load Balancing does not require specialized hardware; all Dell PowerEdge servers are excellent choices for implementing this functionality.

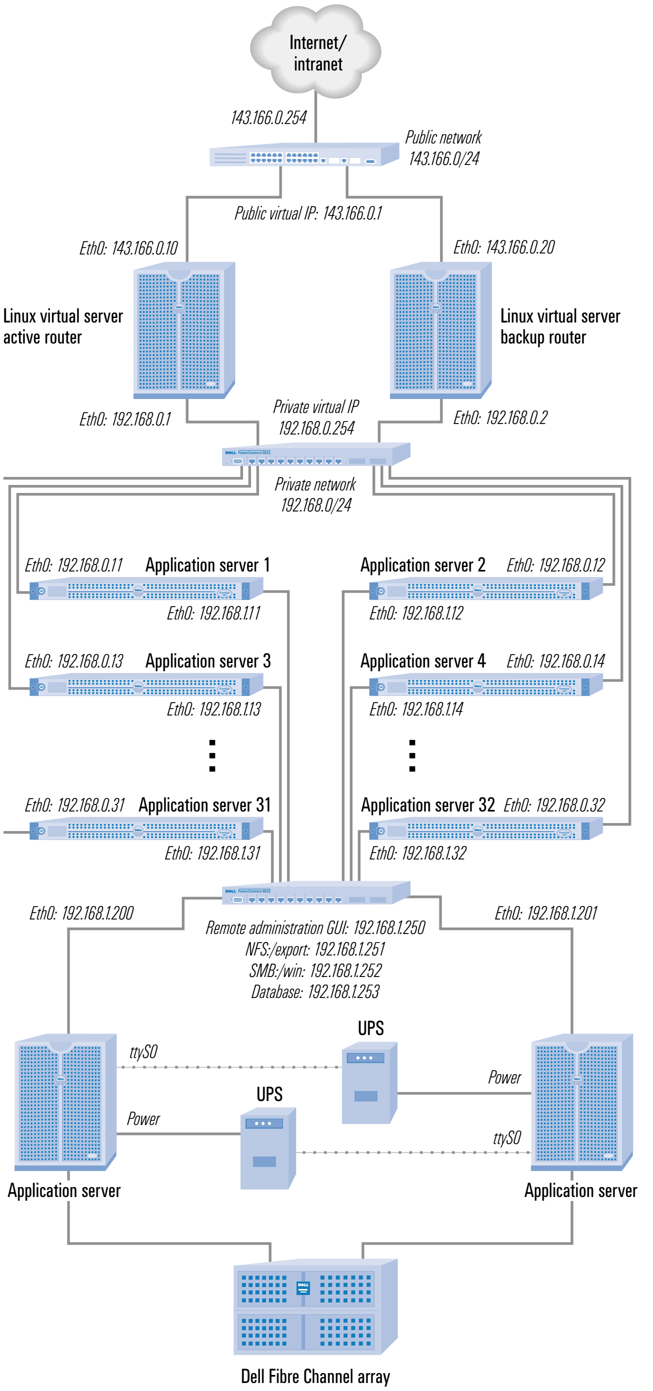


Figure 3. High-availability computing cluster with IP Load Balancing and Cluster Manager

Achieving highly available, scalable environments

High-availability clusters can help decrease downtime, save money, and increase productivity. Load-balancing clusters distribute incoming IP network requests across multiple servers, helping to increase performance, scalability, and fault tolerance of the network. Integrating both approaches into one cluster reaps the advantages of both high availability and load balancing. A cluster running on Red Hat Enterprise Linux AS that combines Cluster Manager and IP Load Balancing functionality can provide excellent network services to clients. Figure 3 shows an example configuration. To system administrators, such a configuration—providing a single place for network services—is easy to maintain. To users, this cluster configuration appears as a single virtual server.

> Load-balancing clusters distribute incoming IP network requests across multiple servers, increasing performance, scalability, and fault tolerance of the network.

By combining Red Hat Enterprise Linux AS, Dell PowerEdge servers, Dell PowerVault storage, Dell PowerConnect switches, and high-speed NICs, IT organizations can achieve highly available, scalable environments for services such as Domain Name System (DNS), mail, proxy, NFS share, SMB share, and Dynamic Host Configuration Protocol (DHCP). These components also can provide the foundation for clusters running applications that require scalability and high availability, such as e-commerce Web sites, Web farms, and application centers.

Danny Trinh (danny_trinh@dell.com) is a senior analyst in the Linux Development Group at Dell. He tests and certifies all Dell PowerEdge servers and peripherals for compatibility with Red Hat Linux. Danny has an associate degree and is a Certified Novell Engineer® (CNE®), Microsoft® Certified Systems Engineer (MCSE), and Red Hat Certified Engineer® (RHCE®).

FOR MORE INFORMATION

iptables: http://www.redhat.com/docs/manuals/linux/RHL-9-Manual/ref-guide/ch-iptables.html

Linux Virtual Server Project: http://www.linuxvirtualserver.org

Piranha Configuration Tool: http://www.redhat.com/docs/manuals/enterprise/RHEL-AS-2.1-Manual/install-guide/ch-lvs-piranha.html

Red Hat Cluster Manager Installation and Administration Guide: http://www.redhat.com/docs/manuals/enterprise/RHEL-AS-2.1-Manual/cluster-manager

Red Hat Enterprise Linux AS: http://www.redhat.com/software/rhel/as

Failure Modes and Effects Analysis of Oracle9*i* RAC

The failure of a mission-critical Oracle9*i*™ Real Application Clusters (RAC) database can spell disaster for the business continuity of an enterprise. Because even the most robust components fail over time, administrators must be able to predict the behavior of the cluster database and take steps to ensure graceful failovers and disaster recovery.

BY SANJEET SINGH, ZAFAR MAHMOOD, AND PAUL RAD

Triple-nine, or 99.9 percent, availability means that a system is down for approximately 8 hours and 46 minutes per year. Depending on what the system does, this downtime could cost an enterprise millions of dollars. For instance, a database that houses ticket prices for an online travel Web site could cause an economic loss of this magnitude if it were to fail for just a few hours.

Oracle9*i*™ Real Application Clusters (RAC), the cluster database option for the Oracle9*i* database server, is often used to cluster business-critical databases that require high availability and predictability. This article documents the behavior of an Oracle9*i* RAC cluster when components fail, and describes best practices to prevent and recover from failures. Proper planning can help prevent failures or—in the worst case, when they do occur—enable optimal failover to restore system uptime.

Understanding Oracle9*i* RAC databases

A Dell™ and Oracle® implementation for Oracle9*i* RAC comprises up to eight Dell server nodes, a shared storage system composed of either a storage area network (SAN) or Dell PowerVault™ SCSI storage, the Red Hat® Enterprise Linux® AS 2.1 operating system, and Oracle9*i* database software with RAC. If SAN storage is used, then a Fibre Channel network is required and the storage enclosure is connected to the nodes through one or more Fibre Channel switches (see Figure 1). For SCSI storage, a SCSI cable connection between the nodes and storage enclosure suffices.

An Oracle9*i* RAC cluster requires a private and a public network. The private network uses Gigabit Ethernet[1] network interface cards (NICs) on the nodes, which can be connected by a Dell PowerConnect™ switch. Primary communication among the cluster nodes takes place over the private network. Clients and other application servers access the database over the public network.

Joint engineering effort

Individual components of the Dell and Oracle implementation for Oracle9*i* RAC are put through an independent, comprehensive test cycle by their respective vendors. The base operating system is tested and maintained by Red Hat. Both Dell and EMC develop the hardware components,

[1] This term indicates compliance with IEEE® standard 802.3ab for Gigabit Ethernet, and does not connote actual operating speed of 1 Gbps. For high-speed transmission, connection to a Gigabit Ethernet server and network infrastructure is required.

including the servers and the storage. Dell works with Oracle to develop the installation routines for the Dell and Oracle components. The Oracle9*i* RAC configuration is tightly bound, so users should be careful when installing new software (including drivers and utilities) to ensure that they are compatible with the kernel version and other software already installed.

In addition to developing some of the hardware pieces and the installation routines for the implementation, Dell runs the entire integrated configuration through a comprehensive test cycle. To provide test coverage for all the supported components, Dell tests many different platforms with the different back-end storage systems to help ensure full functionality. Dell testing methodologies help ensure that interconnects are working properly and that no contentions exist among the components of the configuration. Extensive stress tests are conducted before the configuration is released to customers, thus providing a fully functional database implementation.

Preparing for component failures

Despite thorough testing and adherence to strict quality standards, over an extended time all working components of such complex implementations can be expected to fail. The following sections detail the expected behavior of component failures in an Oracle9*i* RAC cluster and explain how to minimize the effects of such failures.

Server components and server nodes

In a cluster, the internal disks, the CPU, and even the nodes themselves represent potential failure points that can cause performance degradation for an Oracle9*i* RAC database implementation.

Internal disk. If the cluster database is running with a single disk and the disk fails, the administrator must replace the internal disk and install Oracle9*i* RAC software on the node. After installing this software on the node, the administrator must re-add the node to the cluster as if it were a newly deployed node. Administrators can consult the *Oracle9*i *Deployment Guide* for details on how to add a node to the cluster after Oracle9*i* RAC is installed.[2]

The internal disks hosting the operating system and the database software should be installed in a RAID-10 configuration, and administrators should plan ahead by keeping a few spare disks on hand in case of a failure. All supported Dell platforms have hot-plug hard drives. If a single disk in such a RAID configuration fails, the disk can be removed and replaced with another disk on a live system without affecting cluster operation. Using tools such as Dellmgr, administrators can rebuild a new disk within a few hours depending on the disk's total capacity.

