



Campus Clusters Based on Sun™ Cluster 3.0 Software

Hartmut Streppel, Enterprise Solutions Ambassador

Sun BluePrints™ OnLine—November 2002



<http://www.sun.com/blueprints>

Sun Microsystems, Inc.
4150 Network Circle
Santa Clara, CA 95045 USA
650 960-1300

Part No.: 817-0369-10
Revision 1.0
Edition: November 2002

Copyright 2002 Sun Microsystems, Inc. 4150 Network Circle, Santa Clara, California 95045 U.S.A. All rights reserved.

Sun Microsystems, Inc. has intellectual property rights relating to technology embodied in the product that is described in this document. In particular, and without limitation, these intellectual property rights may include one or more of the U.S. patents listed at <http://www.sun.com/patents> and one or more additional patents or pending patent applications in the U.S. and in other countries.

This product or document is protected by copyright and distributed under licenses restricting its use, copying, distribution, and decompilation. No part of this product or document may be reproduced in any form by any means without prior written authorization of Sun and its licensors, if any. Third-party software, including font technology, is copyrighted and licensed from Sun suppliers.

Parts of the product may be derived from Berkeley BSD systems, licensed from the University of California. UNIX is a registered trademark in the United States and other countries, exclusively licensed through X/Open Company, Ltd.

Sun, Sun Microsystems, the Sun logo, Sun BluePrints, Sun BluePrints Online, Solaris, Sun Cluster, SunPlex, Sun StorEdge, Sun StorEdge Availability Suite Software, Sun StorEdge Instant Image, Solaris Resource Manager, Solaris Volume Manager, Solstice DiskSuite, Sun Enterprise, Sun Fire, and Sun Management Center are trademarks or registered trademarks of Sun Microsystems, Inc. in the United States and other countries. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. in the US and other countries. Products bearing SPARC trademarks are based upon an architecture developed by Sun Microsystems, Inc.

The OPEN LOOK and Sun™ Graphical User Interface was developed by Sun Microsystems, Inc. for its users and licensees. Sun acknowledges the pioneering efforts of Xerox in researching and developing the concept of visual or graphical user interfaces for the computer industry. Sun holds a non-exclusive license from Xerox to the Xerox Graphical User Interface, which license also covers Sun's licensees who implement OPEN LOOK GUIs and otherwise comply with Sun's written license agreements.

U.S. Government Rights—Commercial use. Government users are subject to the Sun Microsystems, Inc. standard license agreement and applicable provisions of the FAR and its supplements.

DOCUMENTATION IS PROVIDED "AS IS" AND ALL EXPRESS OR IMPLIED CONDITIONS, REPRESENTATIONS AND WARRANTIES, INCLUDING ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT, ARE DISCLAIMED, EXCEPT TO THE EXTENT THAT SUCH DISCLAIMERS ARE HELD TO BE LEGALLY INVALID.

Copyright 2002 Sun Microsystems, Inc., 4150 Network Circle, Santa Clara, Californie 95045 Etats-Unis. Tous droits réservés.

Sun Microsystems, Inc. a les droits de propriété intellectuels relatants à la technologie incorporée dans le produit qui est décrit dans ce document. En particulier, et sans la limitation, ces droits de propriété intellectuels peuvent inclure un ou plus des brevets américains énumérés à <http://www.sun.com/patents> et un ou les brevets plus supplémentaires ou les applications de brevet en attente dans les Etats-Unis et dans les autres pays.

Ce produit ou document est protégé par un copyright et distribué avec des licences qui en restreignent l'utilisation, la copie, la distribution, et la décompilation. Aucune partie de ce produit ou document ne peut être reproduite sous aucune forme, par quelque moyen que ce soit, sans l'autorisation préalable et écrite de Sun et de ses bailleurs de licence, s'il y en a. Le logiciel détenu par des tiers, et qui comprend la technologie relative aux polices de caractères, est protégé par un copyright et licencié par des fournisseurs de Sun.

Des parties de ce produit pourront être dérivées des systèmes Berkeley BSD licenciés par l'Université de Californie. UNIX est une marque enregistrée aux Etats-Unis et dans d'autres pays et licenciée exclusivement par X/Open Company Ltd.

Sun, Sun Microsystems, le logo Sun, Sun BluePrints, Solaris, Sun Cluster, SunPlex, Sun StorEdge, Sun StorEdge Availability Suite Software, Sun StorEdge Instant Image, Solaris Resource Manager, Solaris Volume Manager, Solstice DiskSuite, Sun Enterprise, Sun Fire, et Sun Management Center sont des marques de fabrique ou des marques déposées, ou marques de service, de Sun Microsystems, Inc. aux Etats-Unis et dans d'autres pays. Toutes les marques SPARC sont utilisées sous licence et sont des marques de fabrique ou des marques déposées de SPARC International, Inc. aux Etats-Unis et dans d'autres pays. Les produits portant les marques SPARC sont basés sur une architecture développée par Sun Microsystems, Inc.

L'interface d'utilisation graphique OPEN LOOK et Sun™ a été développée par Sun Microsystems, Inc. pour ses utilisateurs et licenciés. Sun reconnaît les efforts de pionniers de Xerox pour la recherche et le développement du concept des interfaces d'utilisation visuelle ou graphique pour l'industrie de l'informatique. Sun détient une licence non exclusive de Xerox sur l'interface d'utilisation graphique Xerox, cette licence couvrant également les licenciés de Sun qui mettent en place l'interface d'utilisation graphique OPEN LOOK et qui en outre se conforment aux licences écrites de Sun.

LA DOCUMENTATION EST FOURNIE "EN L'ÉTAT" ET TOUTES AUTRES CONDITIONS, DECLARATIONS ET GARANTIES EXPRESSES OU TACITES SONT FORMELLEMENT EXCLUES, DANS LA MESURE AUTORISEE PAR LA LOI APPLICABLE, Y COMPRIS NOTAMMENT TOUTE GARANTIE IMPLICITE RELATIVE A LA QUALITE MARCHANDE, A L'APTITUDE A UNE UTILISATION PARTICULIERE OU A L'ABSENCE DE CONTREFAÇON.



Please
Recycle



Adobe PostScript

Campus Clusters Based on Sun™ Cluster 3.0 Software

This article describes how to use Sun™ Cluster 3.0 software as part of a comprehensive disaster recovery solution to ensure continuous service availability. This article provides basic guidelines to consider when deploying a campus-cluster solution and offers helpful tips for setting up sound administrative practices.

