Understanding and Designing Serviceguard Disaster Tolerant Architectures

Manufacturing Part Number: T1906-90022
December 2007

Legal Notices

Contents

Contents

Printing History

Table 1 **Editions and Releases**

Printing Date	Part Number	Edition	Operating System Releases (see Note below)
December 2006	B7660-90018	Edition 1	HP-UX 11i v1 and 11i v2
September 2007	B7660-90020	Edition 2	HP-UX 11i v1, 11i v2 and 11i v3
December 2007	T1906-90022	Edition 3	HP-UX 11i v1, 11i v2 and 11i v3

The printing date and part number indicate the current edition. The printing date changes when a new edition is printed. (Minor corrections and updates which are incorporated at reprint do not cause the date to change.) The part number changes when extensive technical changes are incorporated. New editions of this manual will incorporate all material updated since the previous edition.

NOTE This document describes a group of separate software products that are released independently of one another. Not all products described in this document are necessarily supported on all the same operating system releases. Consult your product's *Release Notes* for information about supported platforms.

HP Printing Division:

ESS Software Division
Hewlett-Packard Co.
19111 Pruneridge Ave.
Cupertino, CA 95014

Preface

The following guides describe disaster tolerant clusters solutions using Serviceguard, Serviceguard Extension for RAC, Metrocluster Continuous Access XP, Metrocluster Continuous Access EVA, Metrocluster EMC SRDF, and Continentalclusters:

- ***Understanding and Designing Serviceguard Disaster Tolerant Architectures***
- ***Designing Disaster Tolerant HA Clusters Using Metrocluster and Continentalclusters***

The ***Understanding and Designing Serviceguard Disaster Tolerant Architectures*** user's guide provides an overview of Hewlett-Packard Disaster Tolerant high availability cluster technologies and how to configure an extended distance cluster using Serviceguard and Serviceguard Extension for RAC (SGeRAC). It is assumed you are already familiar with Serviceguard high availability concepts and configurations.

The contents are as follows:

- Chapter 1, *Disaster Tolerance and Recovery in a Serviceguard Cluster*, is an overview of disaster tolerant cluster configurations.
- Chapter 2, *Building an Extended Distance Cluster Using Serviceguard*, shows the creation of disaster tolerant cluster solutions using extended distance Serviceguard cluster configurations.

The ***Designing Disaster Tolerant HA Clusters Using Metrocluster and Continentalclusters*** user's guide provides detailed, task-oriented documentation on how to configure, manage, and set up disaster tolerant clusters using Metrocluster Continuous Access XP, Metrocluster Continuous Access EVA, Metrocluster EMC SRDF, Continentalclusters, and Three Data Center Architecture.

The contents are as follows:

- Chapter 1, *Designing a Metropolitan Cluster*, shows the creation of disaster tolerant cluster solutions using the metropolitan cluster products.

- Chapter 2, *Designing a Continental Cluster*, shows the creation of disaster tolerant solutions using the Continentalclusters product.
- Chapter 3, *Building Disaster Tolerant Serviceguard Solutions Using Metrocluster with Continuous Access XP*, shows how to integrate physical data replication via Continuous Access XP with metropolitan and continental clusters.
- Chapter 4, *Building Disaster Tolerant Serviceguard Solutions Using Metrocluster with Continuous Access EVA*, shows how to integrate physical data replication via Continuous Access EVA with metropolitan and continental clusters.
- Chapter 5, *Building Disaster Tolerant Serviceguard Solutions Using Metrocluster with EMC SRDF*, shows how to integrate physical data replication via EMC Symmetrix disk arrays. Also, it shows the configuration of a special continental cluster that uses more than two disk arrays.
- Chapter 6, *Designing a Disaster Tolerant Solution Using the Three Data Center Architecture*, shows how to integrate synchronous replication (for data consistency) and Continuous Access journaling (for long-distance replication) using Serviceguard, Metrocluster Continuous Access XP, Continentalclusters and HP StorageWorks XP 3DC Data Replication Architecture.

A set of appendixes and a glossary provide additional reference information.

Table 2 outlines the types of disaster tolerant solutions and their related documentation.

Guide to Disaster Tolerant Solutions Documents

Use the following table as a guide for locating specific Disaster Tolerant Solutions documentation:

Table 2 Disaster Tolerant Solutions Document Road Map

To Set up	Read
Extended Distance Cluster for Serviceguard/ Serviceguard Extension for RAC	***Understanding and Designing Serviceguard Disaster Tolerant Architectures*** • Chapter 1: *Disaster Tolerance and Recovery in a Serviceguard Cluster* • Chapter 2: *Building an Extended Distance Cluster Using Serviceguard*
Metrocluster with Continuous Access XP	***Understanding and Designing Serviceguard Disaster Tolerant Architectures*** • Chapter 1: *Disaster Tolerance and Recovery in a Serviceguard Cluster* ***Designing Disaster Tolerant HA Clusters Using Metrocluster and Continentalclusters*** • Chapter 1: *Designing a Metropolitan Cluster* • Chapter 3: *Building Disaster Tolerant Serviceguard Solutions Using Metrocluster with Continuous Access XP*
Metrocluster with Continuous Access EVA	***Understanding and Designing Serviceguard Disaster Tolerant Architectures*** • Chapter 1: *Disaster Tolerance and Recovery in a Serviceguard Cluster* ***Designing Disaster Tolerant HA Clusters Using Metrocluster and Continentalclusters*** • Chapter 1: *Designing a Metropolitan Cluster* • Chapter 4: *Building Disaster Tolerant Serviceguard Solutions Using Metrocluster with Continuous Access EVA*

Table 2 Disaster Tolerant Solutions Document Road Map (Continued)

To Set up	Read
Metrocluster with EMC SRDF	***Understanding and Designing Serviceguard Disaster Tolerant Architectures*** • Chapter 1: *Disaster Tolerance and Recovery in a Serviceguard Cluster* ***Designing Disaster Tolerant HA Clusters Using Metrocluster and Continentalclusters*** • Chapter 1: *Designing a Metropolitan Cluster* • Chapter 5: *Building Disaster Tolerant Serviceguard Solutions Using Metrocluster with EMC SRDF*
Continental Cluster	***Understanding and Designing Serviceguard Disaster Tolerant Architectures*** • Chapter 1: *Disaster Tolerance and Recovery in a Serviceguard Cluster* ***Designing Disaster Tolerant HA Clusters Using Metrocluster and Continentalclusters*** • Chapter 2: *Designing a Continental Cluster*
Continental Cluster using Continuous Access XP data replication	***Understanding and Designing Serviceguard Disaster Tolerant Architectures*** • Chapter 1: *Disaster Tolerance and Recovery in a Serviceguard Cluster* ***Designing Disaster Tolerant HA Clusters Using Metrocluster and Continentalclusters*** • Chapter 2: *Designing a Continental Cluster* • Chapter 3: *Building Disaster Tolerant Serviceguard Solutions Using Metrocluster with Continuous Access XP*

Table 2 Disaster Tolerant Solutions Document Road Map (Continued)

To Set up	Read
Continental Cluster using Continuous Access EVA data replication	***Understanding and Designing Serviceguard Disaster Tolerant Architectures*** • Chapter 1: *Disaster Tolerance and Recovery in a Serviceguard Cluster* ***Designing Disaster Tolerant HA Clusters Using Metrocluster and Continentalclusters*** • Chapter 2: *Designing a Continental Cluster* • Chapter 4: *Building Disaster Tolerant Serviceguard Solutions Using Metrocluster with Continuous Access EVA*
Continental Cluster using EMC SRDF data replication	***Understanding and Designing Serviceguard Disaster Tolerant Architectures*** • Chapter 1: *Disaster Tolerance and Recovery in a Serviceguard Cluster* ***Designing Disaster Tolerant HA Clusters Using Metrocluster and Continentalclusters*** • Chapter 2: *Designing a Continental Cluster* • Chapter 5: *Building Disaster Tolerant Serviceguard Solutions Using Metrocluster with EMC SRDF*
Continental Cluster using other data replication	***Understanding and Designing Serviceguard Disaster Tolerant Architectures*** • Chapter 1: *Disaster Tolerance and Recovery in a Serviceguard Cluster* ***Designing Disaster Tolerant HA Clusters Using Metrocluster and Continentalclusters*** • Chapter 2: *Designing a Continental Cluster*

Table 2 Disaster Tolerant Solutions Document Road Map (Continued)

To Set up	Read
Three Data Center Architecture	***Understanding and Designing Serviceguard Disaster Tolerant Architectures*** • Chapter 1: *Disaster Tolerance and Recovery in a Serviceguard Cluster* ***Designing Disaster Tolerant HA Clusters Using Metrocluster and Continentalclusters*** • Chapter 1: *Designing a Metropolitan Cluster* • Chapter 2: *Designing a Continental Cluster* • Chapter 6: *Designing a Disaster Tolerant Solution Using the Three Data Center Architecture*
Maintenance Mode, Disaster Recovery Rehearsal, Data Replication Storage Failover Preview	***Understanding and Designing Serviceguard Disaster Tolerant Architectures*** • Chapter 1: *Disaster Tolerance and Recovery in a Serviceguard Cluster* ***Designing Disaster Tolerant HA Clusters Using Metrocluster and Continentalclusters*** • Chapter 1: *Designing a Metropolitan Cluster* • Chapter 2: *Designing a Continental Cluster*

On-line versions of the above documents and other HA documentation are available at **`http://docs.hp.com`** -> High Availability.

Related Publications

The following documents contain additional useful information:

- *Clusters for High Availability: a Primer of HP Solutions, Second Edition*. Hewlett-Packard Professional Books: Prentice Hall PTR, 2001 (ISBN 0-13-089355-2)
- *Designing Disaster Tolerant HA Clusters Using Metrocluster and Continentalclusters* user's guide (B7660-90021)
- *Managing Serviceguard Fourteenth Edition* (B3936-90117)
- *Using Serviceguard Extension for RAC* (T1859-90048)
- *Using High Availability Monitors* (B5736-90025)
- *Using the Event Monitoring Service* (B7612-90015)

When using VxVM storage with Serviceguard, refer to the following:

- *VERITAS Volume Manager Administrator's Guide*. This contains a glossary of VERITAS terminology.
- *VERITAS Volume Manager Storage Administrator Administrator's Guide*
- *VERITAS Volume Manager Migration Guide*
- *VERITAS Volume Manager for HP-UX Release Notes*

Use the following URL for access to a wide variety of HP-UX documentation:

- **`http://docs.hp.com/hpux`**

Problem Reporting

If you have problems with HP software or hardware products, please contact your HP support representative.

1 Disaster Tolerance and Recovery in a Serviceguard Cluster

This guide introduces a variety of Hewlett-Packard high availability cluster technologies that provide disaster tolerance for your mission-critical applications. It is assumed that you are already familiar with Serviceguard high availability concepts and configurations.

This chapter introduces the basic concepts of disaster tolerance and disaster recovery within the context of a Serviceguard cluster. The following topics are covered:

- Evaluating the Need for Disaster Tolerance
- What is a Disaster Tolerant Architecture?
- Understanding Types of Disaster Tolerant Clusters
- Disaster Tolerant Architecture Guidelines
- Managing a Disaster Tolerant Environment

Evaluating the Need for Disaster Tolerance

Disaster tolerance is the ability to restore applications and data within a reasonable period of time after a disaster. Most think of fire, flood, and earthquake as disasters, but a **disaster** can be any event that unexpectedly interrupts service or corrupts data in an entire data center: the backhoe that digs too deep and severs a network connection, or an act of sabotage.

Disaster tolerant architectures protect against unplanned down time due to disasters by geographically distributing the nodes in a cluster so that a disaster at one site does not disable the entire cluster. To evaluate your need for a disaster tolerant solution, you need to weigh:

- *Risk of disaster.* Areas prone to tornadoes, floods, or earthquakes may require a disaster recovery solution. Some industries need to consider risks other than natural disasters or accidents, such as terrorist activity or sabotage.

  The type of disaster to which your business is prone, whether it is due to geographical location or the nature of the business, will determine the type of disaster recovery you choose. For example, if you live in a region prone to big earthquakes, you are not likely to put your alternate or backup nodes in the same city as your primary nodes, because that sort of disaster affects a large area.

  The frequency of the disaster also plays an important role in determining whether to invest in a rapid disaster recovery solution. For example, you would be more likely to protect from hurricanes that happen every season, rather than protecting from a dormant volcano.

- *Vulnerability of the business.* How long can your business afford to be down? Some parts of a business may be able to endure a 1 or 2 day recovery time, while others need to recover in a matter of minutes. Some parts of a business only need local protection from single outages, such a node failure. Other parts of a business may need both local protection and protection in case of site failure.

  It is important to consider the role the data servers play in your business. For example, you may target the assembly line production servers as most in need of quick recovery. But if the most likely disaster in your area is an earthquake, it would render the assembly

line inoperable as well as the computers. In this case disaster recovery would be moot, and local failover is probably the more appropriate level of protection.