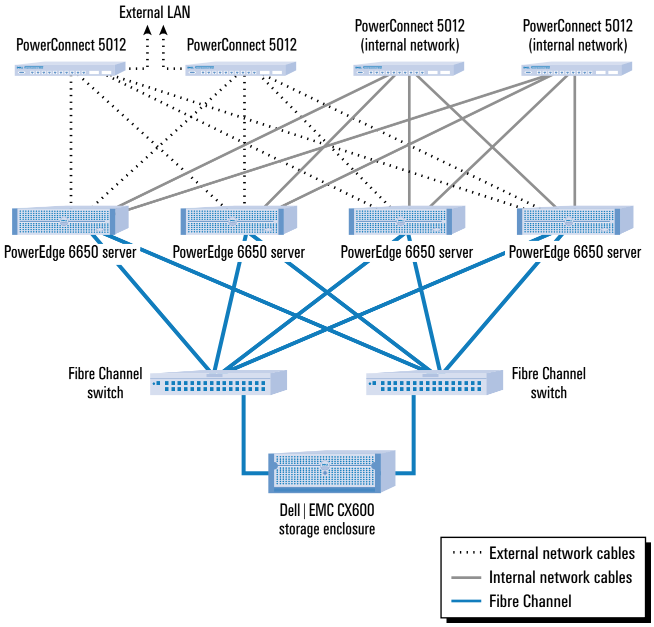


Figure 1. A failure-tolerant Oracle9*i* RAC cluster configuration

CPU. If a node with a single CPU fails, the node will shut down. Oracle9*i* database cluster software can dynamically remove a failed node from the cluster, but the CPU must be immediately replaced to prevent performance degradation caused by having one fewer node in the cluster.

If the node has multiple CPUs and a single CPU fails, the node will also shut down. To allow this node to function temporarily until a replacement CPU is procured, administrators can remove the failed CPU from the node and then restart the system. The only negative effect of running a node with one fewer CPU is lower performance on that particular node.

Once the CPU has been replaced, administrators can consult the *Oracle9*i *Deployment Guide* for instructions on starting the Cluster Manager and the database on this node. The node will join the cluster with no intervention by the administrator.

Server node. If a node fails, Oracle9*i* RAC removes the node from the cluster. The missing node reduces the performance of the entire cluster; therefore, a new node should be procured and installed immediately. Administrators can consult the *Oracle9*i *Deployment Guide* for instructions on adding a node to a cluster. To avoid failures, administrators can monitor the health of individual nodes using Dell Simple Network Management Protocol (SNMP) tools such as Dell OpenManage™ Server Administrator.[3]

[2] The *Oracle9*i *Deployment Guide* is located at http://www.dell.com/us/en/esg/topics/esg_oraclemain_servers_docs_1_pedge_o9irac.htm.

[3] For information on how to automate monitoring, see "Using Server Administrator to Automate the Monitoring of Server Health" by Jianwen Yin, Ph.D.; Alan Daughetee; Bala Beddhannan; and Andrew Wilks in *Dell Power Solutions,* May 2003.

Private and public network connections

An Oracle9*i* RAC cluster requires both a private (internal) and a public (external) network to operate. The failure of either private or public NICs, or the failure of a private network switch, can cause downtime or degrade the performance of the Oracle9*i* RAC database.

Private NIC. If the private interface fails, Oracle9*i* RAC regards that node as a failed node and removes it from the cluster, which leads to lower performance for the entire cluster. To recover from such a failure, administrators should remove the NIC from the system (or disable the NIC in the BIOS if the NIC is on the motherboard) and insert a new NIC. If running Oracle Cluster File System (OCFS), administrators must perform the following procedure after configuring the new interface similarly to the failed one:

1. Stop the cfs and ocmstart services.
2. Delete the last line of the /etc/ocfs.conf file.
3. Rerun the `ocfs_uid_gen -c` command.
4. Start the cfs and ocmstart services.

After confirming that the node can communicate with all the other nodes in the cluster, administrators can start the Cluster Manager and the database on this node. The node will join the cluster with no intervention by the administrator.

To avoid the negative repercussions of an interface failure, Dell recommends the use of NIC teaming to provide a redundant interface for the network connections. If a single interface fails, the presence of an additional interface offers instantaneous failover to keep the cluster functioning. The section in the *Oracle9*i *Deployment Guide* on configuring interconnect redundancy provides information on configuring NIC teaming.

Public NIC. If a public NIC—one connected to the external LAN—fails, external users will lose connectivity to the node in which that NIC resides. The node will not be removed from the cluster, because it continues to respond to other nodes on the private interface. However, the performance of the cluster degrades as the other nodes take over the users previously being routed to the failed node.

To recover from such a failure, administrators should shut down the node and replace the failed NIC—or disable the NIC in the BIOS if the NIC is on the motherboard—and insert a new NIC. After powering up the system and ensuring that it can communicate with all the other nodes on the external network, administrators can start the Cluster Manager and the database on this node. The node will join the cluster with no intervention by the administrator. To guard against external network failures, Dell recommends using NIC teaming and redundant external switches.

The internal disks hosting the operating system and the database software should be installed in a RAID-10 configuration, and administrators should plan ahead by keeping a few spare disks on hand in case of a failure.

Private network switch. If the private network switch fails, the database will shut down on all nodes in the cluster; a user from the external network will not be able to access the database. To recover from such a failure, administrators must replace the switch and, after confirming that all the nodes can communicate with each other, start the Cluster Manager and the database on all the nodes. The nodes will join the cluster with no intervention by the administrator.

As a safeguard to prevent against such failures, NIC teaming can be used in conjunction with two redundant switches. Redundant switches create two separate private networks, so that the failure of one network does not cause the database to shut down on all nodes. If one switch in such a configuration were to fail, the other switch might need up to 90 seconds to take over. Because the Cluster Manager by default is set to shut down a failed node in 200 seconds, switch failover requires no intervention by the administrator.

Fibre Channel switch fabric

The use of SAN storage introduces several more potential points of failure for an Oracle9*i* RAC database, from the host bus adapters (HBAs) and Fibre Channel switches to the storage enclosure itself.

HBA. If the HBA on a node fails, the node cannot connect to the external Fibre Channel storage enclosure. Oracle9*i* RAC automatically removes a node with a failed HBA from the cluster. Administrators must replace the HBA on this node and, using SAN management tools, reconfigure the HBA to join the same storage group to which it previously belonged. Once the node can connect to the shared storage, starting the Cluster Manager and the database on that node will cause it to rejoin the cluster with no intervention by the administrator.

Dell recommends that Oracle9*i* RAC clusters be deployed with at least two HBAs and EMC® PowerPath® software, creating redundant pathways to the SAN from the node.[4] Under normal conditions, the multiple pathways help provide I/O load distribution. During a single-HBA failure, other pathways keep the database from shutting down and enable the system to continue performing—albeit with slightly lower performance—until the HBA is replaced.

[4] For more information, see "Building a Highly Scalable and Available Data Environment for Oracle 9*i* RAC" by Zafar Mahmood, Paul Rad, and Robert Nadon in *Dell Power Solutions*, November 2003.

Fibre Channel switch. Failure of a Fibre Channel switch blocks the pathways from all nodes to the SAN, which causes the database to shut down on all the nodes. After such a failure, administrators can replace the Fibre Channel switch and restart the Cluster Manager and database on all the nodes, after first ensuring that each node can connect to the shared storage.

Configuring two Fibre Channel switches in conjunction with PowerPath creates two separate paths from each node to the SAN, which can mitigate the risks of such a failure. Even if one of the switches fails, the other path can keep the database functional with minimal performance degradation until the malfunctioning switch is replaced.

Storage enclosure. A storage enclosure failure causes the entire cluster database to shut down because no nodes can connect to the storage. The potential for data loss from such an event can be high, depending on the exact cause of the failure.

If the failure is caused by a component that can easily be replaced—a power supply or a storage processor, for example—then the data loss may not be substantial. After replacing the affected component, administrators can easily bring the storage array back online. To allow a degree of hot-failover capability, Dell | EMC CX400 and CX600 storage enclosures have multiple ingress ports for Fibre Channel connections. Thus, if the implementation has been configured with PowerPath, the failure of a port will not affect database operations. However, if some of the storage disks are affected, the potential for data loss can be great.

Dell recommends configuring storage arrays in at least a RAID-10 configuration to afford some data protection. However, a backup is still a requirement to help ensure no loss in case of a failure. A tape backup system can be employed to make regular backups of the SAN. If an enterprise cannot afford to have the database offline for even a minimal amount of time, then the SAN should be mirrored with a hot backup.

Creating a disaster recovery plan

A well-defined disaster recovery plan allows IT organizations to restore database service in the least amount of time—with the most complete database content—after a serious, unanticipated disruption in production service. Both regular tape backups of the SAN and synchronous mirroring are required to prevent data loss in case of a failure.

Tape backup. An inexpensive methodology, tape backup uses third-party tools such as LEGATO® NetWorker® software to make regular backups on a tape drive. Tapes can be stored off-site for an added measure of protection. If data becomes corrupted, the tapes can be used to restore the database to any point in time.[5]

Synchronous mirroring. Applications such as EMC MirrorView™ provide synchronous mirroring of data between different SANs. The secondary backup SAN maintains a copy of the production data on the primary site. Thus, if the primary SAN were to fail, the secondary SAN could be brought online very quickly.

Such a setup allows administrators to maintain a hot backup of the database system. In addition, EMC SnapView™ software enables administrators to capture point-in-time snapshots of the production data. Should a SAN failure be caused by a failed disk drive, the SAN can be quickly restored to the pre-failure stage after administrators replace the failed disk.[6]

Protecting data through contingency planning

Given the certainty that components in a complex infrastructure will fail over time, contingency planning for component failures is required to protect mission-critical Oracle9*i* RAC databases. Creating cluster configurations that follow Dell best practices can enable enterprises to better withstand component failures and eliminate single points of failure. When used with a full disaster recovery plan that includes tape backup and synchronous mirroring, organizations can help protect critical data from loss and enhance productivity by helping to ensure minimal database downtime.

Sanjeet Singh (sanjeet_singh@dell.com) is a software engineer in the Dell Database and Application Engineering Group. Sanjeet has a B.S. in Electrical Engineering and an M.S. in Computer Engineering from Purdue University.

Zafar Mahmood (zafar_mahmood@dell.com) is a software engineer in the Dell Database and Application Engineering Deployment Department of the Dell Product Group. He has been involved in database performance optimization, database systems, and database clustering for more than six years. He is currently working on Oracle9*i* RAC implementations at Dell. Zafar has an M.S. in Electrical Engineering with a specialization in Computer Communications from the City University of New York.

Paul Rad (paul_rad@dell.com) is a senior software engineer in the Dell Database and Application Engineering Department of the Dell Product Group. Paul has master's degrees in both Computer Science and Computer Engineering from the University of Texas at San Antonio.

FOR MORE INFORMATION

*Oracle9*i *Deployment Guide:* http://www.dell.com/us/en/esg/topics/esg_oraclemain_servers_docs_1_pedge_o9irac.htm

Oracle9*i* RAC: http://www.oracle.com/ip/index.html?rac_home.html

Remote SAN backups: http://www.emc.com/techlib/abstract.jsp?id=1224

Tape drive backups: http://portal2.legato.com/products/networker

[5] For a detailed analysis of two different tape backup solutions for an Oracle9*i* RAC database using LEGATO NetWorker, see "Backing Up and Restoring an Oracle Database" by Sanjeet Singh, Robert Nadon, William Collins, and Zafar Mahmood in *Dell Power Solutions,* February 2003.

