

Campus Clusters Based on Sun™ Cluster 3.0 Software

A Technical White Paper



Table of Contents

Introduction1
Evolution of Clusters	2
Cluster Limitations	2
People, Processes, and Products	3
Campus Clusters and the SunPlex Environment	3
Technology Options for Disaster Recovery Solutions4
Quick Checklist for Campus Cluster Deployments7
General Questions	7
Infrastructure Questions	8
People and Processes	9
Campus Cluster Maximum Distances10
Campus Cluster Topologies and Components12
Quorum Devices in Campus Clusters	13
Cluster Interconnect Hardware Configurations	14
Data and Volume Manager Configuration	15
Storage Configurations	15
Wave Division Multiplexers (WDMs)	15
Campus Cluster Configurations16
Sun Enterprise Servers and Sun StorEdge A5x00 Systems	16
Fibre Channel Switch-based Configurations	17
Sun Enterprise Servers With FC Switches	18
Campus Clusters Using Wave Division Multiplexers	18
Other Configuration Considerations	18
IP Addresses and Subnets	18
Remote Access to All Consoles	18
Communication Channels Between Sites	19
Network Considerations for Client Access	19
Performance in a Campus Cluster Environment20
Basic Performance Considerations	20
Oracle Parallel Server and Oracle RAC	21
Performance Recommendations	21
Management Aspects of Campus Clusters22

Processes	22
Administrator Skills	22
Monitoring and Stabilizing the Campus Cluster	23
Changing the Quorum Device	23
Reconfiguring the Volume Manager	24
Back to Normal Operations	24
Conclusion	25
References	26
Glossary	27

Chapter 1

Introduction

In the new economy, any service interruption, however small, corresponds to lost revenue and lost credibility. Ensuring continuous service availability must be a primary goal of the people, processes, and products that make up enterprise information technology (IT). 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.

A scalable, reliable infrastructure utilizing the latest technologies is an essential component of ensuring business continuity in the event of a minor or major disaster. But this infrastructure must also be combined with a set of carefully developed and tested processes, managed and monitored by well trained and motivated staff.

With the Sun™ Cluster 3.0 solution, enterprises can not only achieve high levels of business continuity, they can also implement straightforward disaster recovery processes that minimize loss in the event of a catastrophic disaster. A key component of the SunPlex™ environment, Sun Cluster 3.0 software extends the Solaris™ Operating Environment to provide virtually continuous service levels across geographically-dispersed — or campus — clusters.

This white paper describes how Sun Cluster 3.0 software might be used as part of a comprehensive disaster recovery solution to ensure continuous service availability. It provides basic guidelines for consideration when deploying a campus cluster solution and offers some helpful tips for setting up sound administrative practices.

Evolution of Clusters

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 have become more affordable and more widely used by enterprises of all types, clustering technology has also evolved to provide much greater flexibility, scalability, and manageability along with increasing levels of service availability.

Local clusters (i.e., clusters where all of the nodes¹ and storage subsystems are in the same room) play a major role in achieving business continuity by providing a solid level of continuous service availability. Basic clusters are typically deployed by using a physical interconnect (usually SCSI) to allow two or more servers to share access to data mirrored across storage subsystems. Due to the limits of SCSI technology, the maximum distance between cluster nodes is limited by the maximum cable length between one server, the shared storage, and the other server. While this configuration offers good protection against smaller disasters, such as node disk crashes, it does not protect against major disasters that could destroy or damage the facility site.

With the advent of Fibre Channel technology, it becomes possible to send mirrored data over much greater distances (up to several hundred kilometers), making clusters a more viable solution for wide-scale disaster protection. For example, without changing the software infrastructure, applications, or data, enterprises can extend the distance between nodes in a cluster to different buildings or even different locations within a wide geographic area. Depending on the distance, different combinations of technologies and management practices are used to survive a disaster.

One drawback of the extended cluster is the increased latency introduced when sending data over long fibers. As a result, additional factors such as distance, network backbone connections, application services, manageability, and increased infrastructure costs must also be considered when planning for and deploying a wide-area cluster solution.

Cluster Limitations

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 also cause a major disaster. In those cases, tape backup or some other hard-medium up-to-date copy of the data will prove invaluable for speedy recovery.