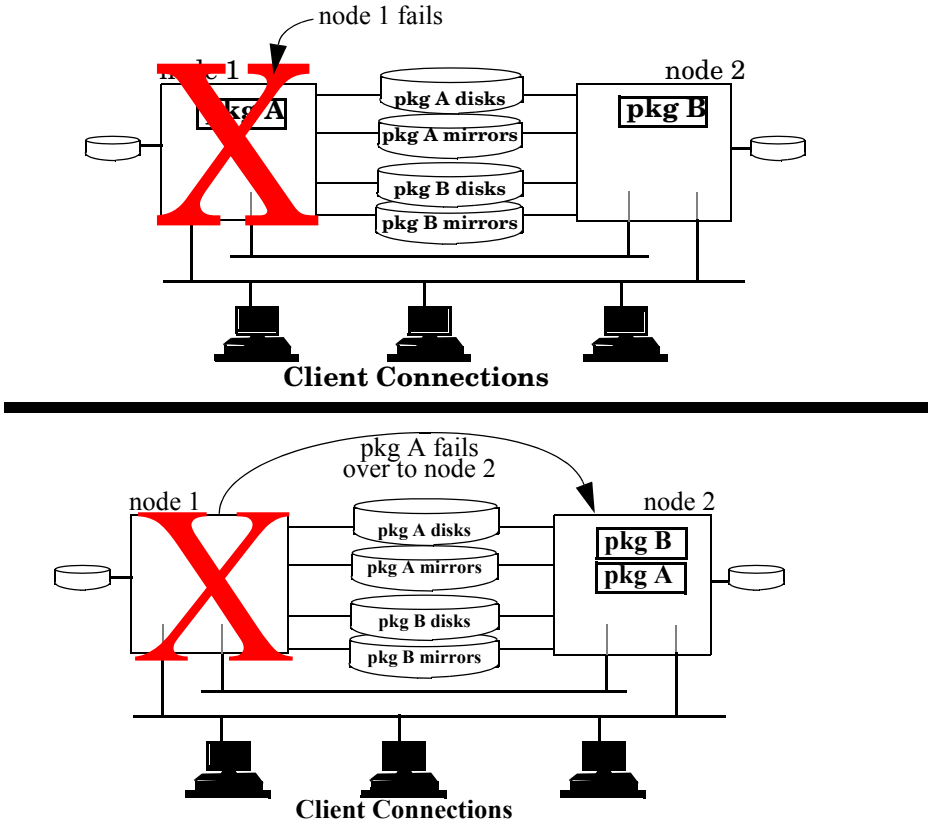
On the other hand, you may have an order processing center that is prone to floods in the winter. The business loses thousands of dollars a minute while the order processing servers are down. A disaster tolerant architecture is appropriate protection in this situation.

Deciding to implement a disaster recovery solution really depends on the balance between risk of disaster, and the vulnerability of your business if a disaster occurs. The following pages give a high-level view of a variety of disaster tolerant solutions and sketch the general guidelines that you should follow in developing a disaster tolerant computing environment.

What is a Disaster Tolerant Architecture?

In an Serviceguard cluster configuration, high availability is achieved by using redundant hardware to eliminate single points of failure. This protects the cluster against hardware faults, such as the node failure in Figure 1-1.

Figure 1-1 High Availability Architecture.

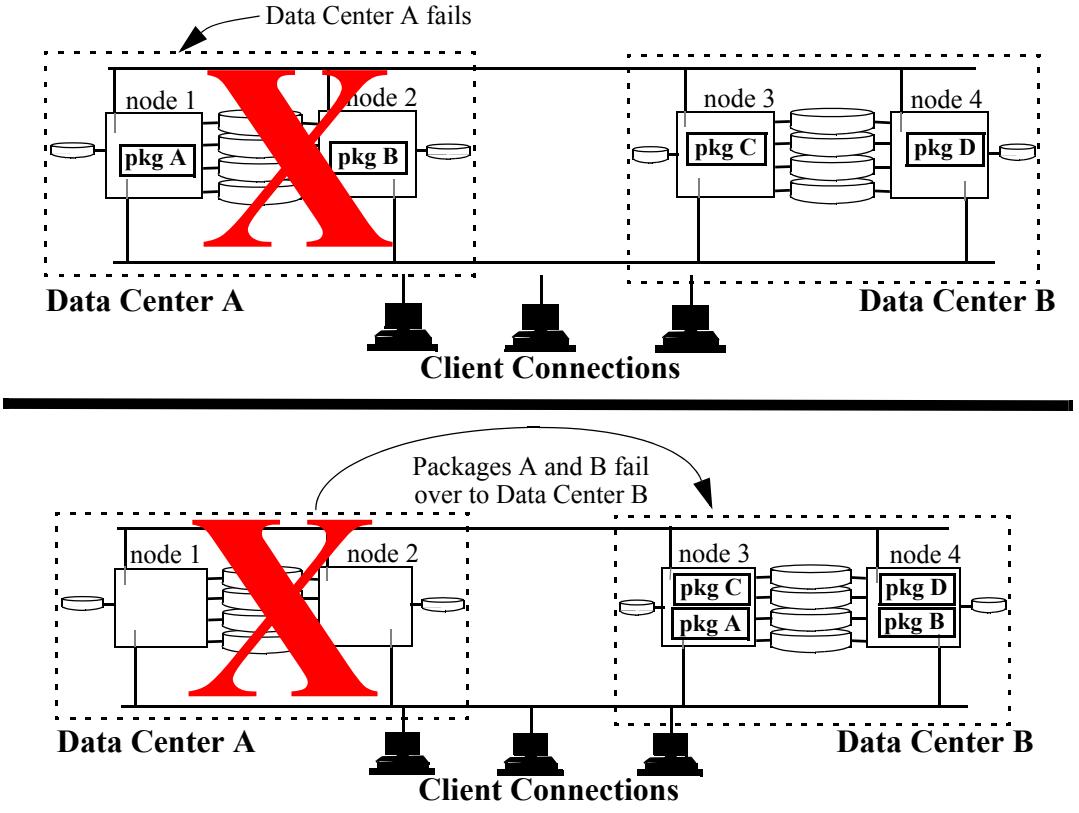


This architecture, which is typically implemented on one site in a single data center, is sometimes called a **local cluster**. For some installations, the level of protection given by a local cluster is insufficient. Consider the order processing center where power outages are common during harsh weather. Or consider the systems running the stock market, where multiple system failures, for any reason, have a significant financial

impact. For these types of installations, and many more like them, it is important to guard not only against single points of failure, but against **multiple points of failure (MPOF)**, or against single massive failures that cause many components to fail, such as the failure of a data center, of an entire site, or of a small area. A **data center**, in the context of disaster recovery, is a physically proximate collection of nodes and disks, usually all in one room.

Creating clusters that are resistant to multiple points of failure or single massive failures requires a different type of cluster architecture called a **disaster tolerant architecture**. This architecture provides you with the ability to fail over automatically to another part of the cluster or manually to a different cluster after certain disasters. Specifically, the disaster tolerant cluster provides appropriate failover in the case where a disaster causes an entire data center to fail, as shown in Figure 1-2.

Figure 1-2 Disaster Tolerant Architecture



Understanding Types of Disaster Tolerant Clusters

To protect against multiple points of failure, cluster components must be geographically dispersed: nodes can be put in different rooms, on different floors of a building, or even in separate buildings or separate cities. The distance between the nodes is dependent on the types of disaster from which you need protection, and on the technology used to replicate data. Three types of disaster-tolerant clusters are described in this guide:

- Extended Distance Clusters
- Metropolitan Cluster
- Continental Cluster

These types differ from a simple local cluster in many ways. Extended distance clusters and metropolitan clusters often require right-of-way from local governments or utilities to lay network and data replication cable. This can complicate the design and implementation. They also require a different kind of control mechanism for ensuring that data integrity issues do not arise, such as a quorum server. Typically, metropolitan clusters use an arbitrator site containing additional cluster nodes instead of the cluster lock disk. Continental clusters span great distances and operate by replicating data between two completely separate local clusters.

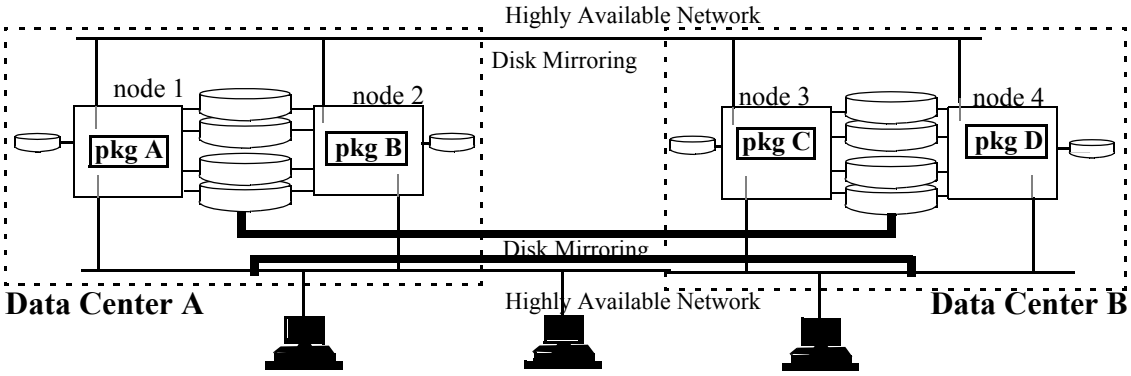
Extended Distance Clusters

The two types of Extended Distance Cluster configurations are **Extended Distance Cluster** and **Extended Distance Cluster for RAC**. Both types use Serviceguard to create disaster tolerant High Availability clusters. The following describes in more detail the key differences between the two types:

An **Extended Distance Cluster** (also known as **extended campus cluster**) is a normal Serviceguard cluster that has alternate nodes located in different data centers separated by some distance. Extended distance clusters are connected using a high speed cable that guarantees network access between the nodes as long as all guidelines for disaster tolerant architecture are followed. Extended distance clusters were

formerly known as **campus clusters**, but that term is not always appropriate because the supported distances have increased beyond the typical size of a single corporate campus. The maximum distance between nodes in an Extended Distance Cluster is set by the limits of the data replication technology and networking limits. An Extended Distance Cluster is shown in Figure 1-3.

Figure 1-3 Extended Distance Cluster



Extended distance clusters can be configured over shorter distances using FibreChannel mass storage, or over distances as great as 100 km using storage and networking routed over links extended via DWDM.

In extended distance architecture, each clustered server is directly connected to all storage in both data centers. With direct access to remote storage devices from a local server, an Extended Distance Cluster with up to four nodes can be designed with two data centers using dual cluster lock disks for cluster quorum. If the cluster size is greater than four nodes, an Extended Distance Cluster can be designed with two data centers and a third location housing arbitrator nodes or quorum server. Architecture and configuration requirements for several types of extended distance clusters are described more fully in Chapter 2, “Building an Extended Distance Cluster Using Serviceguard.”

Benefits of Extended Distance Cluster

- This configuration implements a single Serviceguard cluster across two data centers, and uses either MirrorDisk/UX or Veritas VxVM mirroring from Symantec for data replication. No (cluster) license beyond Serviceguard is required for this solution, making it the least expensive to implement.
- You may choose any storage supported by Serviceguard, and the storage can be a mix of any Serviceguard-supported storage.
- This configuration may be the easiest to understand and manage, as it is similar in many ways to Serviceguard.
- Application failover is minimized. All disks are available to all nodes, so that if a primary disk fails but the node stays up and the replica is available, there is no failover (that is, the application continues to run on the same node while accessing the replica).
- Data copies are peers, so there is no issue with reconfiguring a replica to function as a primary disk after failover.
- Writes are synchronous, unless the link or disk is down, so data remains current between the primary disk and its replica.

Extended Distance Cluster for RAC

An **Extended Distance Cluster for RAC** merges Extended Distance Cluster with Serviceguard Extension for RAC (SGeRAC). SGeRAC is a specialized configuration that enables Oracle Real Application Clusters (RAC) to run in an HP-UX environment on high availability clusters. RAC in a Serviceguard environment lets you maintain a single (Oracle) database image that is accessed by the servers in parallel in an active/active configuration, thereby providing greater processing power without the overhead of administering separate databases.

Benefits of Extended Distance Cluster for RAC

- In addition to the benefits of Extended Distance Cluster, RAC runs in active/active mode in the cluster, so that all resources in both data centers are utilized. The database and data are synchronized and replicated across two data centers up to 100km apart. In the event of a site failure, no failover is required since the instance is already running at the remote site.
- Extended Cluster for RAC implements SLVM so that SGeRAC has a "built-in" mechanism for determining the status of volume group extents in both data centers (that is, the state of the volume groups is kept in memory at the remote site), and SLVM will not operate on non-current data. For Veritas environments, Extended Cluster for RAC is also supported with *Veritas Cluster Volume Manager (CVM)*.

NOTE For the most up-to-date support and compatibility information see the *SGeRAC for SLVM, CVM & CFS Matrix,* and the *Serviceguard Compatibility and Feature Matrix* on **`http://docs.hp.com`** -> High Availability.

Metropolitan Cluster

A **metropolitan cluster** is a cluster that has alternate nodes located in different parts of a city or in adjacent cities. Putting nodes further apart increases the likelihood that alternate nodes will be available for failover in the event of a disaster. The architectural requirements are the same as for an Extended Distance Cluster, with the additional constraint of a third location for arbitrator node(s) or quorum server. And as with an Extended Distance Cluster, the distance separating the nodes in a metropolitan cluster is limited by the data replication and network technology available.

NOTE While it is possible to configure physical data replication through products such as HP's XP Series disk arrays with Continuous Access XP or Symmetrix EMC SRDF, it is still necessary to provide for high availability at the local level through RAID or mirroring.

In addition, there is no hard requirement on how far the third location has to be from the two main data centers. The third location can be as close as the room next door with its own power source or can be as far as in a site across town. The distance between all three locations dictates the level of disaster tolerance a metropolitan cluster can provide.