[6] For more information about creating a remote backup of the SAN, see "Using SnapView and MirrorView for Remote Backup," http://www.emc.com/techlib/abstract.jsp?id=1224.

Integrating Dell PowerVault NAS Systems into UNIX and Linux Environments

Network attached storage (NAS) systems can provide shared storage across heterogeneous environments. This article examines the necessary components and configurations for integrating Dell™ PowerVault™ NAS servers into UNIX®- or Linux®-based networks and provides example configuration instructions for deploying a PowerVault server in a Red Hat® Linux 9.0 environment.

BY WARD WOLFRAM AND DEAN OLIVER

As businesses grow, data storage needs usually increase. Network attached storage (NAS) servers provide an affordable option for scaling out data storage capacity. Physically comparable to traditional file servers, NAS servers offer advanced file-sharing features and shared storage across different types of servers and clients.

Because they are exclusively dedicated to file sharing, NAS servers cannot run general-purpose applications such as Microsoft® SQL Server, Microsoft Exchange, or any other type of production application. NAS servers eliminate the features, functions, and drivers that reside in traditional servers. However, NAS servers usually are more reliable and easier to manage because they are less complex than traditional servers. NAS servers also can support third-party utilities to enhance file sharing, such as antivirus, tape backup, and disk quota management software.

In addition, NAS servers can help organizations achieve cost-effective storage consolidation, fault tolerance, multiplatform and multiprotocol storage sharing, and scalability, as well as moderate performance improvement. Administrators can configure, integrate, and manage NAS servers easily in an existing LAN. NAS servers also can provide non–host bus adapter (HBA) devices, such as low-cost desktop computers, with access to storage area networks (SANs).

Dell PowerVault servers: Meeting the need for NAS

Dell offers three NAS servers: the midrange Dell™ PowerVault™ 775N and PowerVault 770N systems, and the entry-level PowerVault 725N system. Each server runs a customized Microsoft Windows® Powered operating system (OS), which is based on the Windows 2000 Advanced Server OS with Service Pack 3 (SP3). To achieve elevated levels of scalability and availability, administrators can cluster the PowerVault 775N and PowerVault 770N, and use SAN-based storage.

By supporting several file-sharing protocols, Dell PowerVault NAS servers allow a wide variety of clients and servers to access shared storage concurrently. For example, networks with Windows-based hosts implement Common Internet File System (CIFS) to access shared storage, whereas UNIX®- or Linux®-based hosts traditionally use

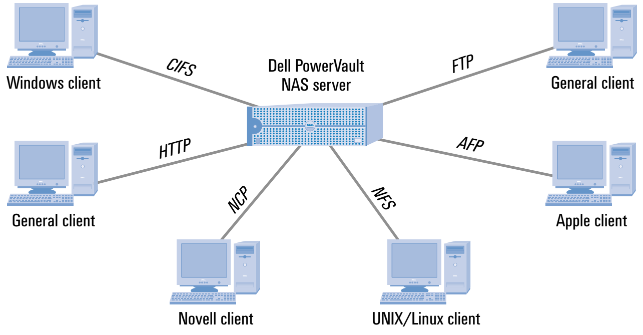


Figure 1. Outlining heterogeneous environments supported by Dell PowerVault NAS servers

Network File System (NFS). PowerVault servers support both CIFS and NFS, as well as Novell® NetWare® Core Protocol (NCP), AppleTalk® Filing Protocol (AFP), HTTP, and FTP (see Figure 1).

Incorporating PowerVault servers into an existing NFS-based UNIX or Linux environment requires Microsoft Services for UNIX (SFU) and may also require a Network Information Services (NIS) domain for file and directory permissions. By implementing NFS file-sharing functions with SFU 2.3, PowerVault servers provide a comprehensive set of services, tools, and utilities to the NAS server—easing integration into an existing UNIX environment.

Example configuration: Introducing PowerVault to Linux

The following scenario describes how to implement a Dell PowerVault NAS server in a Red Hat® Linux 9.0 environment so that the user home directories reside at one central location—the PowerVault server (see Figure 2). In this example, the PowerVault server, the NIS server with NFS automount, and the networked Red Hat Linux 9.0 clients have not been preconfigured. This example presents all required configurations in the proper sequence.

Step one: Configuring the Dell PowerVault NAS server

To implement home directories using a Dell PowerVault NAS server in a Linux environment, administrators first create NFS shares and user directories. After creating these shares and directories, administrators may restrict host access.

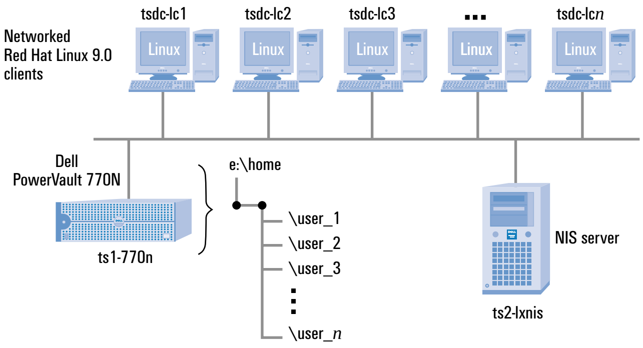


Figure 2. Deploying a Dell PowerVault server in the Red Hat Linux environment

PowerVault servers provide a Web-based GUI management tool called NAS Manager. To launch NAS Manager, administrators enter the IP address or network name of the PowerVault server in a Web browser. NAS Manager displays several menus when administrators log in; the example instructions in this article pertain exclusively to the Shares menu. Within the Shares section, administrators can select one of three submenus: Folders, Shares, and Sharing Protocols. All three submenus are used in this scenario.

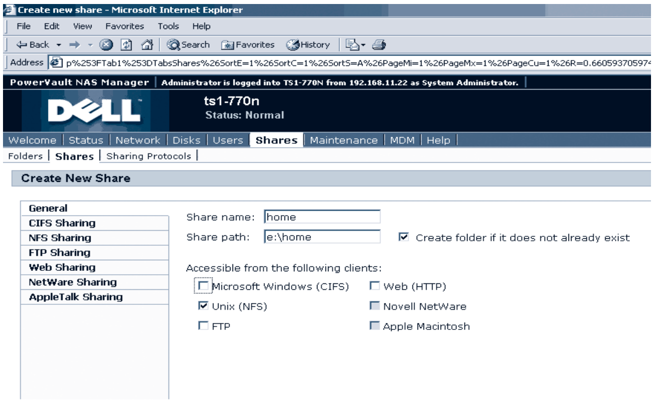
Create NFS shares on the PowerVault NAS server

The method for creating NFS shares on a Dell PowerVault NAS server differs significantly from the way administrators traditionally configure NFS shares on a UNIX- or Linux-based server. Because PowerVault servers are Windows Powered systems, administrative and management tasks follow typical Windows conventions.

In this example, the E: drive on the PowerVault server will host the home NFS share and all user subdirectories. Beginning on the Shares submenu, select "New" under the Tasks column. As shown in Figure 3, enter "e:\home" for the share path and select "Create folder if it does not already exist." Select the UNIX (NFS) protocol for client accessibility; disable all other file-sharing protocols by clearing the check box adjacent to each protocol name. Press the OK button to save and apply the configuration changes.

Within the Folders submenu, open Volume E: and then the Home folder. For each existing or potential Linux user, create a new folder, or *user home directory,* within the home NFS share by selecting "New" under the Tasks column (see Figure 4). Enter the Linux login name of the user as the name of the user home directory so that Windows will apply minimal access permissions ("Everyone" access) to the directories.

For each existing or potential Linux user, create a matching Windows account on the PowerVault server; matching the account names will simplify the mapping of Linux clients to Windows user



p%253FTab1%253DTabsShares%26SortE=1%26SortC=1%26SortS=A%26PageMi=1%26PageMx=1%26PageCu=1%26R=0.66059370597

PowerVault NAS Manager | Administrator is logged into TS1-770N from 192.168.11.22 as System Administrator.

ts1-770n
Status: Normal

Create New Share

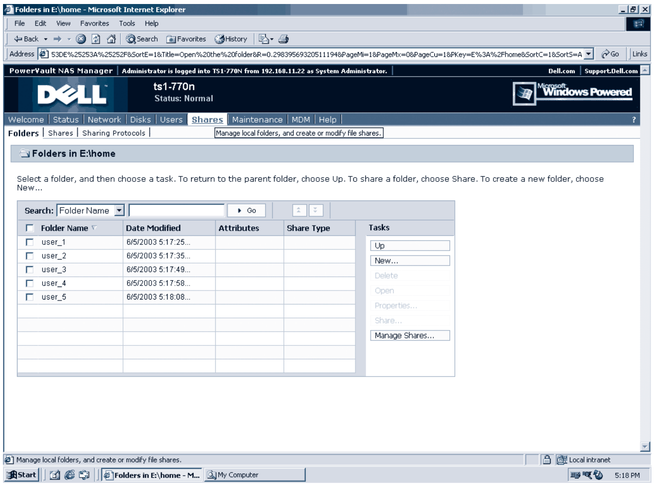
Share name: home

Share path: e:\home ☑ Create folder if it does not already exist

Accessible from the following clients:

☐ Microsoft Windows (CIFS) ☐ Web (HTTP)
☑ Unix (NFS) ☐ Novell NetWare
☐ FTP ☐ Apple Macintosh

Figure 3. Configuring the PowerVault server to host the home NFS share


Folders in E:\home - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address 53DE%25253A%25252F&SortE=1&Title=Open%20the%20folder&R=0.29839569320511194&PageMi=1&PageMx=0&PageCu=1&PKey=E%3A%2Fhome&SortC=1&SortS=A
Go Links
PowerVault NAS Manager
Dell.com
Support.Dell.com
Welcome Status Network Disks Users Shares Maintenance MDM Help
Folders Shares Sharing Protocols
Search: Folder Name Go
Tasks
Up
New...
Delete
Open
Properties...
Share...
Manage Shares...
Local intranet
Start
Folders in E:\home - M...
My Computer
5:18 PM

Administrator is logged into TS1-770N from 192.168.11.22 as System Administrator.

ts1-770n
Status: Normal

Manage local folders, and create or modify file shares.