Even cluster software can fail, especially in the case of a major disaster affecting the cluster infrastructure. For example, a campus cluster in which 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 will want to deploy a multifaceted solution to ensure continuous service availability.

1. Terminology note: Throughout this paper, 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.

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 will not only deploy a comprehensive solution, it will also verify its processes, staff training, and technologies through regular (at least annual) testing. When personnel are prepared and trained to the highest level, best practices established and verified, and sound technologies deployed, enterprises can deliver the high levels of service continuity required to remain competitive in today's economy.

Campus Clusters and the SunPlex Environment

Built around the Sun Cluster 3.0 solution, the Solaris Operating Environment, 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.

Sun Cluster 3.0 software is designed to protect against single hardware or software failures such as node crashes or service interruptions. The software monitors the status of hardware and software components and initiates appropriate actions if a problem or failure occurs. (In the case of a catastrophic disaster, such as a loss of a total site, some recovery procedures must be manually initiated). For better reliability and performance, Sun Cluster 3.0 software is tightly integrated with the Solaris Operating Environment. This speeds up error detection times and makes the whole software stack more robust. However, the Sun Cluster 3.0 solution cannot address infrastructures that span extremely long distances, such as across a continent.

Depending on the failure, Sun Cluster 3.0 software will either failover the affected services to another node in the cluster or try to restart them. In all cases, the software's highest priority is to maintain data integrity regardless of whatever happens. This requirement drives the layout of the infrastructure and all of the algorithms in the product. 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 very high 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.

Chapter 2

Technology Options for Disaster Recovery Solutions

Extended — or campus — cluster solutions can form the base for comprehensive disaster recovery solutions that may also include any one or combination of the following technologies:

Backup and Recovery

Backup scripts or management software automatically generate hard-medium copies of data on a regular basis, which are then archived safely on site or at a remote location.

Outsourced Data Services

Some enterprises may choose to use an outside vendor to provide a replica of the production data center that can be installed quickly in case of a disaster.

Database Replication

Databases may be constructed 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’s SRDF and Hitachi Data Systems’ Hitachi TrueCopy.

Each of these options is adequate for certain situations, but most enterprises will likely choose a combination of several technologies in order to deploy a complete disaster recovery solution. Each option must be judged on how well it meets the 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.

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 also carries a risk: any defect that exists in the data is synchronously mirrored to the other copy, making both copies useless immediately. In this case, having a nonsynchronous copy of data, such as with backup and restore, would be 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 a time gap has elapsed. This would make 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 also 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. People, processes, and products must all play a significant role in ensuring business continuity.

TABLE 1 Comparison of Disaster Recovery Technologies

	Independent Data	Clustered	Automatic Recovery	Recovery Time	Maximum Distance (a)
Remote Backup & 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	No	No	No	Medium	Network
Campus Clusters	No	Yes	Yes	Fast	10 km (b)

(a) Distance negatively affects performance.

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

Chapter 3

Quick Checklist for Campus Cluster 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, enterprises need to consider their requirements, resources, and risks. The following checklist provides an overview of factors to consider when determining which level of solution is best for a particular enterprise.

General Questions

1. 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 50 kilometers away rather than 10.

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

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

Understanding the financial impacts of a disaster on business operations will help define an adequate solution. There are at least three factors to consider:

- 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, the 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 in place. 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 occurrence.

3. Do you understand that campus clusters do not address all aspects of a disaster recovery solution?

A campus cluster alone provides the appropriate infrastructure against many types of disasters, but it does not offer complete protection against all disasters. In addition, it must be accompanied by other mechanisms, such as backup and restore to help ensure business continuity. Finally, it needs the right people and processes in place to make the solution complete.

Infrastructure Questions

1. Are you prepared to accept performance degradation because of the 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 if the latency introduced by the long wires is only a small part of the overall latency, it adds up to a point where it is measurable. The additional components involved, such as transceivers, switches, and multiplexers, also add to the latency.

2. Does your data center infrastructure provide two or more sites?

Quorum devices help the cluster decide which nodes may form a new subcluster 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 in 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 in the third site, so that the loss of one site would not affect the majority of quorum votes. (Chapter 5 discusses quorum devices in more detail.)

3. 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. (Chapter 6 provides more detail).