Metropolitan cluster architecture is implemented through the following HP products:

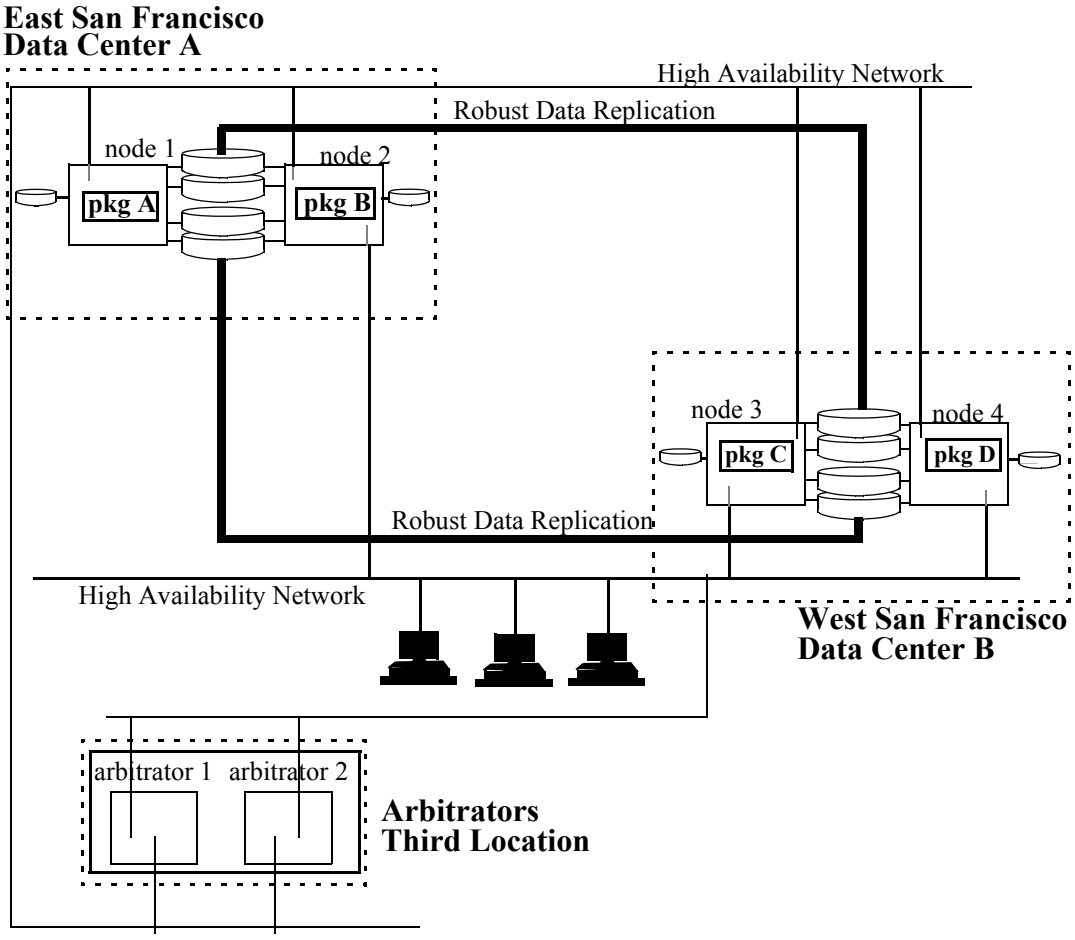
- Metrocluster with Continuous Access XP
- Metrocluster with Continuous Access EVA
- Metrocluster with EMC SRDF

The above products are described fully in Chapters 3, 4, and 5 of the *Designing Disaster Tolerant HA Clusters Using Metrocluster and Continentalclusters* user's guide.

On-line versions of the above document and other HA documentation are available at **`http://docs.hp.com`** -> High Availability.

Metropolitan cluster architecture is shown in Figure 1-4.

Figure 1-4 Metropolitan Cluster



A key difference between extended distance clusters and metropolitan clusters is the data replication technology used. The Extended Distance Cluster uses FibreChannel and HP-UX supported software mirroring for data replication. Metropolitan clusters provide extremely robust hardware-based data replication available with specific disk arrays based on the capabilities of the HP StorageWorks Disk Array XP series, StorageWorks EVA, or the EMC Symmetrix array.

Benefits of Metrocluster

- Metrocluster offers a more resilient solution than Extended Distance Cluster, as it provides full integration between Serviceguard's application package and the data replication subsystem. The storage subsystem is queried to determine the state of the data on the arrays.
  Metrocluster knows that application package data is replicated between two data centers. It takes advantage of this knowledge to evaluate the status of the local and remote copies of the data, including whether the local site holds the primary copy or the secondary copy of data, whether the local data is consistent or not and whether the local data is current or not. Depending on the result of this evaluation, Metrocluster decides if it is safe to start the application package, whether a resynchronization of data is needed before the package can start, or whether manual intervention is required to determine the state of the data before the application package is started.
  Metrocluster allows for customization of the startup behavior for application packages depending on your requirements, such as data currency or application availability. This means that by default, Metrocluster will always prioritize data consistency and data currency over application availability. If, however, you choose to prioritize availability over currency, you can configure Metrocluster to start up even when the state of the data cannot be determined to be fully current (but the data is consistent).
- Metrocluster Continuous Access XP and Metrocluster EMC SRDF support synchronous and asynchronous replication modes, allowing you to prioritize performance over data currency between the data centers.
- Because data replication and resynchronization are performed by the storage subsystem, Metrocluster may provide significantly better performance than Extended Distance Cluster during recovery. Unlike Extended Distance Cluster, Metrocluster does not require any additional CPU time, which minimizes the impact on the host.
- There is little or no lag time writing to the replica, so the data remains current.
- Data can be copied in both directions, so that if the primary site fails and the replica takes over, data can be copied back to the primary site when it comes back up.

- Disk resynchronization is independent of CPU failure (that is, if the hosts at the primary site fail but the disk remains up, the disk knows it does not have to be resynchronized).
- Metrocluster Continuous Access XP is supported in a Three Data Center solution, providing the data consistency of synchronous replication and the capability of CA journaling replication to protect against local and wide-area disasters.

  The Three Data Center solution integrates Serviceguard, Metrocluster Continuous Access XP, Continentalclusters and HP StorageWorks XP 3DC Data Replication Architecture. This configuration consists of two Serviceguard clusters. The first cluster, which is basically a Metrocluster, has two data centers namely Primary data center (DC1) and Secondary data center (DC2). The second cluster, a normal Serviceguard cluster, has only one data center namely Third data center (DC3). Continuous Access synchronous replication is used within the Metrocluster region and Continuous Access long-distance journal replication is used between the Metrocluster and recovery cluster regions.
- Metrocluster supports **Data Replication Storage Failover Preview**; allows you to preview the preparation for the storage of the data replication environment in a Metrocluster failover or Continentalclusters recovery. See “Data Replication Storage Failover Preview”.

Differences Between Extended Distance Cluster and Metrocluster

The *major* differences between an Extended Distance Cluster and a Metrocluster are:

- The methods used to replicate data between the storage devices in the two data centers. The two basic methods available for replicating data between the data centers for HP-UX clusters are either host-based or storage array-based. Extended Distance Cluster always uses host-based replication (either MirrorDisk/UX or Veritas VxVM mirroring). Any (mix of) Serviceguard supported storage can be implemented in an Extended Distance Cluster. Metrocluster always uses array-based replication/mirroring, and requires storage from the same vendor in both data centers (that is, a pair of XPs with Continuous Access, a pair of Symmetrix arrays with SRDF, or a pair of EVAs with Continuous Access).

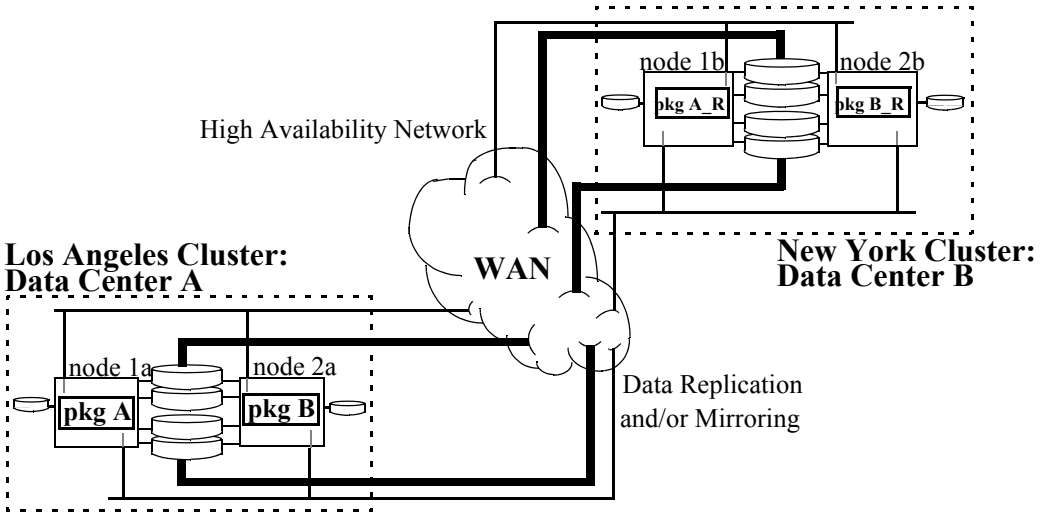
- Data centers in an Extended Distance Cluster can span up to 100km, whereas the distance between data centers in a Metrocluster is defined by the *shortest* of the following distances:
  - the maximum distance that guarantees a network latency of no more than 200ms
  - the maximum distance supported by the data replication link
  - the maximum supported distance for DWDM as stated by the provider
- In an Extended Distance Cluster, there is no built-in mechanism for determining the state of the data being replicated. When an application fails over from one data center to another, the package is allowed to start up if the volume group(s) can be activated. A Metrocluster implementation provides a higher degree of data integrity; that is, the application is only allowed to start up based on the state of the data and the disk arrays.
- Extended Distance Cluster supports active/active access by implementing SGeRAC, whereas Metrocluster supports active/standby access.
- Extended Distance Cluster disk reads may outperform Metrocluster in normal operations. On the other hand, Metrocluster data resynchronization and recovery performance are better than Extended Distance Cluster.

Continental Cluster

A **continental cluster** provides an alternative disaster tolerant solution in which distinct *clusters* can be separated by large distances, with wide area networking used between them. Continental cluster architecture is implemented via the Continentalclusters product, described fully in Chapter 2 of the *Designing Disaster Tolerant HA Clusters Using Metrocluster and Continentalclusters* user's guide. The design is implemented with distinct Serviceguard clusters that can be located in different geographic areas with the same or different subnet configuration. In this architecture, each cluster maintains its own quorum, so an arbitrator data center is not used for a continental cluster. A continental cluster can use any WAN connection via a TCP/IP protocol; however, due to data replication needs, high speed connections such as T1 or T3/E3 leased lines or switched lines may be required. See Figure 1-5.

NOTE A continental cluster can also be built using clusters that communicate over shorter distances using a conventional LAN.

Figure 1-5 Continental Cluster



Continentalclusters provides the flexibility to work with any data replication mechanism. It provides pre-integrated solutions that use HP StorageWorks Continuous Access XP, HP StorageWorks Continuous Access EVA, or EMC Symmetrix Remote Data Facility for data replication via the Metrocluster products.

The points to consider when configuring a continental cluster over a WAN are:

- Inter-cluster connections are TCP/IP based.
- The physical connection is one or more leased lines managed by a common carrier. Common carriers cannot guarantee the same reliability that a dedicated physical cable can. The distance can introduce a time lag for data replication, which creates an issue with data currency. This could increase the cost by requiring higher speed WAN connections to improve data replication performance and reduce latency.

- Operational issues, such as working with different personnel trained on different processes, and conducting failover rehearsals, are made more difficult the further apart the nodes are in the cluster.

Benefits of Continentalclusters

- Continentalclusters provides the ability to monitor a high availability cluster and fail over mission critical applications to another cluster if the monitored cluster should become unavailable.
- Continentalclusters supports mutual recovery, which allows for different critical applications to be run on each cluster, with each cluster configured to recover the mission critical applications of the other.
- You can virtually build data centers anywhere and still have the data centers provide disaster tolerance for each other. Since Continentalclusters uses multiple clusters, theoretically there is no limit to the distance between the clusters. The distance between the clusters is dictated by the required rate of data replication to the remote site, level of data currency, and the quality of networking links between the two data centers.
- In addition, inter-cluster communication can be implemented with either a WAN or LAN topology. LAN support is advantageous when you have data centers in close proximity to each other, but do not want the data centers configured into a single cluster. One example may be when you already have two Serviceguard clusters close to each other and, for business reasons, you cannot merge these two clusters into a single cluster. If you are concerned with one of the centers becoming unavailable, Continentalclusters can be added to provide disaster tolerance. Furthermore, Continentalclusters can be implemented with an existing Serviceguard cluster architecture while keeping both clusters running, and provide flexibility by supporting disaster recovery failover between two clusters that are on the same subnet or on different subnets.
- You can integrate Continentalclusters with any storage component of choice that is supported by Serviceguard. Continentalclusters provides a structure to work with any type of data replication mechanism. A set of guidelines for integrating other data replication schemes with Continentalclusters is included in the *Designing Disaster Tolerant HA Clusters Using Metrocluster and Continentalclusters* user's guide.

- Besides selecting your own storage and data replication solution, you can also take advantage of the following HP pre-integrated solutions:
  - Storage subsystems implemented by Metrocluster are also pre-integrated with Continentalclusters. Continentalclusters uses the same data replication integration module that Metrocluster implements to check for data status of the application package before package start up.
  - If Oracle DBMS is used and logical data replication is the preferred method, depending on the version, either Oracle 8i Standby or Oracle 9i Data Guard with log shipping is used to replicate the data between two data centers. HP provides a supported integration toolkit for Oracle 8i Standby DB in the Enterprise Cluster Management Toolkit (ECMT).
- RAC is supported by Continentalclusters by integrating it with SGeRAC. In this configuration, multiple nodes in a single cluster can simultaneously access the database (that is, nodes in one data center can access the database). If the site fails, the RAC instances can be recovered at the second site.
- RAC using Veritas **Cluster Volume Manager** (CVM) or Veritas **Cluster File System** (CFS) are supported by Continentalclusters by integrating it with SGeRAC. In this configuration, Oracle RAC instances are supported in the Continentalclusters environment for physical replication using HP StorageWorks Continuous Access XP, or EMC Symmetrix Remote Data Facility (SRDF) using HP SLVM or Veritas Cluster Volume Manager (CVM) or Cluster File Systems (CFS) from Symantec for volume management.