This article highlights key technologies involved in spreading a SunPlex™ environment (enabled by Sun Cluster 3.0 software) across a company campus or distributed sites. It describes the processes that need to be incorporated by data centers to leverage the capability of campus clusters.

This article is targeted at IT architects and technical staff who want to understand, evaluate, and address single-site level disaster recovery for their data centers.

This article contains the following topics:

- “Introduction” on page 2
- “Technology Options for Disaster Recovery Solutions” on page 5
- “Quick Checklist for Deployments” on page 7
- “Campus Cluster Maximum Distances” on page 10
- “Campus Cluster Topologies and Components” on page 12
- “Campus Cluster Configurations” on page 18
- “Performance in a Campus Cluster Environment” on page 23
- “Management Aspects of Campus Clusters” on page 25
- “Glossary” on page 28
- “About the Author” on page 29
- “Related Resources” on page 29

Introduction

Disasters do not happen often, but when they do occur, they are likely to have a significant impact on business in terms of lost revenue and service availability. Ensuring business continuity requires that enterprises deploy a multifaceted solution that includes several levels of disaster prevention and recovery technologies and well-documented procedures.

As part of a comprehensive, flexible, and scalable disaster recovery solution, campus clusters based on Sun Cluster 3.0 and newer versions can help protect service availability. With the SunPlex environment, enterprises can deliver higher service levels while helping to protect their critical business services from unavoidable risks—from small interruptions, such as power failures, to major catastrophes such as earthquakes and fires.

Yet technology alone does not address all aspects of continuous service availability. While most enterprises deploy some type of disaster recovery technology to protect against hardware failures or isolated incidents, protecting against a major catastrophe requires a well-planned, comprehensive solution. To ensure the highest levels of business continuity, enterprises must invest in three essential components—people, processes, and products. A well-trained staff armed with thoroughly tested procedures and a robust cluster infrastructure such as Sun Cluster 3.0 is the best defense against detrimental service interruptions.

Cluster Evolution

The concept of clustering two or more redundant servers and related storage arrays was originally introduced to ensure higher levels of availability in mission-critical or compute-intensive environments. These original clusters were expensive to manage, complex to administer, and difficult to extend as needs changed. Consequently, their use was limited. As high-end servers became more affordable and more widely used by enterprises of all types, clustering technology evolved to provide much greater flexibility, scalability, and manageability with increasing levels of service availability.

Local clusters (for example, clusters where all of the nodes¹ and storage subsystems are at the same site) play a major role in achieving business continuity by providing a solid level of continuous service availability. In the early days of clustering technology, share storage subsystems were usually attached using SCSI technology. Due to technology limitations, the maximum distance between cluster nodes was limited by the maximum cable length between a server, the shared storage, and the other server, which could not exceed 50m.

1. Terminology note: Throughout this document, the term “node” describes the server or a single domain within a server that supports multiple domains. The term “server” is always used to describe the physical enclosure.

While this configuration offers good protection against events such as node disk crashes, it does not protect against events that could destroy or damage the facility site.

With the advent of Fibre Channel technology, it became possible to replicate data over much greater distances. Now, enterprises can deploy cluster nodes and storage in different buildings or even at different sites without changing the software infrastructure, applications, data, and storage subsystems, thus building extended clusters.

Cluster Limitations

One drawback of long wires between nodes or storage is increased latency, which can decrease performance dramatically. When latency increases because of longer distances (for example, across a country), other types of data replication (for example, asynchronous mirroring) and other types of high availability or disaster recovery solutions (for example, a cluster of clusters) needed to be developed to solve these problems.

Although extended clusters offer significant protection against disasters, they are not a complete disaster recovery solution. A cluster that has only one logical copy of data is still vulnerable against inconsistencies that might be introduced by faulty software or hardware, even if that data is mirrored. Common user errors such as erroneously deleting database tables may cause a major disaster. In those cases, tape backup or some other up-to-date copy of data is invaluable for recovery.

Even cluster software can fail, especially in the case of a major disaster affecting the cluster infrastructure. For example, a campus cluster where all the nodes are located within a few kilometers may be subject to a major earthquake, knocking out utilities or otherwise affecting its operation. To protect against this possibility, most enterprises deploy a multifaceted solution to ensure continuous service availability.

People, Processes, and Products

Campus clusters are one of the best examples where people, processes, and products must work together for the solution to deliver its maximum benefits. The entire integrated stack of products, from servers and storage subsystems to the operating environment and clustering software, forms only the base of a highly available campus cluster infrastructure. Well-trained, dedicated people must then administer the infrastructure. Processes that cover all aspects of disaster prevention and recovery must be in place. A well-prepared enterprise not only deploys a comprehensive solution, it verifies its processes, trains its staff, and tests the technologies regularly (at least annually). When personnel are trained, best practices implemented, and sound technologies deployed, companies can deliver the high levels of service continuity required to remain competitive in today's economy.

SunPlex Environment

Built around the Sun Cluster 3.0 solution, the Solaris™ Operating Environment (Solaris OE), and Sun™ server, storage, and network connectivity products, the SunPlex environment helps increase business service levels while decreasing the costs and risks of managing complex enterprise networks. Through the SunPlex environment, devices, file systems, and networks can operate seamlessly across a tightly coupled pool of resources, making it easy to deploy extended or campus clusters without changing the underlying infrastructure or applications. A campus cluster based on Sun Cluster 3.0 software is a cluster where nodes are separated by distance in at least two sites.

Sun Cluster 3.0 software is designed to protect against single hardware or software failures such as node crashes or service interruptions. For greater reliability and performance, Sun Cluster 3.0 software is tightly integrated with the Solaris OE. This integration speeds up error detection time and makes the whole software stack more robust.

Depending on the failure, Sun Cluster 3.0 software either fails over the affected services to another node in the cluster or tries to restart them. In either case, the software's highest priority is to maintain data integrity regardless of what happens. This requirement drives the layout of the infrastructure and all of the algorithms in the product. This requirement is the reason why in certain disaster scenarios it might be necessary to initiate recovery procedures manually, in order not to jeopardize data integrity.

Standard monitoring agents are available for many best-of-breed databases and ERP applications. Agents for other services can be developed and deployed using either sophisticated APIs or easy-to-use utilities such as the SunPlex Agent Builder tool.

The Sun Cluster 3.0 software framework and associated algorithms do not change when deployed in a campus cluster. Service availability with data integrity is the primary goal. Depending on the actual requirements, the Sun Cluster 3.0 solution can form an excellent base for a disaster recovery solution, especially when combined with additional technologies, trained personnel, and well-developed management processes.