Folders in E:\home

Select a folder, and then choose a task. To return to the parent folder, choose Up. To share a folder, choose Share. To create a new folder, choose New...

Folder Name	Date Modified	Attributes	Share Type
user_1	6/5/2003 5:17:25...		
user_2	6/5/2003 5:17:35...		
user_3	6/5/2003 5:17:49...		
user_4	6/5/2003 5:17:58...		
user_5	6/5/2003 5:18:08...		

Manage local folders, and create or modify file shares.

Figure 4. Creating user home directories for Linux clients within the home NFS share

accounts in a later step. Select "Local Users" under the Users menu; do not create a Windows Home Directory Path, because doing so will automatically apply permissions for the user name to this directory. Permissions cannot be applied yet, and must be applied utilizing the NIS server, which is discussed later in this article.

Restrict host access

The method for setting NFS rights and access permissions on Dell PowerVault systems also differs greatly from the method used on UNIX- and Linux-based servers. Administrators can control PowerVault NFS rights and access permissions on two levels: the hosts and users.

To configure a PowerVault NAS system without applying host restrictions, select the "home" share within the Shares submenu. Next, select the Properties option under the Task column and then the NFS Sharing tab. Make sure the Permissions entry is "ALL MACHINES Read-Write + Root" and select "Enable anonymous access" (see Figure 5). Press the OK button to save and apply the changes. At this point, the configuration of the Dell PowerVault NAS server is almost complete. Next, administrators must configure two Linux services: NIS and NFS automount.

Step two: Configuring the Linux server

A NIS server provides information vital to all machines on the network. NIS services run on a host that maintains a central database of information about users. In this example scenario, the NIS server distributes user login names, passwords, and home directory locations. A NIS server requires two background processes, or *daemons,* to be configured: ypserv and ypbind; ypbind must be configured on each Linux client as well. NFS automount services also require configuration modifications to the NIS server and each Linux client.

Administrators can implement file- and directory-level access permissions for PowerVault NFS shares either through PCNFS or a NIS server; this example scenario uses a NIS server and assumes there is no preexisting NIS server.

Create a NIS server with NFS automount

The following steps explain how administrators can configure a Linux server to act as the NIS server and provide NFS automount. Note that UNIX and Linux operating systems, unlike Windows, are case sensitive. Therefore, administrators must use lowercase characters for all UNIX or Linux variables, path names, definitions, and so forth to help ensure that the PowerVault server correctly recognizes them.

1. Set the domain name variable and include this entry in /etc/rc.d/rc.local:

```
domainname nisdomain
```

2. Now, activate the updated NIS domain name:

```
[root@ts2-lxnis /]# source /etc/rc.d/rc.local
```

3. Edit the file /var/yp/Makefile to include the following entry in the "all:" section of the file:

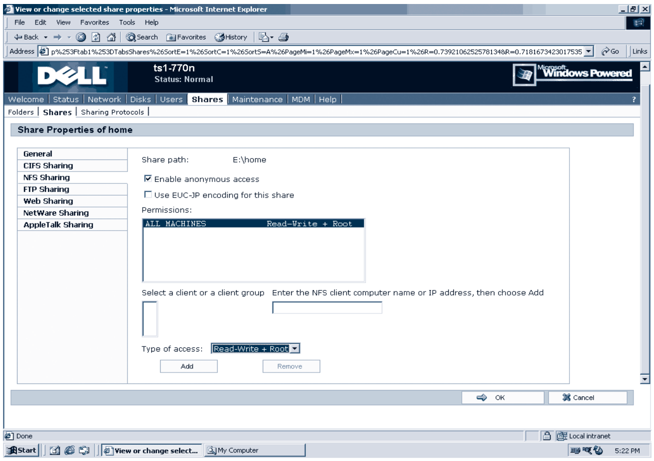
```
auto.master auto.home
```

4. To configure the NFS mounts from the PowerVault server, include the following entry in /etc/auto.master on the NIS server:

```
/home     /etc/auto.home --timeout=60
```

5. Create a mount point for mounting the PowerVault NFS share on the NIS server:

```
[root@ts2-lxnis /]# mkdir /home/users
```


View or change selected share properties - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address p%253Ftab1%253DTabsShares%26SortE=1%26SortC=1%26SortS=A%26PageMi=1%26PageMx=1%26PageCu=1%26R=0.7392106252578134&R=0.7181673423017535
Go Links
Welcome Status Network Disks Users Shares Maintenance MDM Help
Folders Shares Sharing Protocols
General
CIFS Sharing
NFS Sharing
FTP Sharing
Web Sharing
NetWare Sharing
AppleTalk Sharing
Add
Remove
OK
Cancel
Done
Local intranet
Start
View or change select...
My Computer
5:22 PM

ts1-770n
Status: Normal

Share Properties of home

Share path: E:\home

☑ Enable anonymous access

□ Use EUC-JP encoding for this share

Permissions:

```
ALL MACHINES          Read-Write + Root
```

Select a client or a client group

Enter the NFS client computer name or IP address, then choose Add

Type of access: Read-Write + Root

Figure 5. Configuring the PowerVault server with no host access restrictions

6. In /etc/auto.home, add the entry:

   ```
   users -fstype=nfs,rw ts1-770n:/home
   ```

7. Edit the line in /etc/yp.conf to reflect the NIS host name and domain:

   ```
   domain nisdomain server ts2-lxnis
   ```

8. Also edit this file to include the following entry:

   ```
   domain nisdomain broadcast
   ```

9. To automatically launch NIS services at boot, add the following entries in /etc/rc.local:

   ```
   ypserv
   ypbind
   ```

10. Then issue this command to start the ypserv and ypbind services:

    ```
    [root@ts2-lxnis etc]# source rc.local
    ```

11. Create a NIS database with its associated files (use this command only when first creating the NIS server):

    ```
    [root@ts2-lxnis yp]# /usr/lib/yp/ypinit -m
    ```

12. Follow the instructions that are given by the `ypinit` command.

13. Edit the /etc/nsswitch.conf file and add “nis” to the beginning of each entry that the NIS server will be distributing across the network. For example:

    ```
    passwd: nis files dns
    shadow: nis files dns
    group:  nis files dns
    hosts:  nis files dns
    .
    .
    .
    ```

    Adding “nis” will direct name resolution lookups to the NIS domain first, then to local hosts in /etc/hosts, and finally to DNS servers.

14. Start the automount service:

    ```
    [root@ts2-lxnis /]# service autofs restart
    [root@ts2-lxnis /]# chkconfig --level 35 autofs on
    ```

15. In /etc/exports, add the mount point entry:

    ```
    /home/users *(rw,no_root_squash)
    ```

16. Next, reload the new NFS configuration:

    ```
    [root@ts2-lxnis /]# exportfs -r
    ```

17. Finally, check the NFS configuration:

    ```
    [root@linux-nis /]# exportfs
    ```

Step three: Configuring the Linux clients

After configuring the Linux server, administrators should configure each Linux client in the environment. The following steps describe how to implement NIS services, automounts, and NFS on each client system:

1. Edit the line in /etc/yp.conf to reflect the NIS domain and host name:

   ```
   domain nisdomain server ts2-lxnis
   ```

2. Also edit this file to include the following entry:

   ```
   domain nisdomain broadcast
   ```

3. To start the client services in proper order, add the following entries to the /etc/sysconfig/network file:

   ```
   domainname nisdomain
   ypbind
   ```

4. Set the daemons to start automatically at boot:

   ```
   [root@linux-cl1 /]# chkconfig —level 345 nfs on
   [root@linux-cl1 /]# chkconfig —level 345 autofs on
   ```

5. Add the following entry into /etc/auto.master:

   ```
   /home  yp:auto.home
   ```

6. Restart the automount service:

   ```
   [root@linux-cl1 /]# service autofs restart
   ```

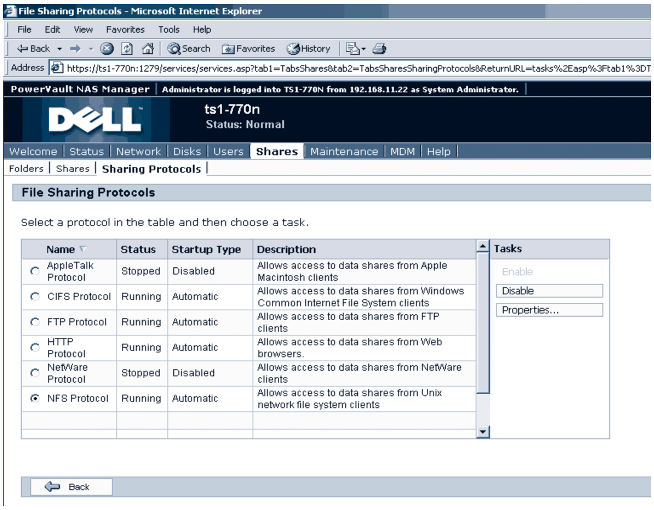
7. Edit the /etc/nsswitch.conf file and add “nis” to the beginning of each entry that will be provided by the NIS server. For example:

   ```
   passwd: nis files dns
   shadow: nis files dns
   group:  nis files dns
   hosts:  nis files dns
   .
   .
   .
   ```

   Adding “nis” will direct name resolution lookups to the NIS domain first, then to local hosts in /etc/hosts, and finally to DNS servers.

8. Add the following entry to the /etc/passwd file:

   ```
   +::::::
   ```


File Sharing Protocols - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address https://ts1-770n:1279/services/services.asp?tab1=TabsShares&tab2=TabsSharesSharingProtocols&ReturnURL=tasks%2Easp%3Ftab1%3DT
PowerVault NAS Manager
DELL
Welcome | Status | Network | Disks | Users | Shares | Maintenance | MDM | Help
Folders | Shares | Sharing Protocols
Tasks
Enable
Disable
Properties...
Back

Administrator is logged into TS1-770N from 192.168.11.22 as System Administrator.

ts1-770n
Status: Normal

File Sharing Protocols

Select a protocol in the table and then choose a task.

	Name	Status	Startup Type	Description
○	AppleTalk Protocol	Stopped	Disabled	Allows access to data shares from Apple Macintosh clients
○	CIFS Protocol	Running	Automatic	Allows access to data shares from Windows Common Internet File System clients
○	FTP Protocol	Running	Automatic	Allows access to data shares from FTP clients
○	HTTP Protocol	Running	Automatic	Allows access to data shares from Web browsers.
○	NetWare Protocol	Stopped	Disabled	Allows access to data shares from NetWare clients
●	NFS Protocol	Running	Automatic	Allows access to data shares from Unix network file system clients

Figure 6. Setting PowerVault directory and file access permissions for Linux clients

9. Add the following entry to the /etc/group file to direct queries for user and group information to the NIS server:
   +:::

   If placed at the bottom of the file(s), this entry will cause local files to be read first.