  For more information on configuring applications in CFS/CVM environments in Continentalclusters, refer to the "*Configuring Single Instance Applications in CFS/CVM Environments in Continentalclusters*" white paper on the high availability documentation web site at **`http://docs.hp.com`** -> High Availability -> Continentalcluster.
- Single instance applications using Veritas Cluster Volume Manager (CVM) or Veritas Cluster File System (CFS) are supported by Continentalclusters.
- Configuration of multiple recovery pairs is allowed. A recovery pair in a continental cluster consists of two Serviceguard clusters. One functions as a primary cluster and the other functions as recovery cluster for a specific application. In the multiple recovery pair

configuration, more than one primary cluster (where the primary packages are running) can be configured to share the same recovery cluster (where the recovery package is running).

- Continentalclusters maximum node support for Serviceguard/Serviceguard Extension for RAC depends upon storage management type (that is, LVM, SLVM, CVM, CFS).
- Failover for Continentalclusters is semi-automatic. If a data center fails, the administrator is advised, and is required to take action to bring the application up on the surviving cluster.
- Continentalclusters supports **Maintenance mode**; allows a recovery group in maintenance mode to be exempt from a recovery.
- Continentalclusters supports **Disaster Recovery (DR) Rehearsal**; detects configuration discrepancies at the recovery cluster and hence improves the "DR preparedness" of the recovery cluster.
- Continentalclusters supports **Data Replication Storage Failover Preview**; previews the preparation of the data replication environment for the storage, in a Metrocluster failover or in Continentalclusters recovery.

Support for Maintenance Mode in a Continentalclusters Environment

On the recovery cluster, recovery groups can be individually moved into maintenance mode. Continentalclusters does not allow recovery of those recovery groups that are in maintenance mode, such as `cmrecovercl`, or `cmrunpkg`/`cmmodpkg`, thus preventing the recovery package startup.

At initial configuration, by default, all recovery groups will be out of maintenance mode. However, when a recovery group is in maintenance mode the availability of the primary packages are not impacted that is, the primary package can be started up or can failover locally at the primary cluster.

NOTE Maintenance mode is an optional feature. To enable maintenance mode, configure a shared disk (non-replicated) with a file system on all recovery clusters and the Continentalclusters configuration file should be specified with the `CONTINENTAL_CLUSTER_STATE_DIR`.

A recovery group is moved into maintenance mode, by default, only if its primary package is running. However, if the site is unreachable or primary package is shutdown down, you can move a recovery group into maintenance mode by using the force option.

CAUTION Do not move a recovery group into maintenance mode, as in the case of the force option, if it is already recovered. This will prevent subsequent startups of the recovery package. Also, when used in DR Rehearsals, this will not prevent DR Rehearsal startups on the production data.

For more information on how to setup or use the maintenance mode feature, see the *Designing Disaster Tolerant HA Clusters Using Metrocluster and Continentalclusters* user's guide.

Support for Disaster Recovery Rehearsal

For a successful recovery in a Continentalclusters environment, it is critical that the configurations on all the systems, both primary and recovery cluster, are in sync.

The configuration, that is subject to change, after the initial setup may not be updated on all systems. Hence, this configuration inconsistency would prevent a recovery attempt on a specific node. For example, a recovery attempt could fail if the Metrocluster environment file changed on the primary cluster hosts and was not updated to the hosts at the recovery cluster.

The DR (Disaster Recovery) rehearsal feature "rehearses" the recovery without impacting the availability of the primary package. The DR rehearsal detects configuration discrepancies at the recovery cluster and hence improves the "DR preparedness" of the recovery cluster.

Continentalclusters, for DR Rehearsals, allows recovery groups to be configured with a special rehearsal package, which is specified as part of the recovery group definition. The DR Rehearsal starts the rehearsal package which has a package configuration that is similar to that of the recovery package and thereby verifying the recovery environment and procedure. The `cmrecovercl` option `{-r -g <recovery group>}` is used to start rehearsal for a recovery group on the recovery cluster.

NOTE DR Rehearsal startup is allowed only if the recovery group is in maintenance mode. This is a protection which ensures that while rehearsal is in progress, recovery is prevented. Since the recovery and rehearsal package have similar package configuration (that is, share resources), allowing both of them to start will result in resource collision and impact data integrity.

For more information on how to setup and run DR Rehearsal, see the *Designing Disaster Tolerant HA Clusters Using Metrocluster and Continentalclusters* user's guide.

Data Replication Storage Failover Preview

Data Replication Storage Failover Preview allows you to preview the preparation for the storage of the data replication environment in a Metrocluster failover or Continentalclusters recovery. This is done with the `cmdrprev` command, which also verifies the data replication environment that may cause a Metrocluster failover or Continentalclusters recovery to fail.

For more information on the use of `cmdrprev`, see the *Designing Disaster Tolerant HA Clusters Using Metrocluster and Continentalclusters* user's guide.

Continental Cluster With Cascading Failover

A continental cluster with **cascading failover** uses three main data centers distributed between a metropolitan cluster, which serves as a primary cluster, and a standard cluster, which serves as a recovery cluster.

Cascading failover means that applications are configured to fail over from one data center to another in the primary cluster and then to a third (recovery) cluster if the entire primary cluster fails. Data replication also follows the cascading model. Data is replicated from the primary disk array to the secondary disk array in the Metrocluster, then replicated to the third disk array in the Serviceguard recovery cluster.

For more information on Cascading Failover configuration, maintenance, and recovery procedures, refer to the "*Cascading Failover in a Continental Cluster*" white paper on the high availability documentation web site at **`http://docs.hp.com`** -> High Availability -> Continentalclusters.

Cascading Failover Using Metrocluster

This configuration uses three data replication groups, two of which are part of the metropolitan cluster and the other attached to the recovery cluster. The data centers are distributed as follows:

- Primary—on the site that holds the primary copy of the data, located in the primary cluster.
- Secondary—on the site that holds a remote mirror copy of the data, located in the primary cluster.
- Arbitrator or Quorum Server—a third location that contains the arbitrator nodes, or quorum server located in the primary cluster.
- Recovery—on a site that holds a remote mirror copy of the data, located in the recovery cluster.

Figure 1-6 illustrates data centers, clusters, and nodes in a cascading failover configuration, and shows at a high level how the data replication is connected. The primary cluster consists of two storage devices: a source device (connected to the primary site and labeled as device A) and a destination device (connected to the secondary site and labeled as device B). Data is replicated via storage data replication facilities (for example, Continuous Access) continuously from source to destination.

On site 2, a local mirror is associated with the destination devices (labeled as device B'). The mirror technology is storage specific (for example, Business Copy). This local mirror also acts as a source device for recovery during rolling disasters.

A **rolling disaster** is defined as a disaster that occurs before the cluster is able to recover from a non-disastrous failure. An example is a data replication link that fails, then, as it is being restored and data is being resynchronized, a disaster causes an entire data center to fail.

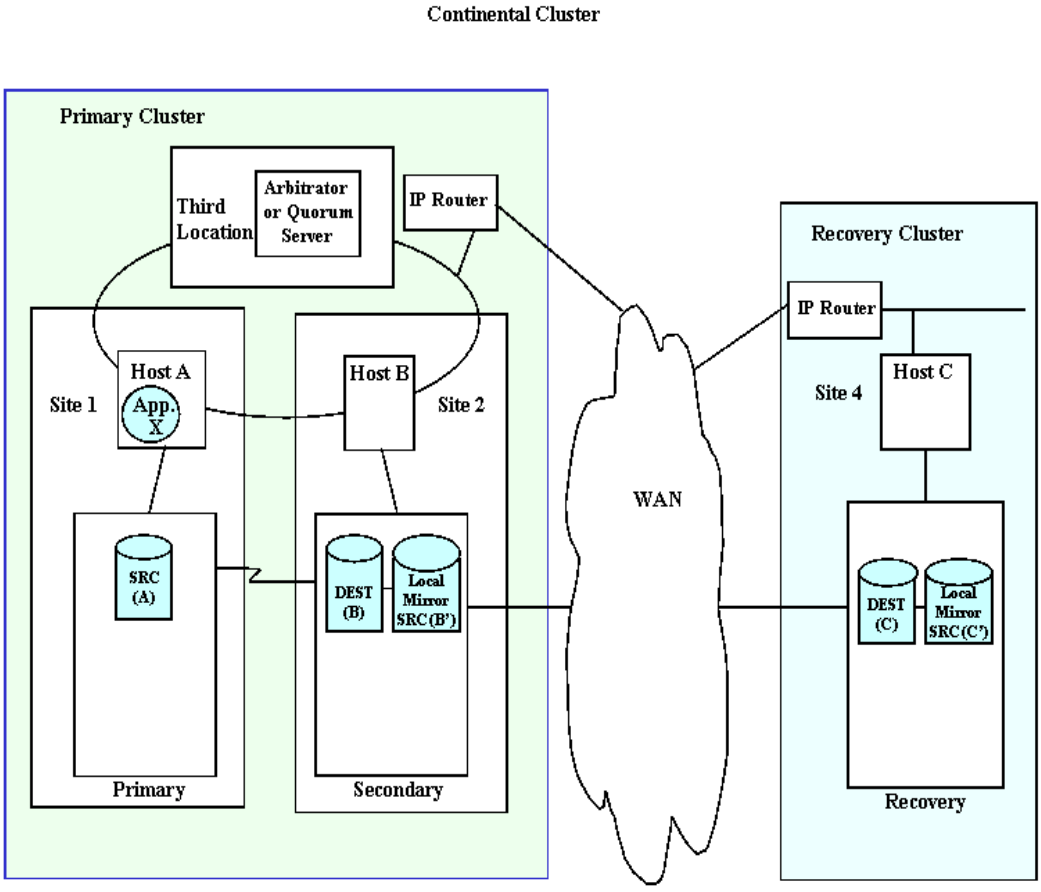
In the recovery cluster, on site 4, the destination device (labeled as device C) is connected to the node in the cluster. Data is periodically replicated to the destination devices via storage data replication technology. A local mirror of the destination device is required on site 4 for cases of rolling disasters (labeled as device C').

Currently, HP StorageWorks XP Continuous Access and EMC Symmetrix SRDF technologies are supported for the multi-site disaster tolerant solution.

Refer to the *Designing Disaster Tolerant HA Clusters Using Metrocluster and Continentalclusters* user's guide for details on setting up data replication for this type of cluster.

The next section provides an overview of a three data center solution, which utilizes both Metrocluster Continuous Access XP and Continentalclusters environments.

Figure 1-6 Cascading Failover Data Center Distribution Using Metrocluster

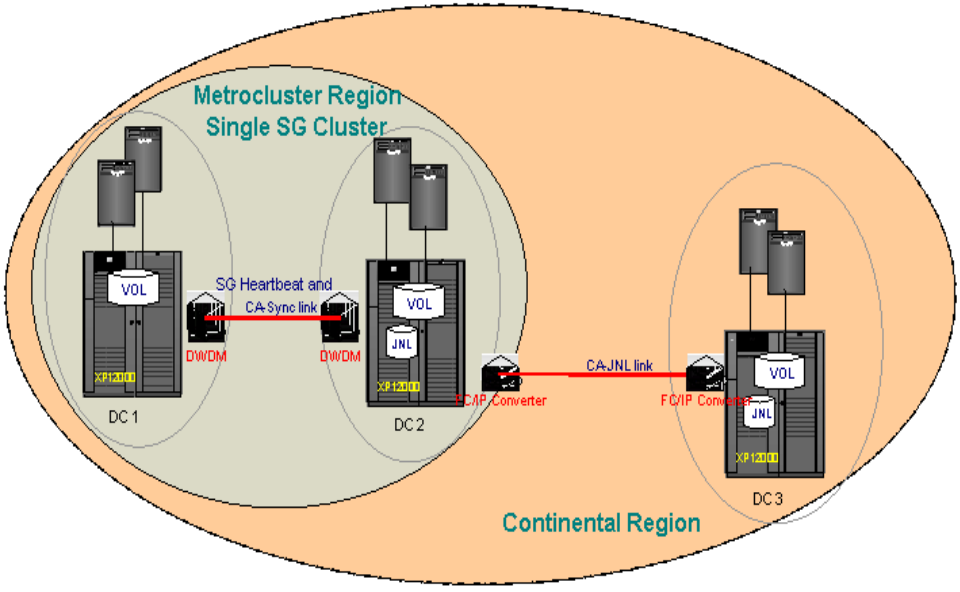


Three Data Center Architecture

A Three Data Center solution integrates Serviceguard, Metrocluster Continuous Access XP, Continentalclusters and HP StorageWorks XP 3DC Data Replication Architecture. This configuration protects against local and wide-area disasters by using both synchronous replication (for data consistency) and Continuous Access journaling (for long-distance replication).

A Three Data Center configuration consists of two Serviceguard clusters. The first cluster, which is a Metrocluster, has two data centers that make up the Primary data center (DC1) and Secondary data center (DC2). The second cluster, typically located at a long distance from the Metrocluster sites, is the Third Data Center (DC3); it is configured as a recovery cluster. These two clusters are configured as a Continental cluster, as shown in Figure 1-7.

Figure 1-7 Three Data Center Solution Overview



HP XP StorageWorks in a Three Data Center Architecture

HP XP StorageWorks Three Data Center architecture enables data to be replicated over three data centers concurrently using a combination of Continuous Access Synchronous and Continuous Access Journaling data replication.

In a XP 3DC design there are two available configurations; *Multi-Target* and *Multi-Hop*. The XP 3DC configuration can switch between the Multi-Target and Multi-Hop configurations at any time during a normal operation. These configurations may be implemented with either two or three Continuous Access links between the data centers.

When there are two Continuous Access links, one link is a Continuous Access Sync and the other is a Continuous Access Journal data replication link. As both supported configurations use two Continuous Access links, they are also referred to as *Multi-Hop-Bi-Link* and *Multi-Target-Bi-Link*.