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

Clusters failover IP addresses to maintain accessibility to high availability services. In order 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 (see Chapter 6).

People and Processes

1. Is your staff well-trained and willing to undergo constant training and exercising?

Having well-trained and experienced administrative staff is one of the key factors for achieving high availability. Ongoing testing of the disaster prevention and recovery processes is essential for minimizing service interruptions and restoring normal operations.

2. 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 already familiar with these processes will understand and internalize the changes faster.

Chapter 4

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 rooms. Conventional wiring technology does not span the distance needed for the cluster and storage interconnects. In many cases, use of third-party fiber networks and wave division multiplexers (WDMs) may be necessary to bridge the required distances of a geographically-dispersed cluster.

The following tables summarize 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/from fiber, both multimode and single mode. Maximum distances should be checked against product specifications for each specific deployment.

TABLE 2 Maximum Distances for Storage Interconnects

	Storage Interconnect
Short Wave Gigabit Internet Converter (SWGIC)	500 m
Long Wave (LWGBIC)	10000 m

TABLE 3 Maximum Distances for Gigabit Ethernet

	Cluster Interconnect Gigabit Ethernet 1000BaseX
SWGBIC (1000BaseSX)	220/550 m (a)
LWGBIC (1000BaseLX)	5000 m (b)

(a) With 62.5/50.0- μ multimode fiber (IEEE 802.3 2000 38.3). Most Gigabit Ethernet adapters today use SWGBICs (1000BaseSX).

(b) With 9.0- μ single-mode fiber (IEEE 802.3 2000 38.4).

TABLE 4 Maximum Distances for Fast Ethernet (f)

	Cluster Interconnect Fast Ethernet 100BaseX
100BaseTX (twisted pair)	100 m (c)
100BaseFX (half duplex, multimode fiber)	412 m (d)
100BaseFX (full duplex, multimode fiber)	2000 m (e)

(c) IEEE 802.3 2000 29.1.1

(d) IEEE 802.3 2000 29.1.1

(e) IEEE 802.3 2000 29.4

(f) IEEE did not set a distance limitation on 100BaseFX over single-mode fiber solutions.

Hardware vendors offer transceivers that can achieve more than 10 km using single-mode fiber.

Chapter 5

Campus Cluster Topologies and Components

There are a number of 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 would consist of four nodes and two-way, host-based mirrors across sites. To protect against local storage failures, controller-based RAID (such as RAID-5) should be used within the storage arrays. Multipathing solutions protect against failures of the storage paths and make it very unlikely that data must be completely resynchronized under normal circumstances. This configuration helps ensure that in most failures, failover would only take place within the same site. This would not cause any additional overhead for the administrative staff to move to another data center. It also would keep full redundancy in case of a total site loss.

Some enterprises may question the expense involved in deploying more than a two-node campus cluster. However, using Solaris Resource Manager software, enterprises 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. Sun Cluster 3.0 campus clusters will initially only support two-node configurations.

Quorum Devices in Campus Clusters

Sun Cluster 3.0 software uses a quorum in critical situations to decide which nodes of the cluster are supposed to form the new subcluster. This quorum mechanism helps ensure data integrity even in cases where cluster nodes cannot talk to each other because of broken interconnects. Only a set of nodes that owns the majority of votes can form a new subcluster. All other nodes not having majority will either shutdown or be prevented from accessing the data disks by means of reservation mechanisms in the storage layer (SCSI-2 reservations, SCSI-3 PGR). Thus, only the nodes of a subcluster with quorum have physical access to the data.