Step four: Creating Linux accounts in the NIS database

Before creating new Linux user accounts in the NIS database, administrators must create user directories on the PowerVault server (see "Step one: Configuring the Dell PowerVault NAS server"). To add user accounts at the NIS server, administrators must use the following command:

```
[root@ts2-lxnis /]# adduser user_1 -d
    /home/users/user_1
```

The -d option specifies the home directory of the new user. To set the new user's password, enter the following:

```
[root@ts2-lxnis /]# passwd user_1
Changing password for user user_1.
New password:
Retype new password:
passwd: all authentication tokens updated
    successfully.
[root@ts2-lxnis /]#
```

When updates to the NIS server are required or when additional users must be added to the NIS database, do not use the ypinit command mentioned previously in the section, "Create a NIS server with NFS automount." Instead, use the make command and change the directory to /var/yp:

```
[root@ts2-lxnis yp]# make
```

Then, from a client machine, attempt to log in as user_1. A successful login and access to the home directory of user_1 verifies a working configuration.

At this point, user accounts exist on both the Linux NIS server and the Windows PowerVault server. The final step is to complete the PowerVault configuration, creating a relationship between the two accounts and applying security measures to the user home directories.

Step five: Mapping Linux users to Windows accounts

Because user home directories exist on the PowerVault Windows Powered server, file and directory access permissions are applied using typical Windows security conventions. That is, security and permissions are set using only the Windows user accounts.

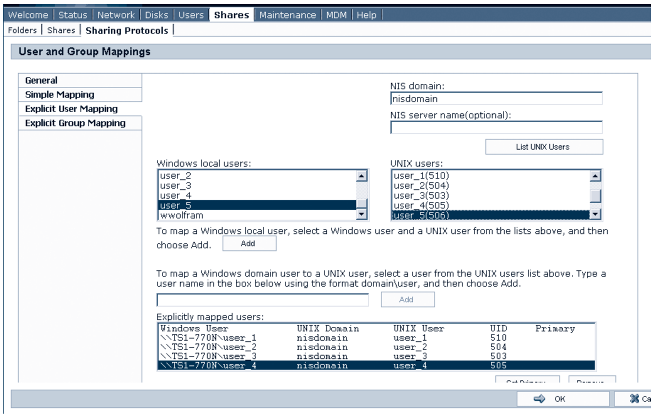
UNIX and Linux operating systems implement file- and directory-level access permissions differently from PowerVault servers; the two security models must coexist and work together in a heterogeneous environment. The PowerVault server must recognize UNIX and Linux user accounts, yet be able to apply file and directory access rights to those accounts using its native Windows conventions. The bonding of the UNIX/Linux and Windows security models is achieved using the Microsoft SFU User Name Mapping utility.

NAS servers can help organizations achieve cost-effective storage consolidation, fault tolerance, multiplatform and multiprotocol storage sharing, and scalability.

The User Name Mapping utility collects UNIX and Linux user accounts from a NIS server, and then administrators can manually or automatically map them to Windows user accounts. The UNIX and Linux user accounts are unmodified and continue to be managed by the NIS server. However, because the User Name Mapping utility affiliates the UNIX and Linux accounts with Windows accounts, file-access security can be applied using the usual Windows conventions.

Set directory and file access permissions

The following instructions conclude the configuration of the Dell PowerVault NAS server for a Red Hat Linux 9.0 environment. Within


Welcome | Status | Network | Disks | Users | Shares | Maintenance | MDM | Help
Folders | Shares | Sharing Protocols
User and Group Mappings
General
Simple Mapping
Explicit User Mapping
Explicit Group Mapping
List UNIX Users
Add
Add
OK
Ca

NIS domain: nisdomain

NIS server name(optional):

Windows local users:
user_2
user_3
user_4
user_5
wwolfram

UNIX users:
user_1(510)
user_2(504)
user_3(503)
user_4(505)
user_5(506)

To map a Windows local user, select a Windows user and a UNIX user from the lists above, and then choose Add.

To map a Windows domain user to a UNIX user, select a user from the UNIX users list above. Type a user name in the box below using the format domain\user, and then choose Add.

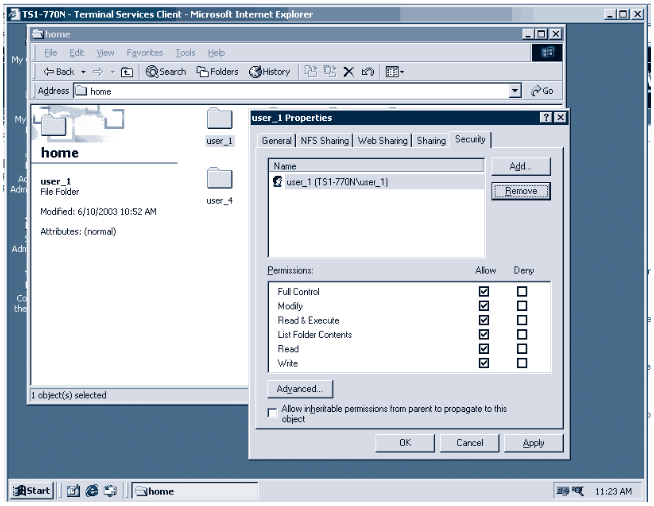
Explicitly mapped users:

```
Windows User          UNIX Domain    UNIX User    UID    Primary
\\TS1-770N\user_1     nisdomain      user_1       510
\\TS1-770N\user_2     nisdomain      user_2       504
\\TS1-770N\user_3     nisdomain      user_3       503
\\TS1-770N\user_4     nisdomain      user_4       505
```

Figure 7. Mapping Linux user accounts to local Windows accounts on the PowerVault server

the Sharing Protocols submenu of the NAS Manager Shares menu, administrators can follow these steps to set directory and file access permissions:

1. Select the NFS Protocol option button and then select the Properties task (see Figure 6).
2. Select the User and Group Mappings option.
3. Enter “nisdomain” as the NIS domain and select the Explicit User Mapping tab on the left side of the screen.
4. Select the List UNIX Users button to update all Linux user accounts from the NIS server.
5. Select a Linux user account and its associated Windows local account and then press the Add button (see Figure 7). This action creates the relationship between a Linux and Windows user account so that file and directory access permissions can be applied using typical Windows conventions.
6. Press the OK button to save and apply the configuration settings.
7. To apply permissions to user home directories, select the Terminal Services option under the Maintenance menu of NAS Manager. Log in to the PowerVault server and launch Microsoft Windows Explorer.
8. Locate the home directory of user_1 and view its properties.
9. Select the Security tab, as shown in Figure 8.
10. Clear the box adjacent to “Allow inheritable permissions from parent to propagate to this object.” If asked, remove this property.
11. If “Everyone” permission exists for the directory, remove it from the permissions list. Then set the appropriate permission level for each user. For example, add “user_1” (user name) and “Administrator” (permission level) to the permissions list and select the Full Control check box for both.



Address: home

user_1
File Folder

Modified: 6/10/2003 10:52 AM

Attributes: (normal)

1 object(s) selected

Name: user_1 (TS1-770N\user_1)

Permissions:	Allow	Deny
Full Control	☑	☐
Modify	☑	☐
Read & Execute	☑	☐
List Folder Contents	☑	☐
Read	☑	☐
Write	☑	☐

☐ Allow inheritable permissions from parent to propagate to this object

Figure 8. Applying access permissions to user home directories

Dell PowerVault: Versatile NAS systems

As the example scenario in this article shows, Windows-based Dell PowerVault servers can be configured easily to accommodate the storage and data access needs of multiplatform, multiprotocol environments. By supporting various file-sharing protocols such as NFS, PowerVault servers can help provide shared storage to many different types of network configuration. NAS systems such as the Dell PowerVault 725N, 770N, and 775N servers can help IT organizations adapt to and meet ever-changing storage needs, without requiring companies to overhaul their entire IT infrastructure.

Ward Wolfram (ward_wolfram@dell.com) is a storage performance and solution engineer on the Dell Solution Enablement Lab and Technology Showcase team. His responsibilities include performance and best-practice analysis for SAN, NAS, and tape backup systems. Ward has an M.S. in Computer Science from the University of Nebraska at Lincoln and a B.S. in Mathematics from Concordia University in Seward, Nebraska.

Dean Oliver (dean_oliver@dell.com) is a senior technical analyst in the Dell Linux Operating System Development Group. His responsibilities include test matrices, troubleshooting, and development of various Red Hat Linux releases on Dell PowerEdge servers. Dean attended LeTourneau University in Longview, Texas, and is a Red Hat Certified Engineer® (RHCE®), Master Certified NetWare Engineer (MCNE), and Microsoft Certified Systems Engineer (MCSE).

FOR MORE INFORMATION

Dell PowerVault servers: http://www.dell.com/nas

Implementing Logical Volumes

on Linux-based Dell PowerEdge Servers

Sistina® Logical Volume Manager (LVM) software provides an abstraction layer that allows administrators to work with logical volumes, helping to simplify storage management. This article describes LVM and presents scenarios for using LVM to resize storage volumes on Dell™ PowerEdge™ and PowerVault™ systems running the Red Hat® Linux® operating system.

BY TESFAMARIAM MICHAEL AND JOSHUA GILES

System administrators who manage disk storage in network environments face several challenges: ensuring that data-intensive applications get the data they need quickly, providing and organizing additional storage, preserving data integrity, and backing up data regularly. Scalable storage management software that enables administrators to reconfigure storage allocations dynamically in response to changing enterprise needs can help meet these challenges. One such open source product is the Sistina® Logical Volume Manager (LVM), which is embedded in the Linux® kernel. LVM combines physical disk drives into easily manageable logical volumes and allows online storage administration through tasks that are transparent to the user and the application.

This article discusses the basic architecture and configuration of LVM running under the Red Hat® Linux operating system. It also explains the steps for using LVM to manage storage in Dell™ PowerEdge™ servers and Dell PowerVault™ systems.