Whether the configuration is multi-hop or multi-target is determined by two factors: where data enters the system (that is, where the application is running) and in what direction the data flows between the XP arrays. In an XP 3DC Multi-Target Bi-Link configuration the data enters the system on a specific XP array and is replicated into multiple directions. In an XP 3DC Multi-Hop Bi-Link configuration the data enters the system on one XP array, is replicated synchronously to the next XP array, and from there is replicated to the last XP array.

A Three Data Center configuration uses HP StorageWorks 3DC Data Replication Architecture in order to replicate data over three data centers, which provides complete data currency and protects against both local and wide-area disasters. Also, a Three Data Center configuration concurrently supports short-distance Continuous Access synchronous replication within the Metrocluster, and long-distance Continuous Access journal replication between the Metrocluster and recovery cluster.

The Three Data Center Architecture is described fully in Chapter 6, *Designing a Disaster Tolerant Solution Using the Three Data Center Architecture* of the *Designing Disaster Tolerant HA Clusters Using Metrocluster and Continentalclusters* user's guide on the high availability documentation web site at **`http://docs.hp.com`** -> High Availability -> Metrocluster or Continentalcluster.

Comparison of Disaster Tolerant Solutions

Table 1-1 summarizes and compares the disaster tolerant solutions that are currently available:

Table 1-1 Comparison of Disaster Tolerant Cluster Solutions

Attributes	Extended Distance Cluster	Extended Distance Cluster for RAC	Metrocluster	Continentalclusters
Key Benefit	Excellent in "normal" operations, and partial failure. Since all hosts have access to both disks, in a failure where the node is running and the application is up, but the disk becomes unavailable, no failover occurs. The node will access the remote disk to continue processing.	Excellent in "normal" operations, and partial failure. Active/active configuration provides maximum data through put and reduces the need for failover (since both data centers are active, the application is already up on the 2nd site).	Two significant benefits: • Provides maximum data protection. State of the data is determined before application is started. If necessary, data resynchroniza -tion is performed before application is brought up. • Better performance than Extended Distance Cluster for resync, as replication is done by storage subsystem (no impact to host).	Increased data protection by supporting unlimited distance between data centers (protects against such disasters as those caused by earthquakes or violent attacks, where an entire area can be disrupted).

Table 1-1 Comparison of Disaster Tolerant Cluster Solutions (Continued)

Attributes	Extended Distance Cluster	Extended Distance Cluster for RAC	Metrocluster	Continentalclusters
Key Limitation	No ability to check the state of the data before starting up the application. If the volume group (vg) can be activated, the application will be started. If mirrors are split or PV links are down, as long as the vg can be activated, the application will be started. Data resynchroniza-tion can have a big impact on system performance, as this is a host-based solution.	SLVM configuration is limited to 2 nodes for distances of up to 100km*. CVM or CFS, which are available with Serviceguard Storage Management Suite Bundles, configuration supports up to 8 nodes. However, 8-node configuration is limited to a distance of 10km*. Data resynchronization can have a big impact on system performance as this is a host-based solution.	Specialized storage required. Currently, XP with continuous access, EVA with continuous access, and EMC's Symmetrix with SRDF are supported.	No automatic failover between clusters.

Table 1-1 Comparison of Disaster Tolerant Cluster Solutions (Continued)

Attributes	Extended Distance Cluster	Extended Distance Cluster for RAC	Metrocluster	Continentalclusters
Maximum Distance	* 100 Kilometers	* 100km (maximum is 2 nodes, with either SLVM or CVM) * 10km (maximum is 2 nodes with SLVM and 8 nodes with CVM and CFS)	Shortest of the distances between: • Cluster network latency (not to exceed 200ms). • Data Replication Max Distance. • DWDM provider max distance.	No distance restrictions.
Data Replication mechanism	Host-based, via MirrorDisk/UX or (Veritas) VxVM. Replication can affect performance (writes are synchronous). Re-syncs can impact performance (full re-sync is required in many scenarios that have multiple failures.)	Host-based, via MirrorDisk/UX or (Veritas) CVM and CFS Replication can impact performance (writes are synchronous). Re-syncs can impact performance (full re-sync is required in many scenarios that have multiple failures).	Array-based, via CAXP or CAEVA or EMC SRDF. Replication and resynchronization performed by the storage subsystem, so the host does not experience a performance hit. Incremental re-syncs are done, based on bitmap, minimizing the need for full re-syncs.	You have a choice of either selecting your own SG-supported storage and data replication mechanism, or implementing one of HP's pre-integrated solutions (including CA XP, CA EVA, and EMC SRDF for array-based, or Oracle 8i Standby for host based.) Also, you may choose Oracle 9i Data Guard as a host-based solution. Contributed (that is, unsupported) integration templates for Oracle 9i.

Table 1-1 Comparison of Disaster Tolerant Cluster Solutions (Continued)

Attributes	Extended Distance Cluster	Extended Distance Cluster for RAC	Metrocluster	Continentalclusters
Application Failover	Automatic (no manual intervention required).	Instance is already running at the 2nd site.	Automatic (no manual intervention required).	Semi-automatic (user must "push the button" to initiate recovery). *Disaster Recovery (DR) Rehearsal* provides a method to identify and fix configuration inconsistency at the recovery cluster. See *Support for Maintenance Mode in a Continentalclusters Environment*. *Data Replication Storage Failover Preview*
Access Mode	Active/Standby	Active/Active	Active/Standby	Active/Standby
Client Transpar-ency	Client detects the lost connection. You must reconnect once the application is recovered at 2nd site.	Client may already have a standby connection to remote site.	Client detects the lost connection. You must reconnect once the application is recovered at 2nd site.	You must reconnect once the application is recovered at 2nd site.

Table 1-1 Comparison of Disaster Tolerant Cluster Solutions (Continued)

Attributes	Extended Distance Cluster	Extended Distance Cluster for RAC	Metrocluster	Continentalclusters
Maximum Cluster Size Allowed	2 to 16 nodes (up to 4 when using dual lock disks).	* 2, 4, 6, or 8 nodes with SLVM or CVM with a maximum distance of 100km. * 2, 4, 6, or 8 nodes to 8 nodes with CVM with a maximum distance of 10km.	3 to 16 nodes	Depends storage management type (that is, LVM, SLVM, CVM, CFS) based on what is being used for Serviceguard/SGeRAC.
Storage	Identical storage is not required (replication is host-based with either MirrorDisk/UX or VxVM mirroring).	Identical storage is not required (replication is host-based with either *MirrorDisk/UX or CVM Mirroring).	Identical Storage is required.	Identical storage is required if storage-based mirroring is used. Identical storage is not required for other data replication implementations.
Data Replication Link	Dark Fiber	Dark Fiber	Dark Fiber Continuous Access over IP Continuous Access over ATM	WAN LAN Dark Fiber (pre-integrated solution) Continuous Access over IP (pre-integrated solution) Continuous Access over ATM (pre-integrated solution)

Table 1-1 Comparison of Disaster Tolerant Cluster Solutions (Continued)

Attributes	Extended Distance Cluster	Extended Distance Cluster for RAC	Metrocluster	Continentalclusters
Cluster Network	Single IP subnet	Single IP subnet	Single IP subnet	Two configurations: Single IP subnet for both clusters (LAN connection between clusters) Two IP subnets – one per cluster (WAN connection between clusters)
DTS Software/ Licenses Required	SG (no other clustering SW is required).	SG + SGeRAC	SG + Metrocluster Continuous Access XP or Metrocluster Continuous Access EVA or Metrocluster EMC SRDF	SG + Continentalclusters + (Metrocluster Continuous Access XP or Metrocluster Continuous Access EVA or Metrocluster EMC SRDF or Enterprise Cluster Master Toolkit) or Customer-selected data replication subsystem CC with RAC: SG + SGeRAC + CVM/CFS + Continentalclusters

NOTE * Refer to Table 1-2 below for the maximum supported distances between data centers for Extended Distance Cluster configurations.

For more detailed configuration information on Extended Distance Cluster, refer to the *HP Configuration Guide* (available through your HP representative).

For the most up-to-date support and compatibility information see the *SGeRAC for SLVM, CVM & CFS Matrix and Serviceguard Compatibility and Feature Matrix* on **`http://docs.hp.com`** -> High Availability -> Serviceguard Extension for Real Application Cluster (ServiceGuard OPS Edition) -> Support Matrixes.

Table 1-2 Supported Distances Extended Distance Cluster Configurations

Cluster type/Volume Manager	Distances up to 10 kilometers	Distances up to 100 kilometers
Serviceguard with LVM and MirrorDisk/UX	Supported for clusters with up to 16 nodes with Serviceguard A.11.16 or greater on HP-UX 11i v1, 11i v2 or 11i v3	Supported for clusters with up to 16 nodes with Serviceguard A.11.16 or greater on HP-UX 11i v1, 11i v2 or 11i v3
Serviceguard with VxVM mirroring	Supported for clusters with up to 16 nodes with Serviceguard A.11.16 or greater on HP-UX 11i v1 or 11i v2	Supported for clusters with up to 16 nodes with Serviceguard A.11.16 or greater on HP-UX 11i v1 or 11i v2
SGeRAC with SLVM and MirrorDisk/UX	Supported for clusters with 2 nodes with SGeRAC A.11.16 or greater on HP-UX 11i v1, 11i v2 or 11i v3	Supported for clusters with 2 nodes with SGeRAC A.11.16 or greater on HP-UX 11i v1, 11i v2 and 11i v3
SGeRAC with CVM 3.5 mirroring	Supported for clusters with 2, 4, 6, or 8 nodes with SGeRAC A.11.16 or greater on HP-UX 11i v1 or 11i v2.	Supported for clusters with 2 nodes with SGeRAC A.11.16 or greater on HP-UX 11i v1 or 11i v2.

Table 1-2 Supported Distances Extended Distance Cluster Configurations

Cluster type/Volume Manager	Distances up to 10 kilometers	Distances up to 100 kilometers
Serviceguard A.11.17 with CVM 4.1 or CFS 4.1 mirroring	Supported for clusters with 2, 4, 6, or 8 nodes with Serviceguard A.11.17 on 11i v2	Supported for clusters with 2, 4, 6 or 8 nodes with Serviceguard A.11.17 on HP-UX 11i v2
SGeRAC A.11.17 with CVM 4.1 or CFS 4.1 mirroring	Supported for clusters with 2, 4, 6, or 8 nodes with Serviceguard A.11.17 on 11i v2. Supported with Oracle RAC 9.2 or 10gR2	Supported for clusters with 2 nodes with Serviceguard A.11.17 on HP-UX 11i v2. Supported with Oracle RAC 9.2 or 10gR2

Disaster Tolerant Architecture Guidelines

Disaster tolerant architectures represent a shift away from the massive central data centers and towards more distributed data processing facilities. While each architecture will be different to suit specific availability needs, there are a few basic guidelines for designing a disaster tolerant architecture so that it protects against the loss of an entire data center:

- Protecting nodes through geographic dispersion
- Protecting data through replication
- Using alternative power sources
- Creating highly available networks

These guidelines are in addition to the standard high-availability guidelines of redundant components such as PV links, network cards, power supplies, and disks.

Protecting Nodes through Geographic Dispersion

Redundant nodes in a disaster tolerant architecture must be geographically dispersed. If they are in the same data center, it is not a disaster tolerant architecture. Figure 1-2 on page 19 shows a cluster architecture with nodes in two data centers: A and B. If all nodes in data center A fail, applications can fail over to the nodes in data center B and continue to provide clients with service.

Depending on the type of disaster you are protecting against and on the available technology, the nodes can be as close as another room in the same building, or as far away as another city. The minimum recommended dispersion is a single building with redundant nodes in different data centers using different power sources. Specific architectures based on geographic dispersion are discussed in “Understanding Types of Disaster Tolerant Clusters” on page 20.

Protecting Data through Replication

The most significant losses during a disaster are the loss of access to data, and the loss of data itself. You protect against this loss through data replication, that is, creating extra copies of the data. Data replication should:

- Ensure **data consistency** by replicating data in a logical order so that it is immediately usable or recoverable. Inconsistent data is unusable and is not recoverable for processing. Consistent data may or may not be current.
- Ensure **data currency** by replicating data quickly so that a replica of the data can be recovered to include all committed disk writes that were applied to the local disks.
- Ensure **data recoverability** so that there is some action that can be taken to make the data consistent, such as applying logs or rolling a database.
- Minimize **data loss** by configuring data replication to address consistency, currency, and recoverability.

Different data replication methods have different advantages with regards to data consistency and currency. Your choice of which data replication methods to use will depend on what type of disaster tolerant architecture you require.

Off-line Data Replication

Off-line data replication is the method most commonly used today. It involves two or more data centers that store their data on tape and either send it to each other (via an express service, if need dictates) or store it off-line in a vault. If a disaster occurs at one site, the off-line copy of data is used to synchronize data and a remote site functions in place of the failed site.

Because data is replicated using physical off-line backup, data consistency is fairly high, barring human error or an untested corrupt backup. However, data currency is compromised by the time delay in sending the tape backup to a remote site.

Off-line data replication is fine for many applications for which recovery time is not an issue critical to the business. Although data might be replicated weekly or even daily, recovery could take from a day to a week

depending on the volume of data. Some applications, depending on the role they play in the business, may need to have a faster recovery time, within hours or even minutes.

On-line Data Replication

On-line data replication is a method of copying data from one site to another across a link. It is used when very short recovery time, from minutes to hours, is required. To be able to recover use of a system in a short time, the data at the alternate site must be replicated in real time on all disks.