Technology Options for Disaster Recovery Solutions

Campus cluster solutions can form the base for comprehensive disaster recovery solutions that include any one or a combination of the following technologies:

- Backup and recovery – Backup scripts and management software automatically generate copies of data regularly, which are then archived safely on site or at a remote location.
- Outsourced data services – Some enterprises choose to use an outside vendor to provide a replica of a production data center that can be installed quickly in case of a disaster.
- Database replication – Databases may be configured to replicate data to a remote server over an IP network. In case of a failure, a manual procedure would put the remote database into production.
- Log shipping – Included in most modern database products, this technology uses the logs produced by the database to “recover” a standby database at a remote site. The remote database is in a standby state, and the logs are applied to it either immediately or after a time gap to prevent logical errors from migrating into the remote database. Logs are usually sent via an IP network, but could also be replicated synchronously using other technologies.
- Data replication over networks using Sun StorEdge™ Availability Suite and Sun StorEdge™ Instant Image – Sun StorEdge Availability Suite replicates arbitrary storage volumes to remote sites over IP networks, while the Sun StorEdge Instant Image product allows a point-in-time copy or snapshot to be used. The combination of these two products can be used to replicate a snapshot of a database to a remote site.
- Data replication (mirroring) using Fibre Channel (FC) – High-end storage products offer the capability to replicate data to a remote storage subsystem of the same type using direct connections without affecting the servers attached to the storage. This replication usually copies data blocks to a remote site. Typical examples of this type of product are EMC SRDF and Hitachi Data Systems TrueCopy.

Each of these options is adequate for certain situations, but most enterprises choose a combination of several technologies to deploy a complete disaster recovery solution. Each option must be evaluated on how well it meets an enterprise's requirements, its short-term (deployment) and long-term (management) costs, and the level of protection it provides. TABLE 1 compares some of the capabilities of these options.

TABLE 1 Comparison of Disaster Recovery Technologies

Technology	Independent Data	Clustered	Automatic Recovery	Recovery Time	Maximum Distance ⁽¹⁾
Remote Backup and Restore	Yes	No	No	High	Network
Log Shipping	Yes	No, but possible	No, but possible	Medium	Network
Sun StorEdge Availability Suite/ Sun StorEdge Instant Image	No	No	No	Medium	Network
Storage-Based Replication	No	No	No	Medium	Storage Interconnect or Network
Database Replication	Yes	No	No	Medium	Network
Campus Clusters	No	Yes	Yes	Fast	10 km ⁽²⁾
Remote Backup and Restore	Yes	No	No	High	Network

1.Distance negatively affects performance.

2.Campus Clusters using wave division multiplexers (WDMs) support longer distances.

One of the main differences between each of these options is the mechanism used to replicate data. Having two or more copies of data in sync at any time is an advantage for fast, automated failover. However, this approach carries a risk: any defect that exists in the data is mirrored to the other copy, making both copies corrupted or useless. In this case, having a independent copy of data, such as with backup and restore, is essential.

A common alternative is to send logical data packets, such as database logs, to the remote site and apply them to the database after time has elapsed. This approach makes it possible to detect errors—logical, administrative errors as well as hardware related inconsistencies—and prevent them from being applied to the remote copy.

Instead of sending database logs via an IP network, database files can be replicated using Sun StorEdge Availability Suite, either working directly on live data or on a snapshot that could be made using Sun StorEdge Instant Image technology.

Regardless of which combination of options an enterprise may choose, the technologies must be supported by rational processes implemented through careful planning and training.

Quick Checklist for Deployments

Sun Cluster 3.0 software is a scalable, flexible solution that can be deployed with equal benefit to small local clusters and larger extended clusters. Before deploying a campus cluster solution, however, consider your enterprise's requirements, resources, and risks. The following checklist provides an overview of factors to consider when determining which level of solution is best for an enterprise.

General Questions

What do you want to protect against?

Certain kinds of failures are more probable than others. For example, if the data center is close to a river, flooding may be a likely risk. The potential risk might impose restrictions on the solution. For example, to protect against site outages due to a major earthquake, a remote site may need to be established 500 kilometers away.

If the planned second site is in another geography, other solutions may have to be applied. If the desired result is total protection against all disasters, a campus cluster may not be adequate.

How much revenue per hour or market credibility could your enterprise lose if your mission-critical services were not available?

Understanding the financial impacts of a disaster on business operations helps define an adequate solution. Consider at least three factors:

- The cost of a disaster recovery solution
- The cost and impact of a disaster, including recovery times, potential data loss, lost revenue, and loss of reputation and credibility in the market
- The probability that disaster will occur

If the potential loss is less than the cost to protect against that loss, a disaster recovery solution does not make sense. Unfortunately, history has shown that enterprises without good disaster recovery solutions suffer far more setbacks than businesses with plans. Campus clusters generally provide for very cost-effective protection against the more probable disaster scenarios, such as the loss of a whole site due to fire or other natural disasters.

Do you understand that a campus cluster does not address all aspects of a disaster recovery solution?

A campus cluster provides the appropriate infrastructure against many types of disasters, however, it does not offer complete protection against all disasters. In addition, it must be accompanied by other mechanisms to help ensure business continuity, for example, to recover from a total loss of data. Finally, a product alone is not enough; a solution needs the right people and processes in place to make it complete.

Infrastructure Questions

Are you prepared to accept performance degradation because of prolonged distances?

Even traveling at the speed of light through a fiber takes time, and the time it takes for data to travel 10 kilometers is 1000 times longer than it takes to go 10 meters. Even the latency introduced by the long wires is only a small part of the overall latency. The additional components involved, such as transceivers, switches, and multiplexers, add to the latency.

Does your data center infrastructure provide two or more sites?

Quorum devices help the cluster decide which nodes may form a new cluster in case of any failure. Thus, the availability of the quorum device is key in a disaster situation. In the case of the very common two-site infrastructure, the quorum device has to be placed at one of the two sites, making the loss of that data center more catastrophic than the loss of the other. In a three-site infrastructure, the quorum device is at the third site, so that the loss of one site would not affect the majority of quorum votes. For more information, refer to “Campus Cluster Topologies and Components” on page 12.

Does your infrastructure provide for independent Fibre Channel and network lines to span the distance between the data centers?

To prevent interference from other components on the same network or storage connections, independent lines or multiplexers should be available. Refer to “Campus Cluster Configurations” on page 18 for more information.