Understanding LVM architecture

LVM comprises five basic structures: volume group (VG), logical volume (LV), logical extent (LE), physical extent (PE), and physical volume (PV). A physical volume is typically a hard disk or a partition—for example, a SCSI disk or a RAID abstraction of a hard disk. Each physical volume comprises chunks of data, called physical extents. All physical extents within a volume group are the same size. In Figure 1, the volume group, called Vol_Grp00, is at the top level of abstraction. This group is obtained by

LVM combines physical disk drives into easily manageable logical volumes and allows online storage administration through tasks that are transparent to the user and the application.

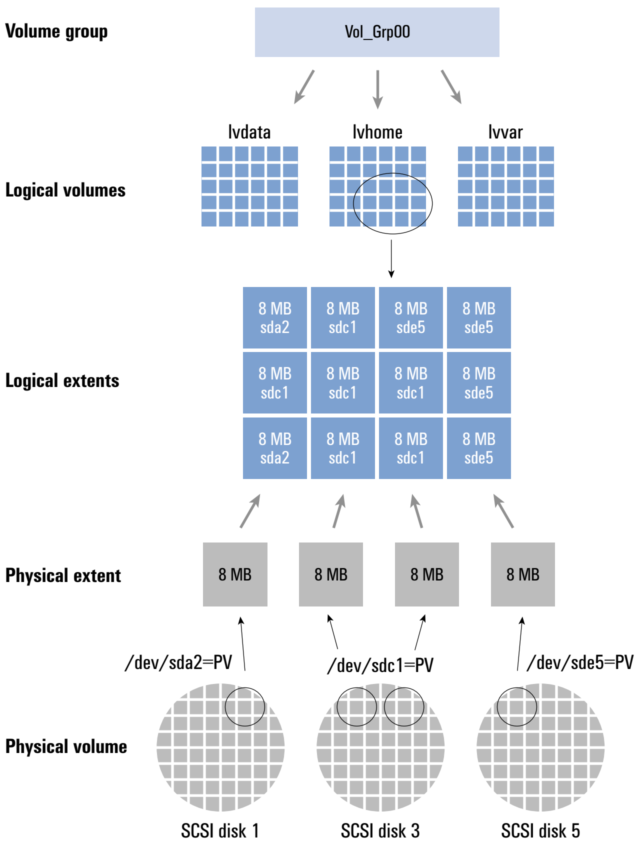


Figure 1. LVM schematic diagram

mapping together the physical volumes /dev/sda2, /dev/sdc1, and /dev/sde5 to create logical volumes that form one manageable group.

In Figure 1, lvdata, lvhome, and lvvar represent logical volumes, which contain the file system and thus can be used as mount points for directories such as /var/ftp, /home, and /usr. Each logical volume is split into chunks of data, called logical extents. Within a volume group, the size of the logical extent is the same as that of the physical extent; that is, a one-to-one mapping exists between the logical extents and the physical extents.

Each physical extent has a unique identification number on a physical volume but does not necessarily have a unique identification number on a logical volume because several different physical volumes can constitute one logical volume. Therefore, the logical extent identification numbers also identify the associated physical extents. Whenever the storage area is accessed, the address or the identification number of the logical extent is used to actually perform I/O on the physical storage.

The volume group descriptor area (VGDA) is stored at the beginning of each physical volume and functions similarly to the partition table for LVM. The VGDA contains one volume group descriptor, one physical volume descriptor, one logical volume descriptor, and several physical extent descriptors. When the system is booted up, the volume group and logical volumes are activated and the VGDA is loaded into memory. The VGDA allows LVM to identify where the logical volumes are actually stored. The one-to-one mapping of physical to logical volumes is necessary to access the physical location and to perform I/O operations.

Because storage management is one of the main challenges system administrators regularly face, most major Linux distributions have included LVM with their installers, simplifying the task of LVM deployment.

Conceptualizing LVM processes

LVM comprises lvm-mod, a kernel module (that is, a device driver) under General Public License (GPL), and applications that use the module to perform storage-related management processes. In a typical scenario, administrators use an LVM command such as `pvcreate` or `vgscan` to perform storage management functions. For example, Figure 2 represents the execution flow of an LVM task to retrieve the status of a particular storage device, as follows:

1. LVM commands invoke the LVM kernel module, lvm-mod, to perform the operating system–level tasks that are required to service the request.
2. The device driver executes the request on the hardware storage device, and the LVM kernel module configures the device dynamically through the /proc file system.
3. LVM dynamically obtains information on the storage devices by using the /proc file system.

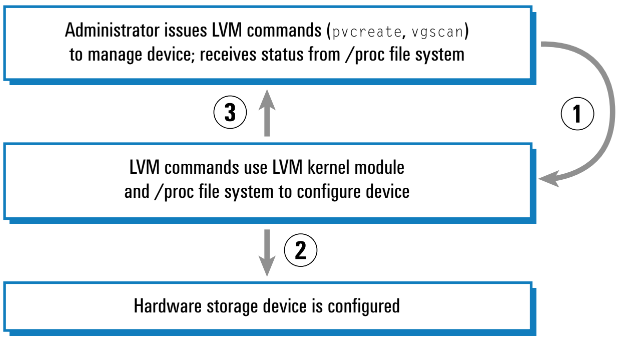

Administrator issues LVM commands (pvcreate, vgscan)
to manage device; receives status from /proc file system
1
3
LVM commands use LVM kernel module
and /proc file system to configure device
2
Hardware storage device is configured

Figure 2. LVM command process

File system mount point	Size (MB)	File system type	LVM group*	Logical volume name
/	6144	ext3	None	None
/boot	128	ext3	None	None
Logical volume mount point	Size (MB)	File system type	LVM group	Logical volume name
/home	4096	ext3	lvm00	lvm_home
/usr	6144	ext3	lvm00	lvm_usr
/var	6144	ext3	lvm00	lvm_var
/data	4096	ext3	lvm00	lvm_data

*Neither the root nor boot partitions use LVM.

Figure 3. Example partition values

Implementing LVM during a Linux installation

Because storage management is one of the main challenges system administrators regularly face, most major Linux distributions have included LVM with their installers, simplifying the task of LVM deployment. For instance, Red Hat Linux included LVM starting with its 8.0 release. Although LVM can be introduced during system installation or any time after, this article discusses only the case of deploying LVM during Red Hat Linux installation and modifying storage volumes afterward. Migration of an already installed system to LVM can be complex, especially if it involves moving any file systems required for system boot, such as root and boot, to an LVM volume.

An important problem administrators must address when installing Linux is how best to partition and allocate system storage. Predicting how these partitions will be used during the lifetime of the system can be difficult. Even though storage on Linux systems can be expanded by adding more disks or storage enclosures, expanding or modifying existing partitions is very risky, if not impossible. LVM greatly simplifies the task of expanding or modifying partitions.

For data protection and integrity, Dell strongly suggests using a Dell PowerEdge Expandable RAID Controller (PERC) or Dell Cost-Effective RAID Controller (CERC) implementing a RAID level of 1, 5, or higher, depending on system needs.

Planning storage partitions and allocations

Before installing Linux, administrators should plan how to partition and allocate their storage. They should also evaluate how much risk and complexity they are willing to handle in order to recover a failed system. For example, creating root and boot partitions outside LVM makes system recovery easier by eliminating the need to activate LVM when rescuing a failed system. Also, those using Linux Loader (LILO) as the preferred boot loader should be aware that it does not support LVM. Administrators using LILO must put the boot partition outside LVM.

Although LVM can be used with several file system types, such as ext2, ext3, and ReiserFS, this article addresses only ext3, the default Red Hat Linux file system. Figure 3 shows an example partitioning scheme; the installation steps described in the following sections use these values.

Using Disk Druid for manual partitioning

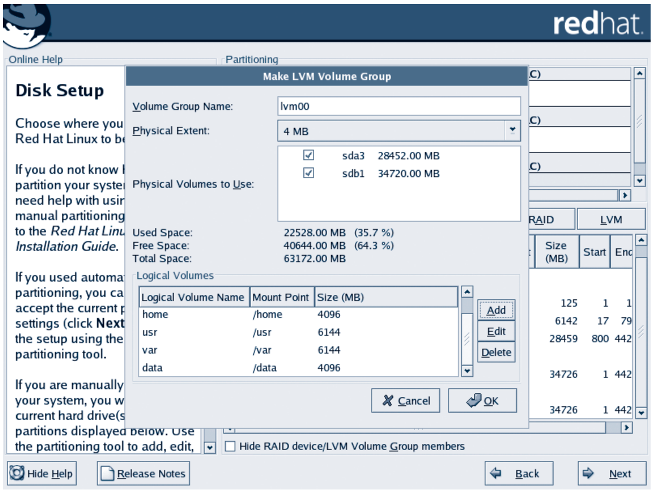
Implementing a partitioning scheme such as the one shown in Figure 3 during installation of Red Hat Linux 9 Professional is straightforward. The LVM feature is available only in the graphical user interface (GUI) installation mode. To use this feature, boot the system by using the Red Hat Linux 9 CD 1. At the boot prompt, press the Enter key to install in GUI mode. At the Disk Partitioning Setup screen, select "Manually partition with Disk Druid," followed by "Next."

When the Disk Druid menu is displayed, administrators can use the New button to create both boot and root partitions on the first hard drive, /dev/sda. Then, selecting the remaining free space of /dev/sda to edit and setting the file system to LVM enables administrators to create physical volumes on the remaining space of the first drive. This step can be repeated for all the remaining hard drives. Selecting the LVM button on the Disk Druid screen launches the Make LVM Volume Group window.

From this window, administrators can add, edit, and delete logical volume groups. Other possible actions include setting the volume group name and physical extent size (the default is 4 MB) and choosing the hard drive on which to create the logical volume. The Add and Edit buttons allow administrators to set the mount point, size, file system type, and name of the planned logical volumes (see Figure 4).

Even though storage on Linux systems can be expanded by adding more disks or storage enclosures, expanding or modifying existing partitions is very risky, if not impossible. LVM greatly simplifies the task of expanding or modifying partitions.

If implemented using the default setting of 4 MB per physical extent, a volume group is limited to 255 GB because only 65,534 physical or logical extents are allowed in a volume group. To provide more than 256 GB in a volume group, administrators must create the physical extents in chunk sizes larger than 4 MB. Physical extents must be a power of 2 and can range anywhere between 8 KB



Make LVM Volume Group

Volume Group Name: lvm00

Physical Extent: 4 MB

Physical Volumes to Use:

☑ sda3 28452.00 MB

☑ sdb1 34720.00 MB

Used Space: 22528.00 MB (35.7 %)

Free Space: 40644.00 MB (64.3 %)

Total Space: 63172.00 MB

Logical Volumes

Logical Volume Name	Mount Point	Size (MB)
home	/home	4096
usr	/usr	6144
var	/var	6144
data	/data	4096

☐ Hide RAID device/LVM Volume Group members

Figure 4. Make LVM Volume Group window in Red Hat Linux installer

and 512 MB. However, the installer limits this range to between 1 MB and 64 MB.