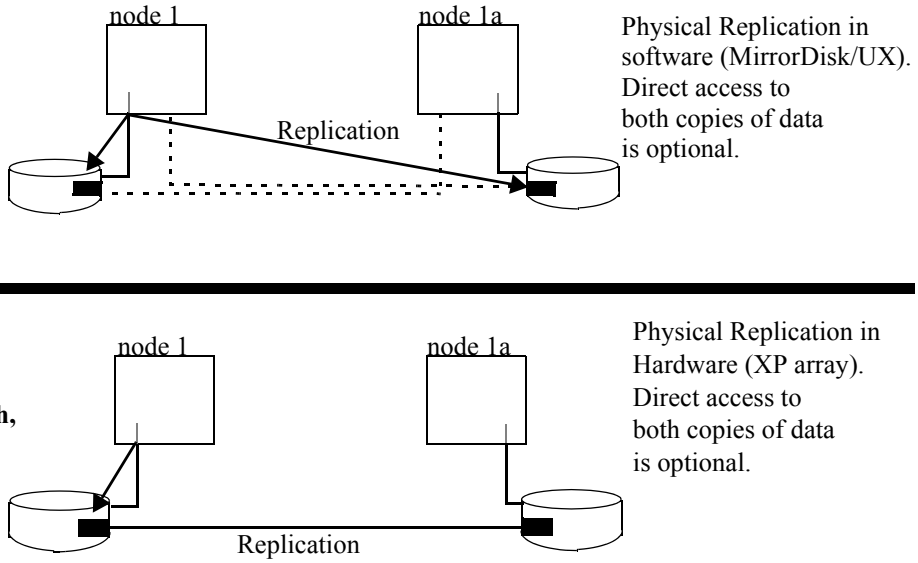
Data can be replicated either synchronously or asynchronously. **Synchronous replication** requires one disk write to be completed and replicated before another disk write can begin. This method improves the chances of keeping data consistent and current during replication. However, it greatly reduces replication capacity and performance, as well as system response time. **Asynchronous replication** does not require the primary site to wait for one disk write to be replicated before beginning another. This can be an issue with data currency, depending on the volume of transactions. An application that has a very large volume of transactions can get hours or days behind in replication using asynchronous replication. If the application fails over to the remote site, it would start up with data that is not current.

Currently the two ways of replicating data on-line are physical data replication and logical data replication. Either of these can be configured to use synchronous or asynchronous writes.

Physical Data Replication Each physical write to disk is replicated on another disk at another site. Because the replication is a physical write to disk, it is not application dependent. This allows each node to run different applications under normal circumstances. Then, if a disaster occurs, an alternate node can take ownership of applications and data, provided the replicated data is current and consistent.

As shown in Figure 1-8, physical replication can be done in software or hardware.

Figure 1-8 Physical Data Replication



The distance between nodes is limited by the link type (SCSI, ESCON, FC) and the intermediate devices (FC switch, DWDM) on the link path

MirrorDisk/UX is an example of physical replication done in the software; a disk I/O is written to each array connected to the node, requiring the node to make multiple disk I/Os. Continuous Access XP on the HP StorageWorks E Disk Array XP series is an example of physical replication in hardware; a single disk I/O is replicated across the Continuous Access link to a second XP disk array.

Advantages of physical replication in hardware are:

- There is little or no lag time writing to the replica. This means that the data remains very current.
- Replication consumes no additional CPU.
- The hardware deals with resynchronization if the link or disk fails. And resynchronization is independent of CPU failure; if the CPU fails and the disk remains up, the disk knows it does not have to be resynchronized.
- Data can be copied in both directions, so that if the primary fails and the replica takes over, data can be copied back to the primary when it comes back up.

Disadvantages of physical replication in hardware are:

- The logical order of data writes is not always maintained in synchronous replication. When a replication link goes down and transactions continue at the primary site, writes to the primary disk are queued in a bit-map. When the link is restored, if there has been more than one write to the primary disk, then there is no way to determine the original order of transactions until the resynchronization has completed successfully. This increases the risk of data inconsistency.

NOTE

Configuring the disk so that it does not allow a subsequent disk write until the current disk write is copied to the replica (synchronous writes) can limit this risk as long as the link remains up. Synchronous writes impact the capacity and performance of the data replication technology.

Also, because the replicated data is a write to a physical disk block, database corruption and human errors, such as the accidental removal of a database table, are replicated at the remote site.

- Redundant disk hardware and cabling are required. This at a minimum doubles data storage costs, because the technology is in the disk itself and requires specialized hardware.
- For architectures using dedicated cables, the distance between the sites is limited by the cable interconnect technology. Different technologies support different distances and provide different “data through” performance.
- For architectures using common carriers, the costs can vary dramatically, and the connection can be less reliable, depending on the Service Level Agreement.

Advantages of physical replication in software are:

- There is little or no time lag between the initial and replicated disk I/O, so data remains very current.
- The solution is independent of disk technology, so you can use any supported disk technology.
- Data copies are peers, so there is no issue with reconfiguring a replica to function as a primary disk after failover.

- Because there are multiple read devices, that is, the node has access to both copies of data, there may be improvements in read performance.
- Writes are synchronous unless the link or disk is down.

Disadvantages of physical replication in software are:

- As with physical replication in the hardware, the logical order of data writes is not maintained. When the link is restored, if there has been more than one write to the primary disk, there is no way to determine the original order of transactions until the resynchronization has completed successfully.

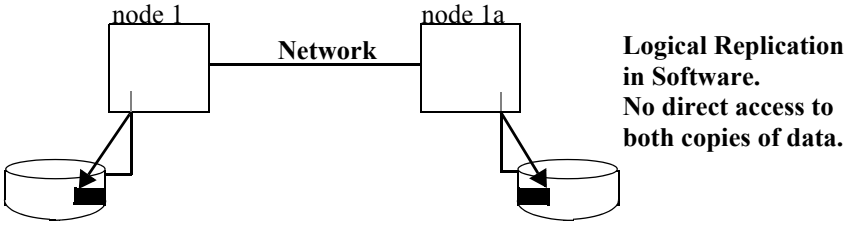
NOTE Configuring the software so that a write to disk must be replicated on the remote disk before a subsequent write is allowed can limit the risk of data inconsistency while the link is up.

- Additional hardware is required for the cluster.
- Distance between sites is limited by the physical disk link capabilities.
- Performance is affected by many factors: CPU overhead for mirroring, double I/Os, degraded write performance, and CPU time for resynchronization. In addition, CPU failure may cause a resynchronization even if it is not needed, further affecting system performance.

Logical Data Replication Logical data replication is a method of replicating data by repeating the sequence of transactions at the remote site. Logical replication often must be done at both the file system level, and the database level in order to replicate all of the data associated with an application. Most database vendors have one or more database replication products. An example is the Oracle Standby Database.

Logical replication can be configured to use synchronous or asynchronous writes. Transaction processing monitors (TPMs) can also perform logical replication.

Figure 1-9 Logical Data Replication



Advantages of using logical replication are:

- The distance between nodes is limited only by the networking technology.
- There is no additional hardware needed to do logical replication, unless you choose to boost CPU power and network bandwidth.
- Logical replication can be implemented to reduce risk of duplicating human error. For example, if a database administrator erroneously removes a table from the database, a physical replication method will duplicate that error at the remote site as a raw write to disk. A logical replication method can be implemented to delay applying the data at a remote site, so such errors would not be replicated at the remote site. This also means that administrative tasks, such as adding or removing database tables, has to be repeated at each site.
- With database replication you can roll transactions forward or backward to achieve the level of currency desired on the replica, although this functionality is not available with file system replication.

Disadvantages of logical replication are:

- It uses significant CPU overhead because transactions are often replicated more than once and logged to ensure data consistency, and all but the most simple database transactions take significant CPU. It also uses network bandwidth, whereas most physical replication methods use a separate data replication link. As a result, there may be a significant lag in replicating transactions at the remote site, which affects data currency.

- If the primary database fails and is corrupt, which results in the replica taking over, then the process for restoring the primary database so that it can be used as the replica is complex. This often involves recreating the database and doing a database dump from the replica.
- Applications often have to be modified to work in an environment that uses a logical replication database. Logic errors in applications or in the RDBMS code itself that cause database corruption will be replicated to remote sites. This is also an issue with physical replication.
- Most logical replication methods do not support personality swapping, which is the ability after a failure to allow the secondary site to become the primary and the original primary to become the new secondary site. This capability can provide increased up time.

Ideal Data Replication The ideal disaster tolerant architecture, if budgets allow, is the following combination:

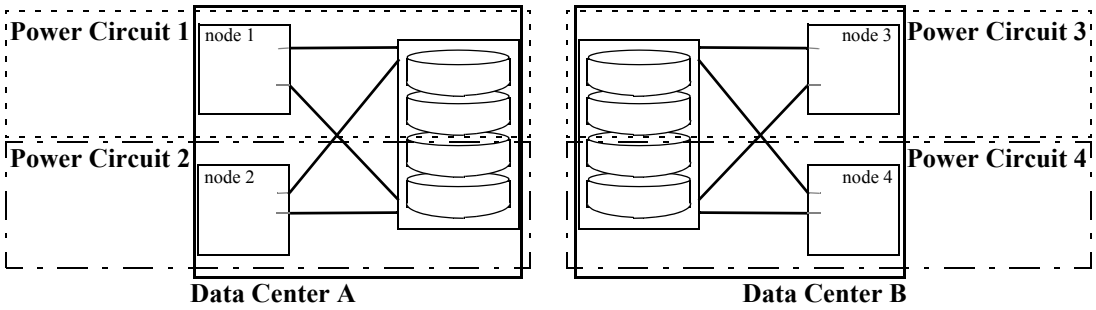
- For performance and data currency—physical data replication.
- For data consistency—either a second physical data replication as a point-in-time snapshot or logical data replication, which would only be used in the cases where the primary physical replica was corrupt.

Using Alternative Power Sources

In a high-availability cluster, redundancy is applied to cluster components, such as PV links, redundant network cards, power supplies, and disks. In disaster tolerant architectures another level of protection is required for these redundancies.

Each data center that houses part of a disaster tolerant cluster should be supplied with power from a different circuit. In addition to a standard UPS (uninterrupted power supply), each node in a disaster tolerant cluster should be on a separate power circuit; see Figure 1-10.

Figure 1-10 Alternative Power Sources



Housing remote nodes in another building often implies they are powered by a different circuit, so it is especially important to make sure all nodes are powered from a different source if the disaster tolerant cluster is located in two data centers in the same building. Some disaster tolerant designs go as far as making sure that their redundant power source is supplied by a different power substation on the grid. This adds protection against large-scale power failures, such as brown-outs, sabotage, or electrical storms.

Creating Highly Available Networking

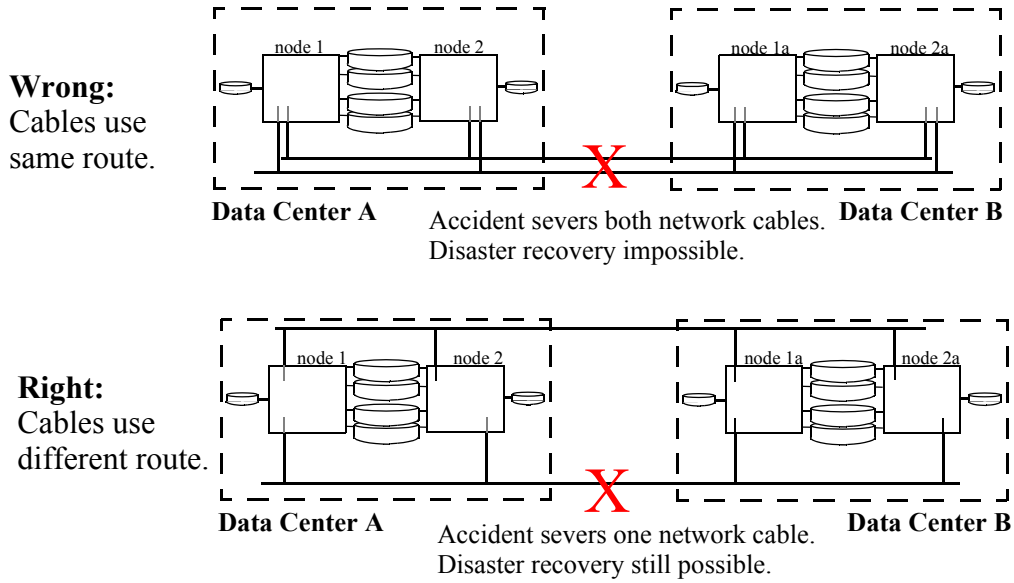
Standard high-availability guidelines require redundant networks. Redundant networks may be highly available, but they are not disaster tolerant if a single accident can interrupt both network connections. For example, if you use the same trench to lay cables for both networks, you do not have a disaster tolerant architecture because a single accident, such as a backhoe digging in the wrong place, can sever both cables at once, making automated failover during a disaster impossible.

In a disaster tolerant architecture, the reliability of the network is paramount. To reduce the likelihood of a single accident causing both networks to fail, redundant network cables should be installed so that they use physically different routes for each network as indicated in Figure 1-11. How you route cables will depend on the networking technology you use. Specific guidelines for some network technologies are listed here.

Disaster Tolerant Local Area Networking

The configurations described in this section are for FDDI and Ethernet based Local Area Networks.