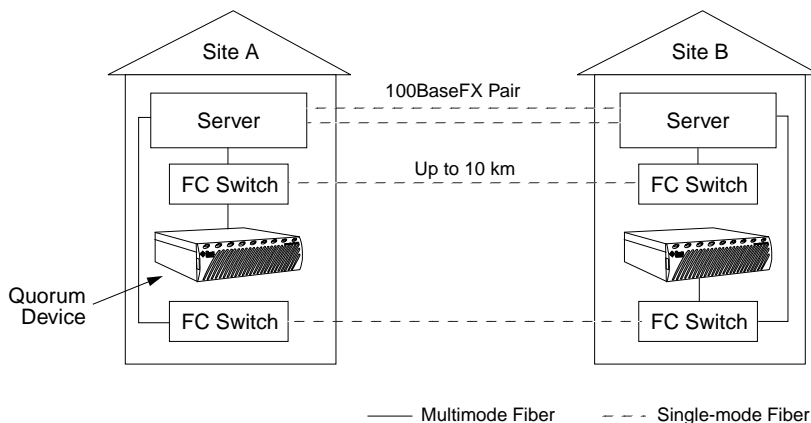
Nodes and disks have quorum votes. By default, nodes have one vote. Dedicated quorum disks have as many votes as the sum of the nodes' votes attached to it minus one; (i.e., dual-ported quorum disks have one vote, a four-node attached quorum disk has three votes). Because there must be a mechanism to break the tie in a two-node cluster, this configuration requires a quorum device. Sun Cluster 3.0 software enforces the definition 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 also designed to protect against total site failures, it is important for enterprises to understand the function of the quorum device in two- and three-room 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 shutdown, leaving the whole cluster — even with a surviving node and a good copy of the data in a local mirror — useless. This indeed makes sense because site B cannot know what has happened to site A. Without the quorum mechanism, both sites could think they were the survivors, form a new subcluster, access the data simultaneously, and produce data corruption immediately.

Figure 5-1: Two-site Campus Cluster Configuration

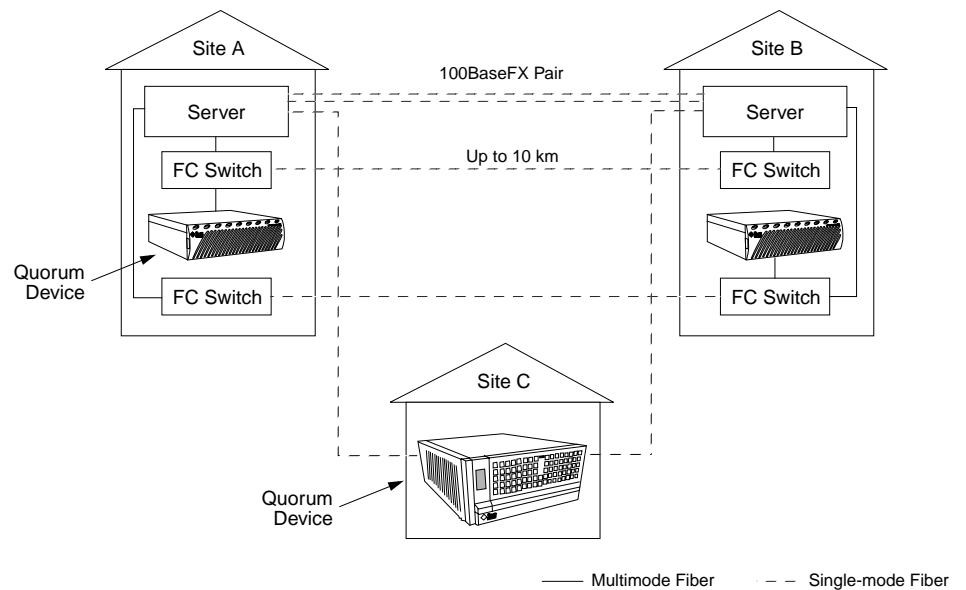


In a configuration where one site is production and the other is either idle or running nonproduction work, the recommended practice is to configure the quorum device in the production site. If the remote site should fail, the production site has enough votes to continue without interruption. Should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

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

Figure 2 illustrates the 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 will remain in operation and provide one quorum vote each, so that the needed quorum of two votes can be gained by the two surviving sites that then will form a new subcluster. Therefore, a three-site configuration is highly recommended for enterprises that require fully automatic failover even in case of a disaster affecting one entire site.

Figure 5-2: Three-site Campus Cluster Configuration



As previously noted, enterprises may also 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 makes it necessary to change the quorum device definition in the cluster configuration repository (CCR) of the other node in the surviving site to a still existing and available quorum device and then reboot this node into cluster mode.

Experience has shown that this technique is very error prone and requires highly skilled personnel to implement correctly. This option should only be considered 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

The network interface cards (NICs) normally installed in cluster nodes can only be linked together using typical data center cabling. However, the maximum distance can be extended by converting the media either to another type of fiber or from copper to fiber.