Managing and maintaining system storage

Once LVM is implemented during installation, managing and maintaining system storage is the next task. If any logical volumes in the system become full, more space can be added by extending the logical volume using several methods, including:

- Add more space from the free unallocated space in the same volume group (see Scenario A in this article)
- Reduce another logical volume (see Scenario B)
- Add more hard drives or Dell PowerVault disk enclosures to the system (see Scenario C)

In the following scenarios that detail these approaches, /home is the logical volume to be expanded. Although LVM does not require these safety measures, administrators should perform a full system backup and bring the system to runlevel 1 (init 1) before starting the expansion; in some approaches, the logical volume being modified should be taken offline. Dell also recommends performing these operations during off-peak hours. For more information about the LVM commands used in these scenarios, see “LVM commands for managing and maintaining system storage” in this article.

Scenario A: Expanding from unallocated space in the same volume group

The “Make LVM Volume Group” window in Figure 4 shows that the volume group lvm00 still has 40 GB of active, unallocated space that can be used to expand any of the logical volumes. The following steps describe how to expand the logical volume /home by 10 GB from this unallocated space:

1. Confirm that the desired logical volume is not being used and take it offline:

   ```
   umount /logical_volume_mounting_point
   ```

   For this scenario, enter:

   ```
   umount /home
   ```

2. To scan for physical and then logical volumes, display all existing physical volumes with information about the physical device (such as /dev/sda3) to volume group mappings, the volume group to which the device belongs, and the present size status of the device:

   ```
   pvscan
   lvscan
   ```

3. To determine the unallocated free space that can be used to expand the logical volumes, display all volume groups and their size allocation status:

   ```
   vgdisplay
   ```

4. Expand the logical volume from an active but unallocated physical volume:

   ```
   lvextend -L +X /dev/lvm_name/logical_volume_name
      /dev/physical_volume
   ```

   where *X* is the expansion size. The results from steps 2 and 3 provide the available size for this expansion. The `-L` parameter extends the logical volume size. In this scenario, the unallocated physical volume is /dev/sda3. Expand the logical volume /home by 10 GB by entering:

   ```
   lvextend -L +10G /dev/lvm00/home /dev/sda3
   ```

5. Verify that the logical volume was extended by comparing the current size to that found in step 2:

   ```
   pvscan
   lvscan
   ```

   These commands should report the updated space information.

6. Resize the file system:

   ```
   e2fsck -f /dev/logical_volume_name
   resize2fs /dev/lvm_name/logical_volume_name
   ```

   For this scenario, enter:

   ```
   e2fsck -f /dev/lvm00/home
   resize2fs /dev/lvm00/home
   ```

7. Bring the logical volume online:

   ```
   mount /logical_volume_mounting_point
   ```

   For this scenario, enter:

   ```
   mount /home
   ```

8. Verify that the added space is accounted for:

   ```
   df -h
   ```

Administrators can also use the `e2fsadm` command to expand the logical volume and resize its file system; this command combines the `lvextend`, `e2fsck`, and `resize2fs` commands.

Scenario B: Reducing another logical volume

Administrators should expand logical volumes by reducing another only when the other volume has sufficient additional space. For instance, if a data volume uses only a fraction of its allocated space and this usage is expected to continue, administrators can safely reduce that volume by some amount and assign that space to another logical volume, /home in this scenario. Administrators should not reduce the logical volume before reducing the file system, because doing so can corrupt the file system on the reduced logical volume. Again, Dell recommends backing up the system before following the approach in this scenario. Performing the following procedure expands one logical volume by reducing another—in this case, by adding 2 GB from /data to /home:

1. Identify a logical volume to shrink that has sufficient free space. The command `df -h` reports information about the system storage usage.

2. Bring the system down to runlevel 1:

   ```
   init 1
   ```

3. Take both logical volumes offline:

   ```
   umount /logical_to_shrink
   umount /logical_to_grow
   ```

   For this scenario, enter:

   ```
   umount /data
   umount /home
   ```

4. Reduce the selected logical volume and its associated file system:

   ```
   e2fsadm -L -X /dev/lvm_name/logical_volume_name
   ```

   where *X* is the size by which the volume should be reduced. Take care not to reduce the size beyond its free space; doing so can cause data loss—or corrupt the file system.

   For this scenario, enter the following command to reduce both the file system and the logical volume /data by 2 GB:

   ```
   e2fsadm -L -2G /dev/lvm00/data
   ```

5. Expand the logical volume by the same amount the other volume was reduced:

   ```
   e2fsadm -L +X /dev/lvm_name/logical_volume_name
   ```

   where *X* is the expansion size. For this scenario, expand /home by 2 GB by entering:

   ```
   e2fsadm -L +2G /dev/lvm00/home
   ```

6. Mount both file systems and bring the system up to runlevel 3 or 5:

   ```
   mount -a
   init desired_runlevel
   ```

7. Verify the changes:

   ```
   df -h
   ```

Administrators can also use the `resize2fs`, `lvreduce`, and `lvextend` commands in place of `e2fsadm` to achieve these results.

LVM COMMANDS FOR MANAGING AND MAINTAINING SYSTEM STORAGE

Figure A shows various LVM commands used in resizing and managing logical volumes. Detailed information about each command is available from the command's man page.

Command	Description
`pvcreate`	Creates a physical volume by initializing a disk or a partition
`pvscan`	Scans all drives for physical volumes and reports drive usage, including available space
`lvscan`	Scans all disks for logical volumes
`lvextend`	Expands the size of a logical volume by adding a given amount of space from a given physical volume
`lvreduce`	Reduces the size of a logical volume (if a file system exists on the volume, the file system must be reduced using `resize2fs` before running this command; `lvreduce` does not permit reducing a logical volume size below its used space)
`vgextend`	Extends a volume group by adding physical volumes
`lvdisplay`	Displays information about the attributes of a logical volume
`vgdisplay`	Displays information about the attributes of volume groups
`e2fsadm`	Expands or reduces an ext2 file system and logical volume together
`resize2fs`	Expands or reduces an ext2 or ext3 file system
`e2fsck`	Checks a Linux ext2

Figure A. LVM commands

Scenario C: Adding more storage

If a Dell PowerEdge server lacks unallocated or unused storage space, administrators must add more storage to expand volumes. This approach requires adding more hard drives or a Dell disk enclosure such as a PowerVault 220S. For background on expanding storage for PowerEdge servers, see "Expanding Storage on Linux-based Servers" by Matt Domsch and Tesfamariam Michael in *Dell Power Solutions,* February 2003 (http://www.dell.com/us/en/esg/topics/power_ps1q03-michael.htm). The article is particularly helpful because the expansion procedure can be very complex, especially for RAID storage, the recommended type. To expand a volume by adding more storage, administrators can follow these steps:

LVM eases storage management in Linux installations by empowering Linux system administrators to expand and reduce storage at an abstracted level.

1. Partition the added drive or RAID volume and set its partition ID to 8e:

   ```
   fdisk /dev/new_added_drive
   ```

   For this scenario, to add the drive /dev/sdc, enter:

   ```
   fdisk /dev/sdc
   ```

2. Create a physical volume on the new partition:

   ```
   pvcreate /dev/partition_of_new_added_drive
   ```

   For this scenario, enter:

   ```
   pvcreate /dev/sdc1
   ```

3. Take the logical volume to be expanded offline:

   ```
   umount /logical_volume_mounting_point
   ```

   For this scenario, enter:

   ```
   umount /home
   ```

4. Expand the volume group:

   ```
   vgextend volume_group_name physical_volume_name
   ```

   For this scenario, enter:

   ```
   vgextend lvm00 /dev/sdc1
   ```

5. Expand the size of the logical volume and its file system:

   ```
   e2fsadm -L +X logical_volume_name
   ```

   where *X* is the expansion size.

   For this scenario, to expand /home by 10 GB, enter:

   ```
   e2fsadm -L +10G /dev/lvm00/home
   ```

   Alternatively, use the `lvextend` and `resize2fs` commands instead of `e2fsadm`. In this case, enter:

   ```
   lvextend -L +10G lvm00
   resize2fs /dev/lvm00/home 10G
   ```

6. Mount the expanded logical volume:

   ```
   mount /logical_volume_mounting_point
   ```

   For this scenario, enter:

   ```
   mount -a
   ```

7. Bring the system back to the desired runlevel (3 or 5):

   ```
   init desired_runlevel
   ```

8. Verify the changes:

   ```
   df -h
   ```

Achieving more manageable storage environments

LVM eases storage management in Linux installations by empowering Linux system administrators to expand and reduce storage at an abstracted level. Using the techniques described in this article, administrators can upgrade, reallocate, and modify storage resources as needed, helping to create a nimble, scalable storage environment with minimal complexity.

Tesfamariam Michael (tesfamariam_michael@dell.com) is a software engineer on the Linux Development Team of the Dell Product Group, which tests Linux on all Dell PowerEdge servers. Tesfamariam has an M.S. in Computer Science from Clark Atlanta University, a B.S. in Electrical Engineering from the Georgia Institute of Technology, and a B.S. in Mathematics from Clark Atlanta University. His areas of interest include operating systems and I/O devices.

Joshua Giles (joshua_giles@dell.com) is a Linux systems management software engineer on the Linux Development Team of the Dell Product Group. His interests include operating systems, grammar- and automata-based programming, and Support Vector Machine (SVM) learning. Joshua has a B.S. from the New Mexico Institute of Mining and Technology.

FOR MORE INFORMATION

Dell and Linux: http://www.dell.com/linux

LVM HowTo page: http://tldp.org/HOWTO/LVM-HOWTO

Red Hat Linux: http://www.redhat.com

Sistina LVM: http://www.sistina.com/products_lvm.htm

New Sage in the Enterprise

Reliable backup procedures are critical to storage management. The DLTSage™ suite of tools helps improve data protection by providing predictive failure alerts, detailed diagnostics, and analytical data capture, and can give administrators more confidence in their backup operations.

BY DIANNE McADAM

Protecting sensitive data and data resources is a primary concern in enterprise computing. Data that is lost or inaccessible—through causes ranging from natural catastrophes to human error—can mean lost revenue and unrealized business opportunities.