Figure 1-11 Reliability of the Network is Paramount

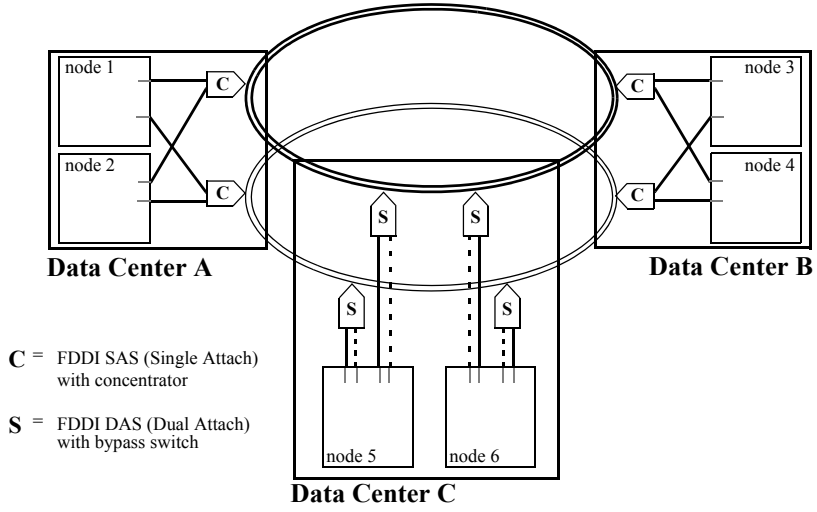


If you use FDDI networking, you may want to use one of these configurations, or a combination of the two:

- Each node has two SAS (Single Attach Station) host adapters with redundant concentrators and a different dual FDDI ring connected to each adapter. As per disaster tolerant architecture guidelines, each ring must use a physically different route.
- Each node has two DAS (Dual Attach Station) host adapters, with each adapter connected to a different FDDI bypass switch. The optical bypass switch is used so that a node failure does not bring the entire FDDI ring down. Each switch is connected to a different FDDI ring, and the rings are installed along physically different routes.

These FDDI options are shown in Figure 1-12.

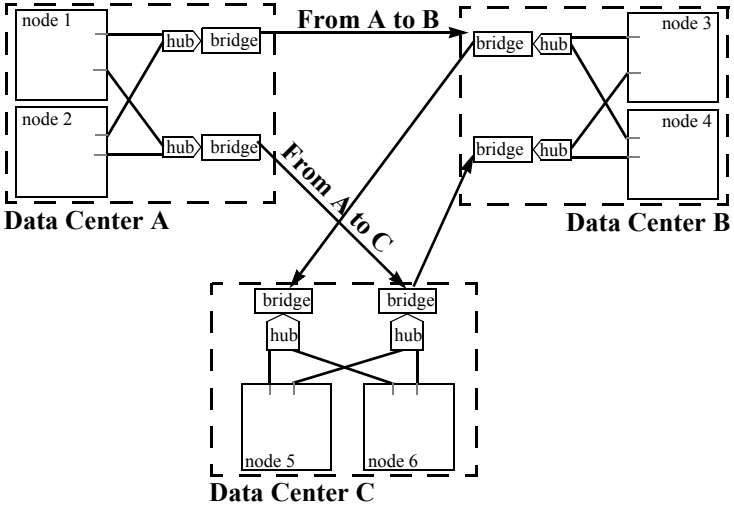
Figure 1-12 Highly Available FDDI Network: Two Options



Ethernet networks can also be used to connect nodes in a disaster tolerant architecture within the following guidelines:

- Each node is connected to redundant hubs and bridges using two Ethernet host adapters. Bridges, repeaters, or other components that convert from copper to fiber cable may be used to span longer distances.
- Redundant Ethernet links must be routed in the opposite directions, otherwise a data center failure can cause a failure of both networks. Figure 1-13 shows one Ethernet link routed from data center A, to B, to C. The other is routed from A, to C, to B. If both links were routed from A, to C, to B, the failure of data center C would make it impossible for clients to connect to data center B in the case of failover.

Figure 1-13 Routing Highly Available Ethernet Connections in Opposite Directions



Disaster Tolerant Wide Area Networking

Disaster tolerant networking for continental clusters is directly tied to the data replication method. In addition to the redundant lines connecting the remote nodes, you also need to consider what bandwidth you need to support the data replication method you have chosen. A continental cluster that handles a high number of transactions per minute will not only require a highly available network, but also one with a large amount of bandwidth.

This is a brief discussion of things to consider when choosing the network configuration for your continental cluster. Details on WAN choices and configurations can be found in a white paper available from **`http://docs.hp.com`** -> High Availability.

- Bandwidth affects the rate of data replication, and therefore the currency of the data should there be the need to switch control to another site. The greater the number of transactions you process, the more bandwidth you will need. The following connection types offer differing amounts of bandwidth:
  - — T1 and T3: low end
  - — ISDN and DSL: medium bandwidth

— ATM: high end

- Reliability affects whether or not data replication happens, and therefore the consistency of the data should you need to fail over to the recovery cluster. Redundant leased lines should be used, and should be from two different common carriers, if possible.
- Cost influences both bandwidth and reliability. Higher bandwidth and dual leased lines cost more. It is best to address data consistency issues first by installing redundant lines, then weigh the price of data currency and select the line speed accordingly.

Disaster Tolerant Cluster Limitations

Disaster tolerant clusters have limitations, some of which can be mitigated by good planning. Some examples of MPOF that may not be covered by disaster tolerant configurations:

- Failure of all networks among all data centers — This can be mitigated by using a different route for all network cables.
- Loss of power in more than one data center — This can be mitigated by making sure data centers are on different power circuits, and redundant power supplies are on different circuits. If power outages are frequent in your area, and down time is expensive, you may want to invest in a backup generator.
- Loss of all copies of the on-line data — This can be mitigated by replicating data off-line (frequent backups). It can also be mitigated by taking snapshots of consistent data and storing it on-line; Business Copy XP and EMC Symmetrix BCV (Business Consistency Volumes) provide this functionality and the additional benefit of quick recovery should anything happen to both copies of on-line data.
- A **rolling disaster** is a disaster that occurs before the cluster is able to recover from a non-disastrous failure. An example is a data replication link that fails, then, as it is being restored and data is being resynchronized, a disaster causes an entire data center to fail. The effects of rolling disasters can be mitigated by ensuring that a copy of the data is stored either off-line or on a separate disk that can quickly be mounted. The trade-off is a lack of currency for the data in the off-line copy.

Managing a Disaster Tolerant Environment

In addition to the changes in hardware and software to create a disaster tolerant architecture, there are also changes in the way you manage the environment. Configuration of a disaster tolerant architecture needs to be carefully planned, implemented and maintained. There are additional resources needed, and additional decisions to make concerning the maintenance of a disaster tolerant architecture:

- *Manage it in-house, or hire a service?*

  Hiring a service can remove the burden of maintaining the capital equipment needed to recover from a disaster. Most disaster recovery services provide their own off-site equipment, which reduces maintenance costs. Often the disaster recovery site and equipment are shared by many companies, further reducing cost.

  Managing disaster recovery in-house gives complete control over the type of redundant equipment used and the methods used to recover from disaster, giving you complete control over all means of recovery.

- *Implement automated or manual recovery?*

  Manual recovery costs less to implement and gives more flexibility in making decisions while recovering from a disaster. Evaluating the data and making decisions can add to recovery time, but it is justified in some situations, for example if applications compete for resources following a disaster and one of them has to be halted.

  Automated recovery reduces the amount of time and in most cases eliminates human intervention needed to recover from a disaster. You may want to automate recovery for any number of reasons:

  — Automated recovery is usually faster.

  — Staff may not be available for manual recovery, as is the case with “lights-out” data centers.

  — Reduction in human intervention is also a reduction in human error. Disasters don’t happen often, so lack of practice and the stressfulness of the situation may increase the potential for human error.

  — Automated recovery procedures and processes can be transparent to the clients.

Even if recovery is automated, you may choose to, or need to recover from some types of disasters with manual recovery. A **rolling disaster**, which is a disaster that happens before the cluster has recovered from a previous disaster, is an example of when you may want to manually switch over. If the data link failed, as it was coming up and resynchronizing data, and the data center failed, you would want human intervention to make judgment calls on which site had the most current and consistent data before failing over.

- *Who manages the nodes in the cluster and how are they trained?*

  Putting a disaster tolerant architecture in place without planning for the people aspects is a waste of money. Training and documentation are more complex because the cluster is in multiple data centers.

  Each data center often has its own operations staff with their own processes and ways of working. These operations people will now be required to communicate with each other and coordinate maintenance and failover rehearsals, as well as working together to recover from an actual disaster. If the remote nodes are placed in a "lights-out" data center, the operations staff may want to put additional processes or monitoring software in place to maintain the nodes in the remote location.

  Rehearsals of failover scenarios are important to keep prepared. A written plan should outline rehearsal of what to do in cases of disaster with a minimum recommended rehearsal schedule of once every 6 months, ideally once every 3 months.

- *How is the cluster maintained?*

  Planned downtime and maintenance, such as backups or upgrades, must be more carefully thought out because they may leave the cluster vulnerable to another failure. For example, in the Serviceguard configurations discussed in Chapter 2, nodes need to be brought down for maintenance in pairs: one node at each site, so that quorum calculations do not prevent automated recovery if a disaster occurs during planned maintenance.

  Rapid detection of failures and rapid repair of hardware is essential so that the cluster is not vulnerable to additional failures.

  Testing is more complex and requires personnel in each of the data centers. Site failure testing should be added to the current cluster testing plans.

Additional Disaster Tolerant Solutions Information

For information on how to build, configure, and manage disaster tolerant cluster solutions using *Metrocluster Continuous Access XP*, *Metrocluster Continuous Access EVA*, *Metrocluster EMC SRDF*, *Continentalclusters,* and *Three Data Center Architecture* refer to the following guide:

- *Designing Disaster Tolerant HA Clusters Using Metrocluster and Continentalclusters*

On-line versions of the above document and other HA documentation are available at **`http://docs.hp.com`** -> High Availability -> Metrocluster or Continentalclusters.

2 Building an Extended Distance Cluster Using Serviceguard

Simple Serviceguard clusters are usually configured in a single data center, often in a single room, to provide protection against failures in CPUs, interface cards, and software. Extended Serviceguard clusters are specialized cluster configurations, which allow a single cluster to extend across two or three separate data centers for increased disaster tolerance. Depending on the type of links employed, distances of up to 100 km between data centers can be achieved.

This chapter discusses several types of Extended Distance Cluster that use basic Serviceguard technology with software mirroring (using MirrorDisk/UX or Veritas VxVM) and Fibre Channel. Both two data center and three data center architectures are illustrated. This chapter discusses the following:

- Types of Data Link for Storage and Networking
- Two Data Center Architecture
- Two Data Center and Third Location Architectures
- Rules for Separate Network and Data Links
- Guidelines on DWDM Links for Network and Data

NOTE Metropolitan and continental clusters, which include specialized data replication techniques and which require the purchase of separate software, are described in Chapters 1 through 4 of the *Designing Disaster Tolerant HA Clusters Using Metrocluster and Continentalclusters* user's guide located at **`http://docs.hp.com`** -> High Availability.

Types of Data Link for Storage and Networking

FibreChannel technology lets you increase the distance between the components in a Serviceguard cluster, thus making it possible to design a disaster tolerant architecture. The following table shows some of the distances possible with a few of the available technologies, including some of the FibreChannel alternatives.

Table 2-1 Link Technologies and Distances

Type of Link	Maximum Distance Supported
Fast/Wide SCSI	25 meters
Gigabit Ethernet Twisted Pair	50 meters
Short Wave FibreChannel	500 meters
FDDI Networking	100 kilometers *
Long Wave FibreChannel	10 kilometers
Finisar Gigabit Interface Converters (GBICs)	80 kilometers
Dense Wave Division Multiplexing (DWDM)	100 kilometers

The development of DWDM technology allows designers to use dark fiber (high speed communication lines provided by common carriers) to extend the distances that were formerly subject to limits imposed by FibreChannel for storage and Ethernet for network links.

NOTE * FDDI Networking is a ring topology therefore, the maximum effective distance supported between data centers is 50 kilometers.

NOTE Increased distance often means increased cost and reduced speed of connection. Not all combinations of links are supported in all cluster types. For a current list of supported configurations and supported distances, refer to the *HP Configuration Guide*, available through your HP representative. As new technologies become supported, they will be described in that guide.

Two Data Center Architecture

The two data center architecture is based on a standard Serviceguard configuration with half of the nodes in one data center, and the other half in another data center. Nodes can be located in separate data centers in the same building, or even separate buildings within the limits of FibreChannel technology. Configurations with two data centers have the following requirements:

- There must be an equal number of nodes (1 or 2) in each data center.
- In order to maintain cluster quorum after the loss of an entire data center, you must configure dual cluster lock disks (one in each data center). Since cluster lock disks are only supported for up to 4 nodes, the cluster can contain only 2 or 4 nodes. The Serviceguard Quorum Server cannot be used in place of dual cluster disks, as the Quorum Server must reside in a third data center. Therefore, a three data center cluster is a preferable solution, if dual cluster lock disks cannot be used, or if the cluster must have more than 4 nodes. When using dual cluster lock disks, there exists a chance of Split Brain Syndrome (where the nodes in each data center form two separate clusters, each with exactly one half of the cluster nodes) if all communication between the two data centers is lost and all nodes remain running.
  The Serviceguard Quorum Server prevents the possibility of split brain, however the Quorum Server must reside in a third site. Therefore a three data center cluster is a preferable solution, to prevent split brain, and the only solution if dual cluster lock disks cannot be used, or if the cluster must have more than 4 nodes.
- Two data center configurations are not supported if SONET is used for the cluster interconnects between the Primary data centers.
- To protect against the possibility of a split cluster inherent when using dual cluster locks, at least two (three preferred) independent paths between the two data centers must be used for heartbeat and cluster lock I/O. Specifically, the path from the first data center to the cluster lock at the second data center must be different than the path from the second data center to the cluster lock at the first data center. Preferably, at least one of the paths for heartbeat traffic should be different from each of the paths for cluster lock I/O.
- No routing is allowed for the networks between data centers.