Does your infrastructure provide for a single IP subnet across the two (or three) sites?

To maintain accessibility to high availability services, clusters failover IP addresses. To configure the same IP addresses on network interface cards (NICs) and networks in different sites, it is necessary to have a single IP subnet across the two or three sites. Refer to “Campus Cluster Configurations” on page 18 for more information.

People and Processes

Is your staff well-trained and willing to undergo continuing training and exercising?

Having well-trained and experienced administrative staff is one of the key factors for achieving high availability. Periodically testing, at least annually, the disaster prevention and recovery processes is essential for training personnel to minimize service interruptions and restore normal operations.

Are there already processes in place that deal with recovery procedures for disasters?

Defining and establishing new processes to be used during disasters is a major task. If similar procedures are already in place, adapting them to the new infrastructure and new products is less effort. Administrative staff who are familiar with the processes understand and incorporate changes faster.

Campus Cluster Maximum Distances

A campus cluster based on Sun Cluster 3.0 software is a cluster where nodes are separated by distance in at least two sites. Conventional wiring (copper or multimode fiber) technology used in data centers does not span the distance needed for the cluster and storage interconnects. In many cases, use of third-party fiber networks and WDMs may be necessary to bridge the required distances of a geographically dispersed cluster.

TABLE 2 summarizes the maximum distances for storage and network interconnects based on the IEEE 802.3 standards and product specifications. By using special third-party hardware, such as transceivers and single-mode fibers, these distances can be further extended. Transceivers are used in SunPlex technology-based campus

clusters to convert to and from fiber, both multimode and single mode. Maximum distances should be checked against product specifications for each specific deployment.

TABLE 2 Maximum Distances for Interconnects

Interconnect Type	Maximum Distance
Storage Interconnect	
SWGBIC (1Gb)	500m
SWGBIC (2Gb)	150m/300m ⁽¹⁾
LWGBIC (1Gb/2Gb)	10000m ⁽²⁾
Cluster Interconnect Gigabit Ethernet 1000BASE-X	
SWGBIC (1000BASE-SX)	220m/550m ⁽³⁾
LWGBIC (1000BASE-LX)	5000m ⁽⁴⁾
Cluster Interconnect Fast Ethernet BASE-X ⁽⁵⁾	
100BASE-TX (twisted pair)	100m ⁽⁶⁾
100BASE-FX (half duplex, multimode fiber)	412m ⁽⁷⁾
100BASE-FX (full duplex, multimode fiber)	2000m ⁽⁸⁾
Storage Interconnect	500m

1. With 62.5µm/50.0µm multimode fiber.

2. Same distance with both versions.

3. With 62.5µm/50.0µm multimode fiber (IEEE 802.3 2000 38.3).

4. With 9.0µm single-mode fiber (IEEE 802.3 2000 38.4) Most Gigabit Ethernet adapters today use short wave Gigabit Interface Converter Modules (GBICs) (1000BASE-SX). This maximum distance can be extended using transceivers or media converters.

5. IEEE did not set a distance limitation on 100BASE-FX over single-mode fiber solutions. Hardware vendors offer transceivers that can achieve more than 10 km using single-mode fiber.

6. IEEE 802.3 2000 29.1.1

7. IEEE 802.3 2000 29.1.1

8. IEEE 802.3 2000 29.4

Campus Cluster Topologies and Components

There are many considerations involved when planning a campus cluster topology, such as:

- Number of cluster nodes
- Type of interconnects
- Type of servers and storage interconnects
- Distance
- Availability of a third site

Although the vast majority of clusters deployed today are two-node clusters, a more robust campus cluster consists of four nodes, two at each site, and two-way, host-based mirrors across sites. To protect against local storage failures, controller-based RAID (such as RAID-5) is used within the storage arrays.

Multipathing solutions protect against failures of the storage paths and make it very unlikely that data needs to be completely resynchronized under normal circumstances. This configuration helps ensure that in most failures, failover would only take place within the same site. This configuration would not cause any additional overhead for the administrative staff to move to another data center. It maintains full redundancy in case of a total site loss.

Some financial decision makers may question the expense involved in deploying more than a two-node campus cluster. However, using Solaris Resource Manager software, corporations can allocate some resources to nonclustered services on the remote systems, making use of otherwise idle resources while still reserving resources for failover in case of a disaster.

Quorum Devices in Campus Clusters

During cluster membership changes (for example, those caused by node outages), Sun Cluster 3.0 software uses a quorum mechanism to decide which nodes of the cluster are supposed to form the new cluster. Only a group of nodes with the majority of quorum votes may form a new cluster. This quorum mechanism helps ensure data integrity even in cases where cluster nodes cannot talk to each other because of broken interconnects. All other nodes not having majority are either shut down or prevented from accessing the data disks by means of reservation mechanisms in the storage layer (SCSI-2 and SCSI-3 persistent group reservations). Thus, only the nodes of a cluster with quorum have physical access to data.

Nodes and disks have quorum votes. By default, nodes have one vote. Dedicated quorum disks have as many votes as the sum of the nodes' votes attached to it, minus one. For example, dual-ported quorum disks have one vote, a four-node attached quorum disk has three votes. Because there must be a mechanism to break the tie in a two-node cluster, this configuration requires a quorum device. Sun Cluster 3.0 software enforces the configuration of a quorum device in these cases.

Quorum rules are not only valid for normal clusters, but for campus clusters as well. Because campus clusters are designed to protect against total site failures, it is important to understand the function of the quorum device in two- and three-site configurations.

For example, consider a typical two-site campus cluster setup. For the sake of simplicity, the quorum device (QD) is represented as a separate disk, which is not a requirement. It could be a disk in the data storage.

As shown in FIGURE 1, the quorum device (QD) is configured in site A. If site A fails completely, two out of three votes would be unavailable, leaving the node in site B with only one vote. The node in site B could not gain quorum and thus would shut down, leaving the whole cluster—even with a surviving node and a good copy of data in a local mirror—useless. When an operational node cannot communicate with its peer, it has no way of telling whether the communication paths or the peer itself are down. This situation is known as split brain. Shutting down node B in this scenario makes sense because it is unclear what has happened to A. If A is still alive, this action prevents data corruption. If it is down, the administrator has to decide how to proceed.

Without the quorum mechanism, each site could think it was the only survivor, then form a new cluster and start HA services. If both sites access the same data simultaneously, there is a high probability that they might cause data corruption.