Local mirroring and remote replication can limit the risk of an outage, but they also can propagate both good and corrupt data. To mitigate the risk of data loss caused by deletion or corrupted data, IT administrators must establish and follow sound backup and recovery procedures.

Backups are an essential part of daily activities in the data center. Backup and restore software can restore corrupted or deleted data to its most recent uncorrupted state. However, backups do not always complete successfully. Vague messages or human error can contribute to an unsuccessful backup operation. In fact, unclear or cryptic error messages from backup applications also can make it difficult, if not impossible, to determine the root cause of backup failures. Was it a tape media failure, a tape drive failure, or a backup software problem? Administrators may lose confidence in the backup application, when the real problem is lack of clear, detailed information to determine why a backup did not complete successfully.

Enter DLTSage: Enriching information gathering

DLTtape™ technology addresses the problem of insufficient information by using a suite of predictive and preventive maintenance tools. These tools provide alerts to help administrators predict failure, supply detailed diagnostics, and capture analytical data. By working with existing Super DLTtape™ (SDLT) drives and other backup and restore applications from independent software vendors (ISVs), DLTSage can help reduce the guesswork required to determine how and why a backup failed.

Like traditional maintenance programs, DLTSage reports on tape drive and media failures. But DLTSage

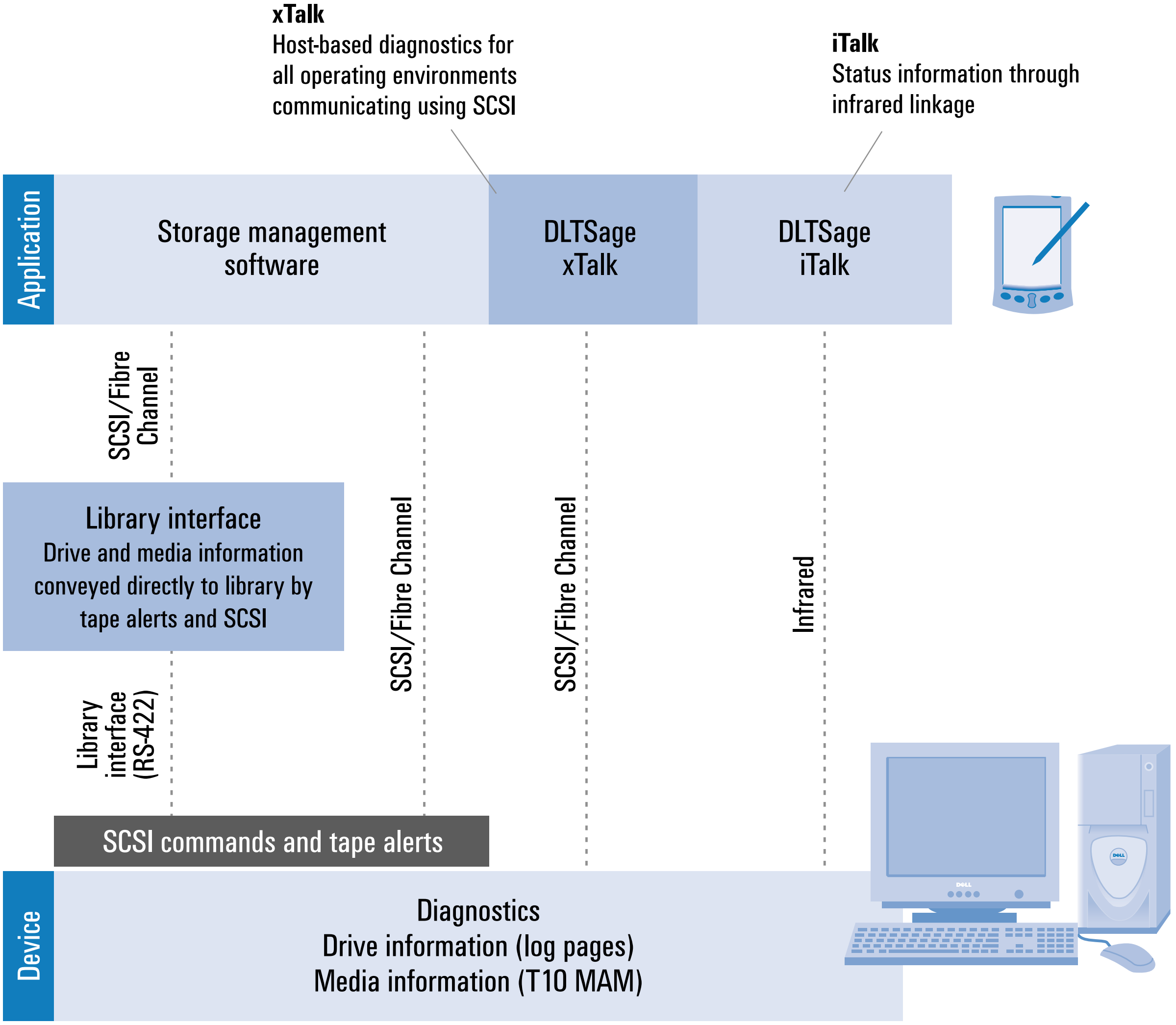

xTalk
Host-based diagnostics for all operating environments communicating using SCSI
iTalk
Status information through infrared linkage
Application
Storage management software
DLTSage xTalk
DLTSage iTalk
SCSI/Fibre Channel
Library interface
Drive and media information conveyed directly to library by tape alerts and SCSI
SCSI/Fibre Channel
SCSI/Fibre Channel
Infrared
Library interface (RS-422)
SCSI commands and tape alerts
Device
Diagnostics
Drive information (log pages)
Media information (T10 MAM)

Figure 1. DLTSage architecture

provides additional functions, as follows:

- Monitoring tape drive and cartridge usage to predict when tapes are nearing the end of their reliable life
- Predicting when a tape drive or cartridge may fail
- Determining whether the potential problem is a tape drive, tape cartridge, or both
- Recommending that an older cartridge be retired from active service or recommending that maintenance be scheduled for a particular drive

DLTSage unites hardware and ISV software reporting schemes into a single administrative interface. Now when a backup error occurs, DLTSage—working with backup applications—can help administrators determine whether the problem is software- or hardware-related. It can then recommend and communicate a course of action to the appropriate administrators for resolution.

How it works: Improving access to information

Today, DLTSage includes two access tools: xTalk and iTalk. These components use different conduits to access information about drives and media (see Figure 1).

The DLTSage xTalk automatically discovers DLTtape drives attached to a storage network and tracks information by device type, SCSI ID, firmware version, and unit serial number. It captures usage statistics about all attached drives and cartridges, and monitors environmental conditions such as temperature of the tape path or the number of power-up hours. In addition, xTalk can determine whether backups are being run at the preset interval and can highlight drives that are underutilized.

DLTSage xTalk is a host agent that currently runs on the Microsoft® Windows®, Hewlett-Packard® HP-UX®, Sun™ Solaris™, and Linux® operating systems, with support for more platforms scheduled in the future. Through a series of menus, xTalk allows administrators to run diagnostics or access tape and drive information through commands that are sent over SCSI or Fibre Channel paths to DLTtape drives.

DLTtape technology addresses the problem of insufficient information by using a suite of predictive and preventive maintenance tools. These tools provide alerts to help administrators predict failure, supply detailed diagnostics, and capture analytical data.

The second component, DLTSage iTalk, runs on a Microsoft Pocket PC or any laptop equipped with an infrared port. A hardware engineer equipped with a Pocket PC can stroll through a computer room and point the device at an SDLT tape drive. The device communicates to the drive through its infrared links and retrieves drive statistics quickly. The result is a greatly simplified storage administration process; running reports on the server or queuing reports for printing is no longer necessary. A future release is expected to include a Web-based management component, DLTSage eTalk, for customers requiring remote access.

Integration with standards

DLTSage can be integrated with other backup and restore applications that support the T10[1] Medium Auxiliary Memory (MAM) and TapeAlert standards. The MAM standard defines how vendors should store status information on a selected area of a tape cartridge. The first 4 KB of status information contain mandatory parameters—such as the serial numbers of the last four drives that used the cartridge—and application information stored by backup software vendors. The MAM standard allocates additional space that hardware vendors or ISVs can use to track information, such as the status of recent backup jobs or the serial numbers of the last 50 drives that read or wrote data to the cartridge.

The DLTtape technology implementation of the MAM standard has been expanded to include this additional information. By tracking such information, DLTtape technology can perform extended trend analysis and increase the accuracy of predictions rendered by DLTSage. ISVs also can take advantage of DLTtape extensions by storing detailed information about the status of their backups.

The T10 TapeAlert standard uses 64 status flags to provide information about tape drives, such as the number of drive write recoverable errors or the number of unrecoverable read errors. Library and tape drive vendors share status information using TapeAlert, and ISVs can query flag status and convert the information to text messages that are easy to understand.

DLTSage helps ensure successful backups

Many IT administrators have experienced the frustration of an incomplete backup, wondering whether the restore will really work—this time. However, the problem is not with the backup operation itself, but with the processes that determine whether a backup operation actually completed successfully.

DLTSage identifies and helps mitigate risk by monitoring the backup process, detecting and reporting failures as they happen. In addition, DLTSage predicts when specific drive or tape failures may occur. These capabilities can give administrators more confidence in their backup and restore software, and help ensure successful restore operations.

Dianne McAdam (dmcadam@datamobilitygroup.com) is senior analyst and partner at Data Mobility Group. With more than three decades of experience in IT, Dianne leads Data Mobility Group's research and advisory services on topics such as replication technology, business continuance, and networked storage. Before joining Data Mobility Group, Dianne led the Information Logistics practice at Illuminata. In addition, Dianne counseled prospects and customers on enterprise storage at the EMC® Executive Briefing Center. At Hitachi Data Systems, Dianne was a performance and capacity planning systems engineer for mainframe environments, and later a systems engineering manager. She also worked at StorageTek as a virtual tape and disk specialist, at Sun Microsystems as an enterprise storage specialist, and at several area companies as a technical services director. Dianne has bachelor's and master's degrees in Mathematics from Hofstra University in Long Island, New York. She is a founding member and treasurer of the Greater Boston Regional Computer Measurement Group.

FOR MORE INFORMATION

DLTSage: http://www.DLTSage.com

DLTtape: http://www.DLTtape.com

[1] T10 is a technical committee of the International Committee on Information Technology Standards (INCITS).