Transceivers for Fast Ethernet adapters plug into the RJ45 or the MII port of the NIC and convert to single-mode fiber cables that can then span up more than 15 km (in this combination). This type of transceiver has already 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, as the public network is not part of the cluster, it is up to the enterprise to extend the network appropriately.

Data and Volume Manager Configuration

Mirroring data across sites helps to ensure that a copy of the 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 more 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 also introduce latency problems in accessing the data. Volume managers offer a property called “preferred plex,” which — if configured correctly — 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. Still, limitations in the maximum distance exist that may limit the usefulness of this technology in certain customer scenarios. Campus clusters using Sun Cluster 3.0 software today support the following:

- Sun Cluster 2.2 software campus cluster configurations using Sun Enterprise™ servers, SOC+ FCAL host bus adapters, LWGBICs and Sun StorEdge A5x00 storage systems
- Storage configurations using cascaded Fibre Channel switches with LWGBICs

Wave Division Multiplexers (WDMs)

In many areas, single-mode fiber is either extremely costly or not available in the quantities needed. 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 could need at least seven single-mode fiber connections.

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

Chapter 6

Campus Cluster Configurations

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

Sun Enterprise Servers and Sun StorEdge A5x00 Systems

Sun Enterprise servers (3500 – 10000) traditionally use Fibre Channel Host Bus Adapters (HBA) (code named SOC+) to attach to storage subsystems. These adapters use Gigabit interface converters (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 500 m using 50.0- μ fiber. The GBICs in these specific 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 hardware-like 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.

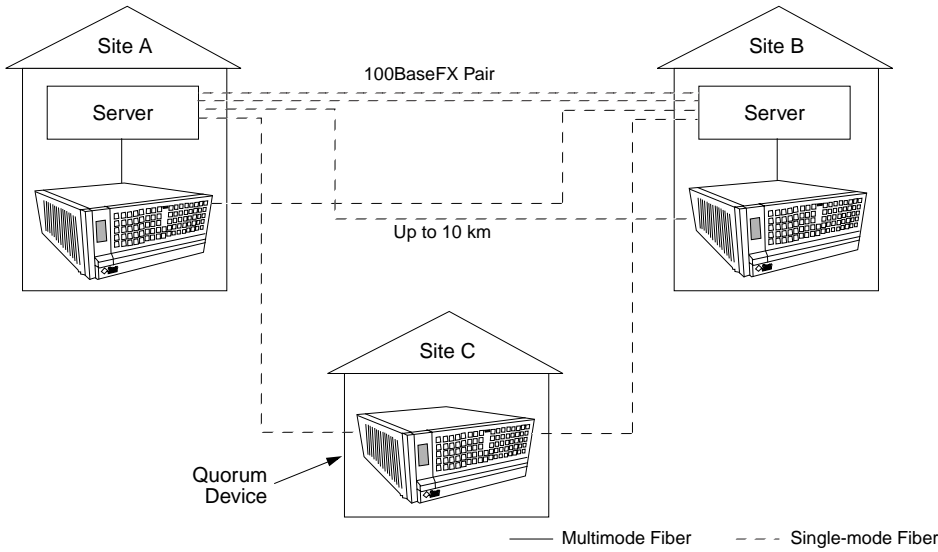


Figure 6-3: Three-site Configuration With Sun Enterprise Servers and Sun StorEdge A5x00

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.