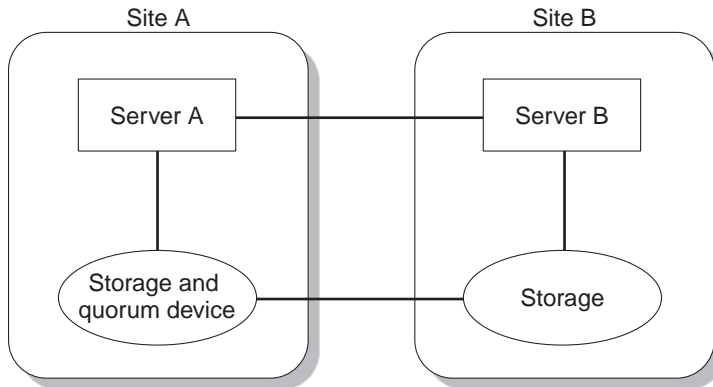


FIGURE 1 Two-Site Campus Cluster Configuration

In a configuration where one site is production and the other is either idle or running nonproduction work, the recommended practice is to configure the quorum device in the production site. If the remote site fails, the production site has enough votes to continue without interruption. If the production site fails, the problem cannot be overcome automatically with a two-site topology.

In this case, two options exist. Either an administrator must initiate a manual procedure to recover the quorum, or implement a third site for the quorum. Both methods are supported with Sun Cluster 3.0 software.

FIGURE 2 illustrates a three-site configuration. The quorum device is in a separate third site, C. In all scenarios where only one site is affected by a disaster, two remain in operation and provide one quorum vote each, so that the needed quorum of two votes is gained by the two surviving sites, that then form a new cluster. Therefore, a three-site configuration is highly recommended for enterprises that require fully automatic failover, even in case of a disaster affecting one site.

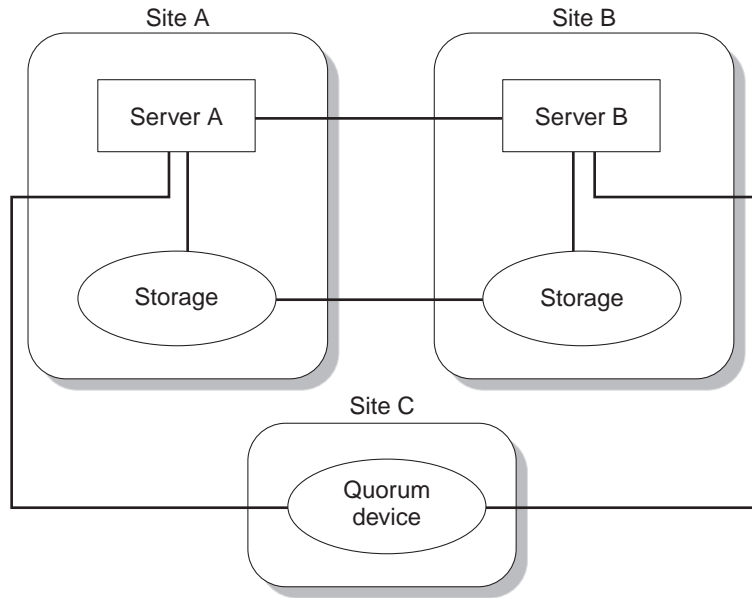


FIGURE 2 Three-Site Campus Cluster Configuration

As previously noted, enterprises may choose to use a manual procedure to recover from a loss of quorum. In this situation, the node that lost quorum is unavailable and cannot boot into cluster mode. This situation makes it necessary to change the quorum device definition in the cluster configuration repository (CCR) of the other node in the surviving site to an available quorum device, then to reboot this node into cluster mode.



Caution – Experience has shown that this technique is very error prone and requires highly skilled personnel to implement correctly. This option should be considered only if the cause of loss of quorum is a total loss of a site and no system in that site is accessing any data.

A final possibility is to use a third server in the third location to serve as the quorum. This node would then serve as the third vote in a three-site configuration.

Cluster Interconnect Hardware Configurations

Typical NICs can only be used with either copper or multimode fiber cables. However, the maximum distance can be extended by converting the media to single-mode fiber.

Transceivers for Fast Ethernet adapters plug into the RJ-45 or the MII port of the NIC and convert to single-mode fiber cables that can then span more than 15 km (in this combination). This type of transceiver has been qualified for campus clusters based on Sun Cluster 2.2 software.

Similar converters exist for Gigabit Ethernet to convert from multimode fiber to single-mode fiber. Using single-mode fiber, the maximum distance can be extended at least to 5 km. However, because the public network is not part of the cluster, it is up to the administrator to extend the network appropriately.

Data and Volume Manager Configuration

Mirroring data across sites helps to ensure that a copy of data survives any disaster. For campus clusters, host-based mirroring using a volume manager is recommended. However, special care should be taken when configuring Solaris™ Volume Manager (formerly Solstice DiskSuite™) software as the volume manager, especially when distributing replicas. (Refer to the Sun product documentation for details.)

Newer releases of volume managers tend to be equipped with more intelligence regarding placement policies for mirrors. Therefore, it is even more important to have control over this placement process. It is highly recommended to use the appropriate controls provided by the volume managers to spread mirrors across sites.

The prolonged distance between sites may introduce latency problems in accessing data. Volume managers offer a property called “preferred plex,” which directs read requests to the preferred local plex, thus avoiding the overhead of going to the remote storage.

Storage Configurations

Since the advent of Fibre Channel, extending the distance between servers and storage devices is no longer a problem. However, limitations in the maximum distance exist that may limit the usefulness of this technology in certain scenarios. Campus clusters using Sun Cluster 3.0 software today support the following:

- Upgrades from Sun Cluster 2.2 software campus cluster configurations using Sun Enterprise™ servers, Fibre Channel host bus adapters (code named SOC+) and Fibre Channel arbitrated loop (FC-AL), long wave GBICs (LWGBICS), and Sun StorEdge A5x00 storage systems
- Storage configurations using cascaded Fibre Channel switches with LWGBICs

Wave Division Multiplexers

In many areas, single-mode fiber is too costly or not available in sufficient quantities. A typical campus cluster configuration requires two wires for the storage, two for the cluster interconnect, and at least one for the public network. Additionally, many configurations have another network for backup purposes and one for administration that is connected to the terminal concentrator and other console ports. In total, a single campus cluster might need seven single-mode fiber connections.

WDMs use certain properties of the fiber to multiplex several streams onto a single fiber. Using WDMs over distances longer than 10 km has been successfully tested. This approach enables enterprises to deploy campus clusters in most geographic locations.

Campus Cluster Configurations

Cluster configurations are determined by the technology used to access the remote storage.

Sun Enterprise Servers and Sun StorEdge A5x00 Systems

Sun Enterprise servers (3500 to 10000 series) traditionally use SOC+ to attach to storage subsystems. These adapters use GBICs to convert signals to and from fiber to other media. The default is to use short wave GBICs with multimode fiber that can span a distance up to 500m using 50.0- μ fiber. The GBICs in these specific host bus adaptors (HBAs) can be replaced by long-wave GBICs that—using 9- μ single-mode fiber—can span a distance of up to 10 km. The same is true for the GBICs in the Sun StorEdge A5x00 subsystems. Using this technique, nodes and storage can be separated by a distance of up to 10 km without additional FC-switches.

Configurations using these technologies are in production today in many campus clusters around the world, based on Sun Cluster 2.2 software.

FIGURE 3 represents a typical campus cluster with Sun Enterprise servers and Sun StorEdge A5x00 systems in a three-site configuration.

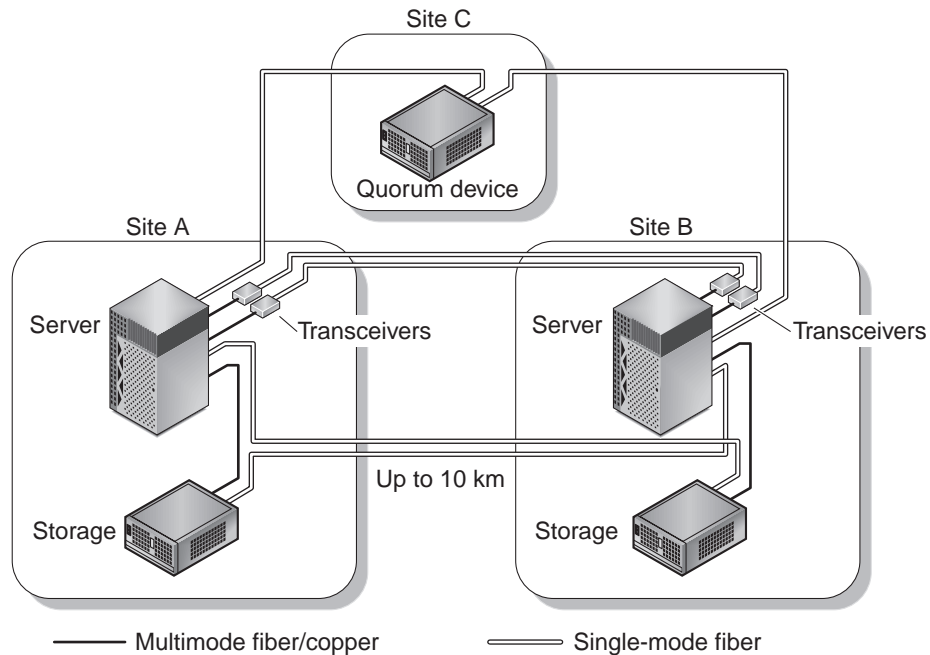


FIGURE 3 Three-Site Configuration With Sun Enterprise Servers and Sun StorEdge A5x00

Fibre Channel Switch-Based Configurations

The newer, fabric-capable Fibre Channel HBAs used by Sun volume servers and the new generation of Sun Fire™ servers do not allow for the replacement of the on-board short wave GBICs. Instead, the long distance is achieved by introducing Fibre Channel switches into the configuration. Sun's switches allow for the replacement of the GBICs to long-wave GBICs, so that one can connect two switches via a single-mode fiber over a distance up to 10 km. FIGURE 4 shows a typical configuration.

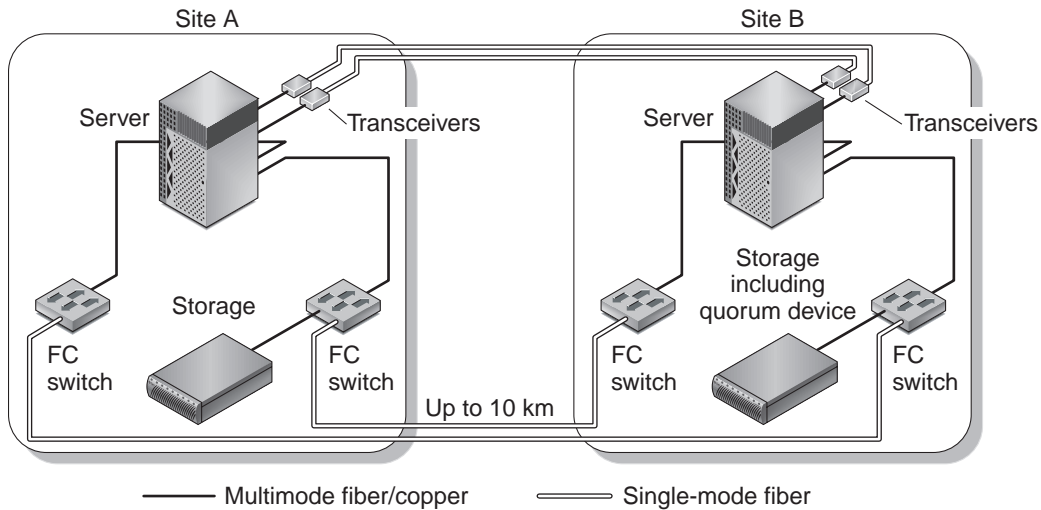


FIGURE 4 Two-Site Campus Cluster With T3WGs

This topology introduces two storage area networks (SANs). Data is mirrored across SANs. Additional storage may be attached to the FC switches. In that case, special care must be taken when choosing the mirror disks in the remote site. Mirrors must be on different SANs in different sites to avoid introducing single point of failure.

The configuration rules must be adhered to regarding firmware revisions, port configurations, topologies, and special restrictions for cascaded switches. Refer to the web site sun.com/storage/san for more details.

Sun Enterprise Servers With FC Switches

A new SBus Fibre Channel card opens up the possibility for the Sun Enterprise servers to be part of a full fabric. This adapter enables enterprises to connect such a server with this HBA to a Fibre Channel switch that is used to span the distance in a campus cluster configuration.