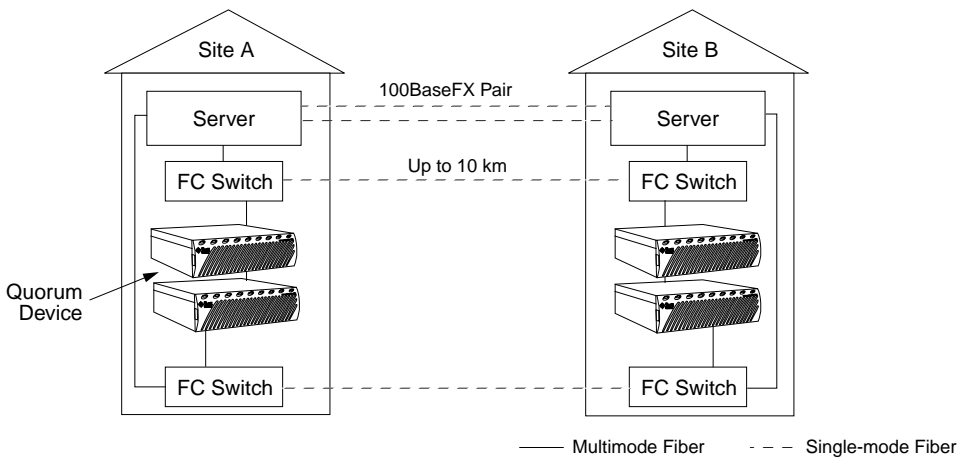


Figure 6-4: Two-site Campus Cluster With T3WGs

This topology introduces two storage area networks (SANs). Data is mirrored across SANs. If more storage than a single T3 has to be deployed at each site, it is initially not necessary to add more Fibre Channel switches to the configuration. If all of the Fibre Channel switches are used to attach additional storage, 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 regarding firmware revisions, port configurations, topologies, and special restrictions for cascaded switches must be adhered to. Refer to sun.com/storage/san for more details.

Sun Enterprise Servers With FC Switches

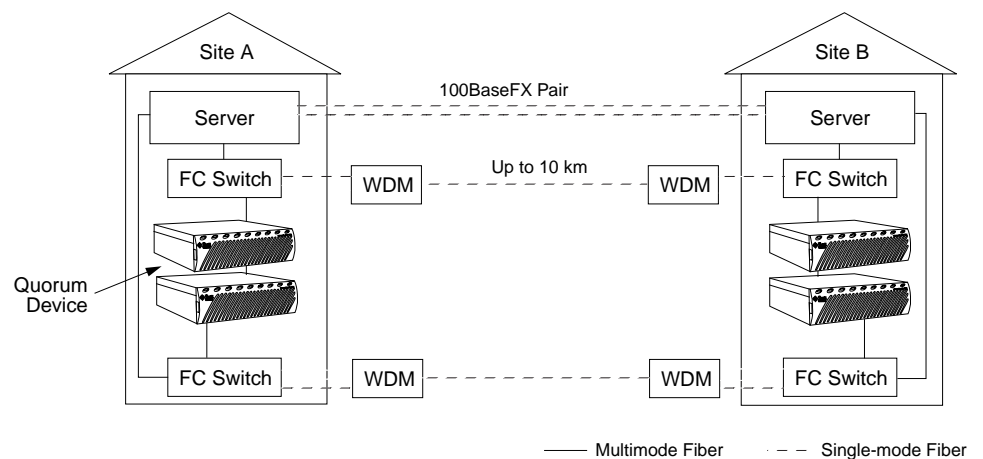
A new SBus Fiber Channel Arbitrated Loop (FCAL) card opens up the possibility for the Sun Enterprise servers to be part of a full fabric. This adapter also enables enterprises to connect such a server to a Fibre Channel switch that is used to span the distance in a campus cluster configuration.

Campus Clusters Using Wave Division Multiplexers

Wave division multiplexers (WDMs) allow for multiplexing several storage and network interconnects over a single fiber. This 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 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 could share the same WDM based infrastructure).

Figure 5 depicts a campus cluster infrastructure built in two sites using WDMs.

Figure 6-5: 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 should fail, this IP address must be failed over to another site. Therefore, all nodes in a cluster must be attached to the same public net (i.e., the same IP subnet). This restriction is also 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 best practice to be able to access the console of any cluster node via the network (e.g., 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 in place 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 has 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:

E-mail

If the e-mail system shares the data center infrastructure, it may not be available or reliable in a disaster.

Telephones

This is also a high risk; 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 also have been affected by the disaster.

Voice Radio

Many security staffs already have voice radios in use. 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.

Chapter 7

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 small distances. For example, a laser beam travels through a fiber optic cable at 5 $\mu\text{sec}/\text{km}$. Thus, a 40 km round trip (2 x 20 km) adds a latency of (40 x 5 μsec) 200 μsec or 0.2 msec to all remote disk I/O operations and network transmissions between the sites. Network equipment may also add to the latency.

Basic Performance Considerations

Because data must be mirrored to the remote site, some performance latency cannot be avoided. 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 (i.e., a write operation is only complete if the write operations to all mirrors have been completed) will always introduce some performance latency. Fortunately, read operations are not affected if the preferred plex property is employed. This property directs the volume manager to use the preferred — or most local — plex for read operations to minimize performance degradations.