Campus Clusters Using Wave Division Multiplexers

WDMs allow for multiplexing several storage and network interconnects over a single fiber. This approach eliminates the need to have more than two fibers between the sites. However, to prevent a single point of failure, two WDMs are needed at

each site. This requirement adds significantly to the initial infrastructure cost, but saves money later due to the limited number of fibers needed to connect sites (other clusters or standalone systems can share the same WDM-based infrastructure).

FIGURE 5 depicts a campus cluster infrastructure built at two sites using WDMs.

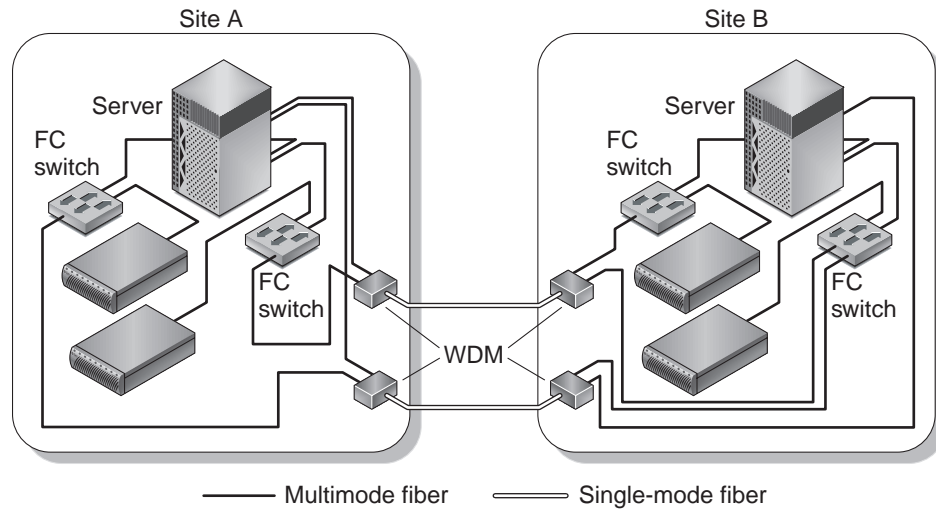


FIGURE 5 Two-Site Configuration Using WDMs

Other Configuration Considerations

IP Addresses and Subnets

High-availability services are generally reached through a unique IP address that is known through a name service to all of the clients. If one site in a campus cluster fails, this IP address must be failed over to another site. Therefore, all nodes in a cluster must be attached to the same public net (for example, the same IP subnet). This restriction is true for a campus cluster. To resolve this issue, all cluster nodes must be in the same broadcast domain.

Remote Access to All Consoles

It is a preferred practice to be able to access the console of any cluster node via the network (for example, using a terminal concentrator). In contrast to Sun Cluster 2.2 software, terminal concentrators are not mandatory in a Sun Cluster 3.0 infrastructure, due to a different failure fencing mechanism.

In case of failures in the production network and in the cluster interconnects, the only way to detect the status of the cluster nodes is either through the management network (if it is still operable) or through accessing the console ports using the terminal concentrator. Having a terminal concentrator at each physical location eliminates the need to inspect the physical systems or attach a terminal or laptop to the console port. If both of these methods fail, manual inspection is the only way to come to a final conclusion about what happened.

Communication Channels Between Sites

In a disaster situation, communication is critical. Planning for reliable communications is an important part of developing an overall campus cluster solution. Optional technologies include:

- Email – If the email system shares the data center infrastructure, it may not be available or reliable in a disaster.
- Telephones – Power supplies and phone lines may be coupled with the other parts of the infrastructure that has been damaged or destroyed.
- Cellular Phones – Although mobile phones are helpful for local disasters, they may be unreliable in larger disasters due to unreachable networks, limited range, and dependency on a complex infrastructure that may have been affected by the disaster.
- Voice Radio – Many security staffs have voice radios. Similar systems may be implemented for emergency communications.

Network Considerations for Client Access

Client access to a production system is often as vital as the production system itself. Additional “networking” connections such as ISDN or fax lines may be critical for normal business operations. Enterprises need to take care of client connectivity to the alternate remote site when designing their campus cluster configurations.

Performance in a Campus Cluster Environment

Data services running in a campus cluster environment may encounter performance degradation due to the latencies imposed by the distance between the nodes and the storage subsystems. Over long distances, the pure signal traveling time increases by a significant factor when compared with short distances. For example, a laser beam travels through a fiber optic cable at 0.2km/μs. Thus, a 40 km round trip adds a latency of 200 μs or 0.2 μs to all remote disk I/O operations and network transmissions between the sites. Network equipment may add to the latency.

Basic Performance Considerations

Because data must be mirrored to the remote site, some latency is unavoidable. The Sun Cluster 3.0 campus cluster environment requires the use of a volume manager product (such as Solaris Volume Manager software) for host-based remote data mirroring. The synchronous mirroring process introduces some latency. For example, a write operation is only complete if the write operations to all mirrors are completed. Fortunately, read operations are not affected if the preferred plex property is employed. This property can be used to direct the volume manager to a local plex for read operations.

Traffic routed over the cluster interconnect is affected by the distance of the nodes in the cluster infrastructure. This traffic includes intracluster traffic, network packets for scalable services, global file service (GFS) operations including data and replication information, and possibly user application traffic.

The GFS—also called the cluster file system (CFS)—uses the concept of a single primary server node for a CFS. Nodes that are not the primary server access the CFS through the cluster interconnect. Most applications today do not make heavy concurrent use of the CFS. A recommended practice for campus cluster environments is to ensure that the primary server node of a CFS runs on the same node as the application that uses that CFS. A special resource type available in Sun Cluster 3.0 software called HAStoragePlus can be used to help ensure collocation of the storage and services using that storage. HAStoragePlus has a special resource property called *AffinityOn* that, if set to “True,” provides exactly this functionality.

The failover file service (FFS through the HAStoragePlus resource type) may be deployed if there is no data requirement with CFS.

Heartbeats that use the cluster interconnect typically have time outs that are magnitudes higher than the latency even over a long distance. Due to the complexity of networks over long distances and the latencies introduced by additional hardware, the probability of a failure is much higher than in a local environment, so monitoring software must be configured accordingly.

Oracle Parallel Server and Oracle9i RAC

Enterprise environments increasingly deploy parallel databases to achieve greater scalability and service availability levels. In campus cluster environments, many companies may deploy the Oracle Parallel Server and Oracle9i Real Application Clusters (RAC) technologies. If the performance of an OPS/RAC configuration is latency bound on the interconnect or the storage, the longer distance between sites in a campus cluster most likely impacts the database performance negatively.

Performance Recommendations

The following recommendations help reduce possible performance impacts of wide-area campus clusters:

- Use the “preferred plex” property of a volume manager to achieve good read performance.
- Use the HASToragePlus resource type to configure the colocation of application and storage (using the AffinityOn resource property) or to configure failover filesystems if the information does not need to be accessed globally.
- Do intensive performance testing, especially for peak application usage levels, to ensure that unexpected performance degradation does not adversely affect the production environment.

Management Aspects of Campus Clusters

Processes

Well-defined processes play a vital role in ensuring timely disaster recovery with minimal data loss. Disasters by nature cannot be predicted, making it absolutely critical to establish tested procedures for recovery. Staff training and expertise, with clear lines of communication and decision-making, are essential. The procedures must be reviewed and audited regularly and updated or refined as necessary. Changes in technology, organizational structure, and other fundamentals must be accommodated as soon as possible. Finally, when a disaster occurs, a post-recovery analysis helps determine what went well, what went wrong, and what needs to be improved.

Administrator Skills

Because clusters are expected to provide very high service levels, most enterprises assign dedicated, specialized staff to administer cluster configurations. Intensive training in Sun Cluster software and other technologies, with detailed procedural run books, are required to help specialized administrators maintain the cluster environment at the highest possible service level. In addition to training and defined processes, the administrative staff needs to exercise some measure of creativity and flexibility to cope with the unexpected and unprecedented complexities of a disaster.

Administrators should have excellent skills in understanding the basic concepts of clustering, especially of the algorithms the cluster uses to decide which nodes are part of a new cluster and which are not. It is equally important to understand how mirroring in a remote environment works, and how it is possible to reconfigure complex storage and volume manager configurations. Most importantly, administrators must be able to apply investigative inquisitiveness in complex error scenarios to rapidly determine the impact of a disaster and take appropriate actions.

Monitoring and Stabilizing the Campus Cluster

Management infrastructure tools such as Sun™ Management Center 3.0 software can be used to help monitor the health of the campus cluster. Used either as a standalone solution or linked into the enterprise management framework, management tools enable administrators to quickly detect potential problems with individual nodes or interconnects.

In the event of a failure, administrators need to act quickly to determine the scenario. Any of the following failures may interrupt service availability:

- One site is totally unavailable
- Network connections, including the cluster interconnect between sites, are broken, but storage connections are still available
- Storage connections are broken, but network connections are up
- Network and storage connections are both unavailable

If the cluster or a new cluster is still operational, stabilizing the cluster mainly requires reconfiguring nodes and storage. Refer to the Sun Cluster 3.0 product documentation for specific reconfiguration procedures.

If a new cluster cannot be formed, (for example, due to loss of quorum), administrators must intervene. Care must be taken to avoid jeopardizing data integrity during manual procedures where cluster mechanisms might temporarily be disabled. Manually stabilizing the cluster prevents the formation of more than one cluster when nodes return to operation. Disabling power and removing all network and storage connections from a failed node are possible mechanisms for stabilizing a cluster. If the failed node is accessible, its cluster configuration should be changed so that the node does not try to rejoin the cluster automatically upon reboot.

Changing the Quorum Device

If a cluster cannot form a new cluster, manual intervention is necessary. Manual intervention may temporarily remove the quorum and failure fencing mechanisms of Sun Cluster 3.0. software. Therefore, it is essential to prevent more than one cluster from running at the same time. Otherwise, more than one node could access the shared data and cause data corruption.

In the case of a slowly approaching disaster, proactive measures should be applied, if possible. Administrators should be thoroughly trained in procedures for evacuating high-availability services and configuration information from the production site. If the quorum device is in the affected site, the administrator's first priority is to change the quorum device to one in the unaffected site. This task can be done using the SunPlex Manager tool, the `scconf` command at the command-line interface, or the `scsetup` menu interface.

Because the quorum reconfiguration must occur quickly, it is advisable to prepare a run book and special scripts for this situation. When deploying the reconfiguration procedure, administrators must choose a device that is positively in the unaffected cluster site or data center. Note that in a two-node cluster, the last quorum cannot be deleted. Administrators need first to add a second quorum, then delete the old one.

Furthermore, this procedure works only if the cluster has quorum. If the quorum and the other node are lost, then certified personnel must change the internal cluster configuration database and define a new quorum device.

Reconfiguring the Volume Manager

If access to storage in all sites (for example, all mirrors) is still available, no special procedures are necessary to protect data integrity. However, if an administrator determines that storage at the remote site is lost and must be replaced as part of the recovery, it is advisable to detach the mirrors located on these storage devices and remove the disks from the volume manager configuration. It is important to remove failed nodes from the cluster configuration.

Back to Normal Operations

Once the cluster is stabilized and data services are again available, the real recovery process can start. If there is no redundancy in the surviving data center, it is essential to decide how to establish this redundancy, especially on the data level, as soon as possible. This task can be achieved either by re-establishing a site or by adding storage and cluster nodes to the remaining site. Ideally, the steps required to re-establish redundancy are included in the preparatory process and documented in a run book.

Glossary

CCR	Cluster configuration repository
CFS	Cluster file system
DWDM	Dense wave division multiplexer
FCAL	Fibre Channel arbitrated loop
GBIC	Gigabit Interface Converter Module
GFS	Global file service
HBA	Host bus adapter
LWGBIC	Long wave GBIC
MII	Media Independent Interface
NIC	Network interface card
OPS	Oracle Parallel Server (before Oracle9i RAC)
PGR	Persistent group reservation
RAC	Real application cluster (Oracle9i version of OPS)
SAN	Storage area network
SCI	Scalable coherent interface
SRDF	Symmetrix remote data facility
VLAN	Virtual local area network
WDM	Wave division multiplexer

About the Author

Hartmut Streppel is an Enterprise Solutions Ambassador and a Sun Cluster Ace based in Germany. This article is the result of a special project and the results have been very successful for Sun customers around the world. Hartmut can be reached at Hartmut.Streppel@Sun.COM.

Related Resources

Publications

Sun Microsystems posts complete information on Sun's hardware and software products and service offerings in the form of data sheets and white papers on its Internet Web page at sun.com. Product documentation can be found at docs.sun.com.

Marcus, Evan and Hal Stern. *Blueprints for High Availability: Designing Resilient Distributed Systems*, John Wiley & Sons, ISBN: 0471356018, January 31, 2000.

Web Sites

- IEEE 802.3 Standards: standards.ieee.org
- Storage area networks (SANs) sun.com/storage/san