Traffic routed over the cluster interconnect will also be affected by the distance of the nodes in the cluster infrastructure. This 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 do not host that primary server node have to 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 HASStorage can be used to help ensure collocation of the storage and the services using that storage. HASStorage has a special property called *AffinityOn* that, if set to “True”, provides exactly this functionality.

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

Heartbeats that also use the cluster interconnect usually 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 Oraclegi RAC

Enterprise environments increasingly deploy parallel databases in order to achieve greater scalability and service availability levels. In campus cluster environments, many companies may deploy the Oracle Parallel Server and Oraclegi Real Application Clusters (RAC) technologies. Because the latency of the interconnects and storage performance are the main factors influencing OPS/RAC performance, a long distance will always impact the database performance in a negative way.

Performance Recommendations

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

- Use the “preferred plex” property of a volume manager to achieve good read performance
- Use the HASStorage resource and its *AffinityOn* property to collocate the application to its storage
- Use of the HASStorage+ (FFS) resource may be considered for data that do not need to be globally accessible
- Do intensive performance testing, especially for peak application usage levels, to ensure that unexpected performance degradation will not adversely affect the production environment

Chapter 8

Management Aspects of Campus Clusters

Processes

Well-defined processes play a vital role in ensuring timely disaster recovery with minimal loss of data. Disaster by nature cannot be predicted, making it absolutely critical to establish tested procedures for recovery. Staff training and expertise, along with clear lines of communication and decision-making, are essential. The procedures must be reviewed and audited on a regular basis and updated or refined as necessary. Changes in technology, organizational structures, and other fundamentals must be accommodated as soon as possible. Finally, when a disaster does strike, a post-recovery analysis can help determine what went well, what went wrong, and what has 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, along with detailed procedural runbooks, are required to help these 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 will be part of a new subcluster and which are out. 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 the disaster and take the appropriate actions to restore service levels.

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 standalone solutions 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 suspected failure, administrators need to act quickly to determine the existing scenario. Any of the following failures may interrupt the service availability of the cluster:

- 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 subcluster 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 subcluster cannot be formed, (e.g., due to the loss of quorum), administrators must intervene. Care must be taken to avoid jeopardizing data integrity during manual maintenance procedures where cluster mechanisms might temporarily be disabled. Manually stabilizing the cluster will prevent the formation of more than one subcluster when nodes return to operation. Removing power or all of the network and storage connections from the failed node are possible mechanisms for stabilizing the 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 the cluster cannot form a new subcluster, 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 subcluster from running at the same time. Otherwise, more than one node could access the shared data and cause inconsistencies.

In the case of a slowly approaching disaster, such as a fire in another part of the building or impending flooding, proactive measures should be applied. 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 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 runbook and special scripts specifically 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 also that in a two-node cluster, the last quorum cannot be deleted. Administrators need to add a second quorum first and then delete the old one.

Furthermore, this procedure only works 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 (i.e., all mirrors) is still available, no special procedures are necessary to protect data integrity. However, if the administrators determine 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 also 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 fast as possible. This can be achieved either by reestablishing a site or by adding storage and cluster nodes to the remaining site. Ideally, the steps required to reestablish redundancy will have been included in the preparatory process and documented in a runbook.

Chapter 9

Conclusion

Disasters do not happen very 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 comprehensive, flexible, and scalable disaster recovery solution, campus clusters based on Sun Cluster 3.0 software 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 earthquake and fire.

Yet technology alone will not address all aspects of continuous service availability. In order to ensure the highest levels of business continuity, enterprises must invest in three essential component — 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.

Chapter 10

References

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

Other references for this document include:

- *Blueprints for High Availability: Designing Resilient Distributed Systems* by Evan Marcus and Hal Stern, ISBN 0-471-35601-8
- IEEE 802.3 Standards (standards.ieee.org)

Chapter 11

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

SUN™ Copyright 2002 Sun Microsystems, Inc., 901 San Antonio Road, Palo Alto, California 94303 U.S.A. All rights reserved.

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 U.S. and other countries, exclusively licensed through X/Open Company, Ltd.

Sun, Sun Microsystems, the Sun logo, Solaris, Sun Enterprise, Sun Fire, Sun Plex, and Sun Storedge are trademarks, registered trademarks, or service marks of Sun Microsystems, Inc. in the U.S. and other countries. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. in the U.S. 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.

RESTRICTED RIGHTS: Use, duplication, or disclosure by the U.S. Government is subject to restrictions of FAR 52.227-14(g)(2)(6/87) and FAR 52.227-19(6/87), or DFAR 252.227-7015(b)(6/95) and DFAR 227.7202-3(a).

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.

SUN™ Copyright 2002 Sun Microsystems, Inc., 901 San Antonio Road, Palo Alto, California 94303 États-Unis. Tous droits réservés.

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 déposée aux États-Unis et dans d'autres pays et licenciée exclusivement par X/Open Company, Ltd.

Sun, Sun Microsystems, le logo Sun, Solaris, Sun Enterprise, Sun Fire, Sun Plex, et Sun Storedge sont des marques de fabrique ou des marques déposées, ou marques de service, de Sun Microsystems, Inc. aux États-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 États-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.

CETTE PUBLICATION EST FOURNIE "EN L'ETAT" ET AUCUNE GARANTIE, EXPRESSE OU IMPLICITE, N'EST ACCORDEE, Y COMPRIS DES GARANTIES CONCERNANT LA VALEUR MARCHANDE, L'APTITUDE DE LA PUBLICATION A REpondre A UNE UTILISATION PARTICULIERE, OU LE FAIT QU'ELLE NE SOIT PAS CONTREFAISANTE DE PRODUIT DE TIERS. CE DENI DE GARANTIE NE S'APPLIQUERAIT PAS, DANS LA MESURE OU IL SERAIT TENU JURIDIQUEMENT NUL ET NON AVENU.



Please
Recycle



Adobe PostScript

Sun Microsystems, Inc. 4150 Network Circle, Santa Clara, CA 95054 USA Phone 800 786-7638 +1 512 434-1577 Web sun.com



Sun Worldwide Sales Offices: Africa (North, West and Central) +33-13-067-4680, Argentina +5411-4317-5600, Australia +61-2-9844-5000, Austria +43-1-60563-0, Belgium +32-2-704-8000, Brazil +55-11-5187-2100, Canada +905-477-6745, Chile +56-2-3724500, Colombia +571-629-2323, Commonwealth of Independent States +7-502-935-8411, Czech Republic +420-2-3300-9311, Denmark +45 4556 5000, Egypt +202-570-9442, Estonia +372-6-308-900, Finland +358-9-525-561, France +33-134-03-00-00, Germany +49-89-46008-0, Greece +30-1-618-8111, Hungary +36-1-489-8900, Iceland +354-563-3010, India-Bangalore +91-80-2298989/2295454; New Delhi +91-11-6106000; Mumbai +91-22-697-8111, Ireland +353-1-8055-666, Israel +972-9-9710500, Italy +39-02-641511, Japan +81-3-5717-5000, Kazakhstan +7-3272-466774, Korea +822-2193-5114, Latvia +371-750-3700, Lithuania +370-729-8468, Luxembourg +352-49 11 33 1, Malaysia +603-21161888, Mexico +52-5-258-6100, The Netherlands +00-31-33-45-15-000, New Zealand-Auckland +64-9-976-6800; Wellington +64-4-462-0780, Norway +47 23 36 96 00, People's Republic of China-Beijing +86-10-6803-5588; Chengdu +86-28-619-9333; Guangzhou +86-20-8755-5900; Shanghai +86-21-6466-1228; Hong Kong +852-2202-6688, Poland +48-22-8747800, Portugal +351-21-4134000, Russia +7-502-935-8411, Singapore +65-6438-1888, Slovak Republic +421-2-4342-9485, South Africa +27 11 256-6300, Spain +34-91-596-9900, Sweden +46-8-631-10-00, Switzerland-German 41-1-908-90-00; French 41-22-999-0444, Taiwan +886-2-8732-9933, Thailand +662-344-6888, Turkey +90-212-335-22-00, United Arab Emirates +9714-3366333, United Kingdom +44-1-276-20444, United States +1-800-555-9SUN or +1-650-960-1300, Venezuela +58-2-905-3800 08/02 FE1866-0