- MirrorDisk/UX mirroring for LVM and VxVM mirroring are supported for clusters of 2 or 4 nodes. However, the dual cluster lock devices can only be configured in LVM Volume Groups.
- There can be separate networking and FibreChannel links between the two data centers, or both networking and Fibre Channel can go over DWDM links between the two data centers. See the section below “Network and Data Replication Links Between the Data Centers” for more details.
- CVM 3.5 and CVM 4.1 mirroring is supported for Serviceguard and Extended Cluster for RAC clusters. However, the dual cluster lock devices must still be configured in LVM Volume Groups. Since cluster lock disks are only supported for up to 4 nodes, the cluster can contain only 2 or 4 nodes.
- MirrorDisk/UX mirroring for Shared LVM volume groups is supported for Extended Cluster for RAC clusters containing 2 nodes.
- FibreChannel Direct Fabric Attach (DFA) is recommended over FibreChannel Arbitrated loop configurations, due to the superior performance of DFA, especially as the distance increases. Therefore Fibre Channel switches are preferred over Fibre Channel hubs.
- Any combination of the following FibreChannel capable disk arrays may be used: HP StorageWorks Virtual Arrays, HP StorageWorks Disk Array XP, Enterprise Virtual Arrays (EVA) or EMC Symmetrix Disk Arrays. Refer to the *HP Configuration Guide* (available through your HP representative) for a list of supported FibreChannel hardware.
- Application data must be mirrored between the primary data centers. If MirrorDisk/UX is used, Mirror Write Cache (MWC) must be the Consistency Recovery policy defined for all mirrored logical volumes. This will allow for resynchronization of stale extents after a node crash, rather than requiring a full resynchronization. For SLVM (concurrently activated) volume groups, Mirror Write Cache must not be defined as the Consistency Recovery policy for mirrored logical volumes (*that is, NOMWC must be used*). This means that a full resynchronization may be required for shared volume group mirrors after a node crash, which can have a significant impact on recovery time. To ensure that the mirror copies reside in different data centers, it is recommended to configure physical volume groups for the disk devices in each data center, and to use Group Allocation Policy for all mirrored logical volumes.

- Due to the maximum of 3 images (1 original image plus two mirror copies) allowed in MirrorDisk/UX, if JBODs are used for application data, only one data center can contain JBODs while the other data center must contain disk arrays with hardware mirroring. Note that having three mirror copies will affect performance on disk writes. VxVM and CVM 3.5 mirroring does not have a limit on the number of mirror copies, so it is possible to have JBODS in both data centers, however increasing the number of mirror copies may adversely affect performance on disk writes.
- Veritas Volume Manager (VxVM) from mirroring is supported for distances of up to 100 kilometers for clusters of 16 nodes. However, VxVM supports up to 10 kilometers for clusters of 16 nodes on supported versions of HP-UX. Ensure that the mirror copies reside in different data centers and the DRL (Dirty Region Logging) feature is used. Raid 5 mirrors are not supported. It is important to note that the data replication links between the data centers VxVM can only perform a full resynchronization (that is, it cannot perform an incremental synchronization) when recovering from the failure of a mirror copy or loss of connectivity to a data center. This can have a significant impact on performance and availability of the cluster if the disk groups are large.
- Veritas CVM version 3.5 mirroring is supported for Serviceguard, Serviceguard OPS Edition, or Serviceguard Extension for RAC clusters (SGeRAC) for distances up to 10 kilometers for 2, 4, 6, or 8 node clusters, and up to 100 kilometers for 2 node clusters.

  S*ince CVM 3.5 does not support multiple heartbeats and allows only one heartbeat network to be defined for the cluster, you must make the heartbeat network highly available, using a standby LAN to provide redundancy for the heartbeat network*. The heartbeat subnet should be a dedicated network, to ensure that other network traffic will not saturate the heartbeat network. The CVM Mirror Detachment Policy must be set to “Global”. CVM 4.1 supports multiple heartbeat subnets.
- For clusters using Veritas CVM 3.5, only a single heartbeat subnet is supported, so it is required to have both Primary and Standby LANs configured for the heartbeat subnet on all nodes. For SGeRAC clusters, it is recommended to have an additional network for Oracle RAC cache fusion traffic. It is acceptable to use a single Standby

network to provide backup for both the heartbeat network and the RAC cache fusion network, however it can only provide failover capability for one of these networks at a time.

- Serviceguard Extension for Faster Failover (SGeFF) is not supported in a two data center architecture, which requires a two-node cluster and the use of a quorum server. For more detailed information on SGeFF, refer to the *Serviceguard Extension for Faster Failover Release Notes* and the "*Optimizing Failover Time in a Serviceguard Environment*" white paper.

NOTE

Refer to Table 1-2 for the maximum supported distances between data centers for Extended Distance Cluster configurations.

For more detailed configuration information on Extended Distance Cluster, refer to the *HP Configuration Guide* (available through your HP representative).

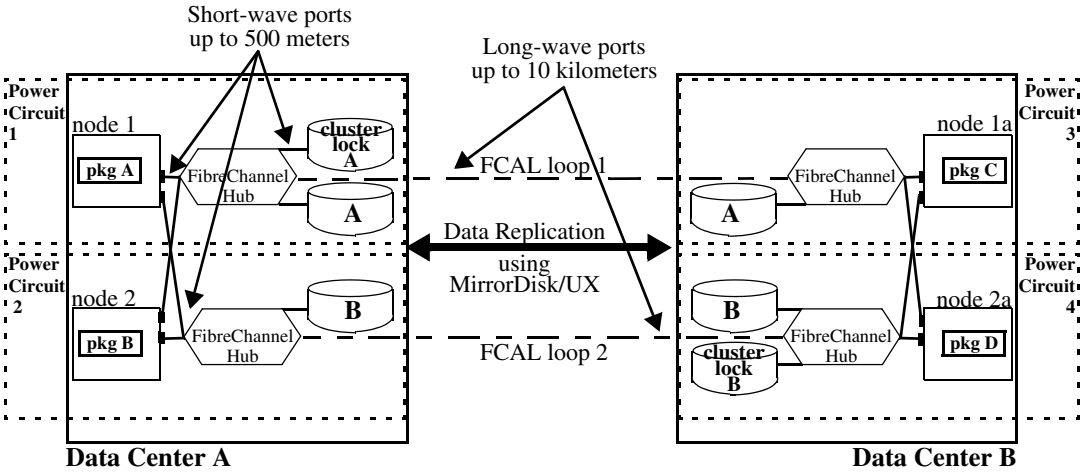
For the most up-to-date support and compatibility information see the *SGeRAC for SLVM, CVM & CFS Matrix and Serviceguard Compatibility and Feature Matrix* on **`http://docs.hp.com`** -> High Availability -> Serviceguard Extension for Real Application Cluster (ServiceGuard OPS Edition) -> Support Matrixes

Two Data Center FibreChannel Implementations

FibreChannel Using Hubs

In a two data center configuration, shown in Figure 2-1, it is required to use a cluster lock disk, which is only supported for up to 4 nodes. This configuration can be implemented using any HP-supported FibreChannel devices. Disks must be available from all nodes using redundant links. Not all links are shown in Figure 2-1.

Figure 2-1 Two Data Centers with FibreChannel Hubs



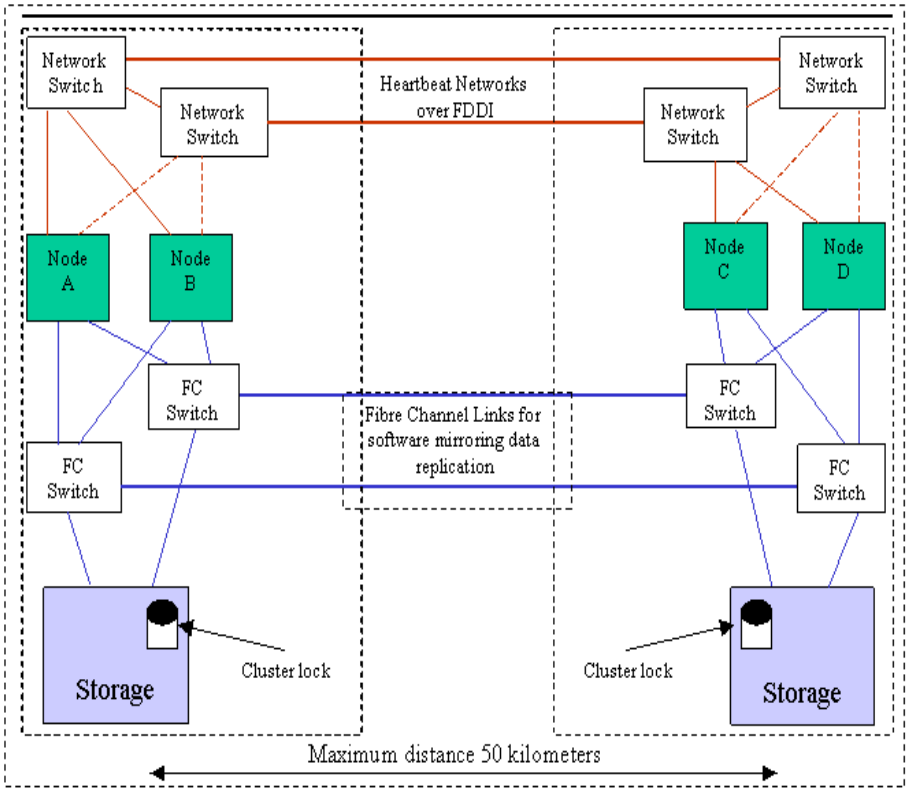
The two cluster lock disks should be located on separate FibreChannel loops to guard against single point of failure. The lock disks can also be used as data disks. They must be connected to all nodes using redundant links (not all links are shown in Figure 2-1).

Nodes can connect to disks in the same data center using short wave ports, and hubs can connect between data centers using long-wave ports. This gives you a maximum distance of 10 kilometers between data centers, making it possible to locate data centers in different buildings.

FibreChannel Using Switches

The two data center architecture is also possible over longer distances using FibreChannel switches. Figure 2-2 is one example of a switched two data center configuration using FibreChannel and FDDI networking.

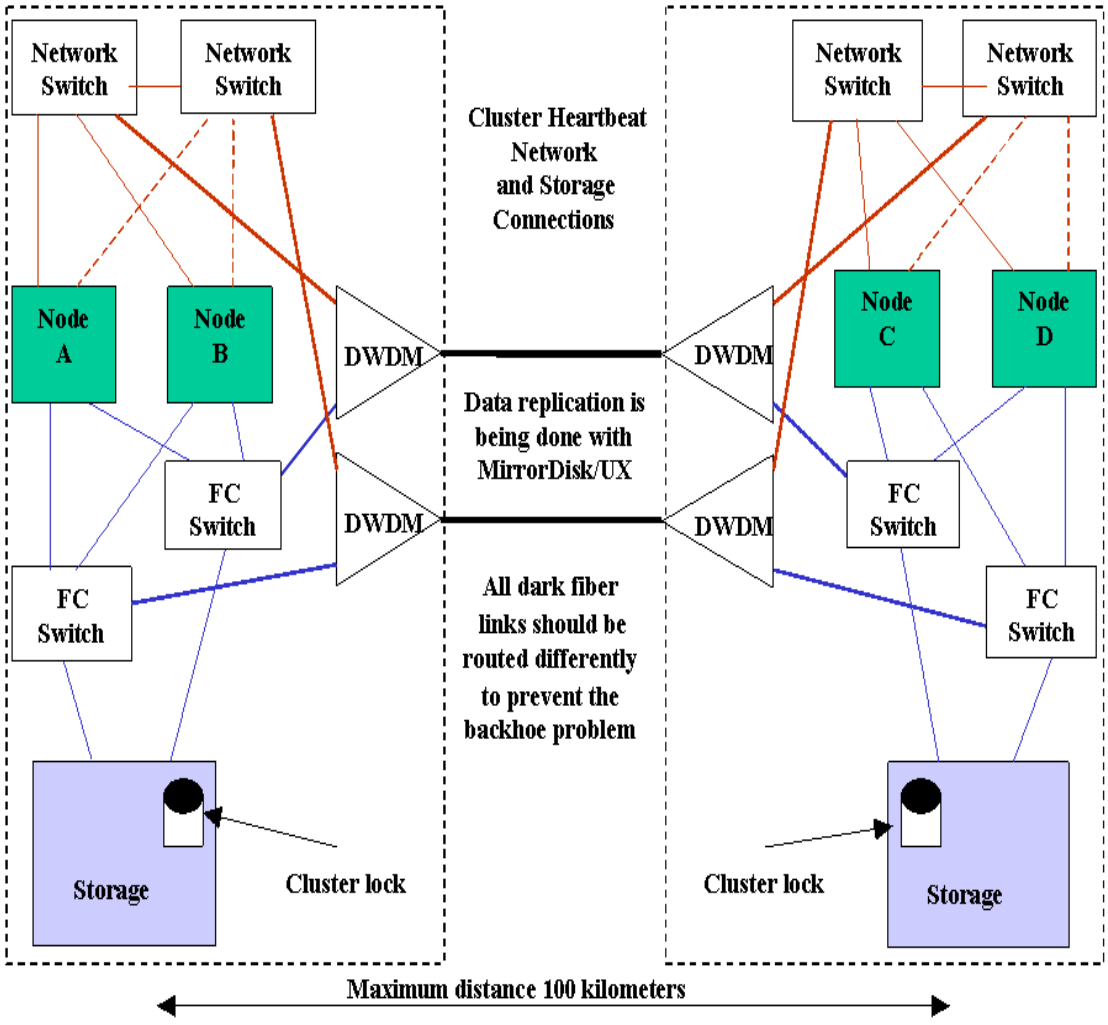
Figure 2-2 Two Data Centers with FibreChannel Switches and FDDI



DWDM with Two Data Centers

Figure 2-3 is an example of a two data center configuration using DWDM for both storage and networking.

Figure 2-3 Two Data Centers with DWDM Network and Storage



Advantages and Disadvantages of a Two Data Center Architecture

The advantages of a two data center architecture are:

- Lower cost.
- Only two data centers are needed, meaning less space and less coordination between operations staff.
- No arbitrator nodes are needed.
- All systems are connected to both copies of data, so that if a primary disk fails but the primary system stays up, there is a greater availability because there is no package failover.

The disadvantages of a two data center architecture are:

- There is a slight chance of split brain syndrome. Since there are two cluster lock disks, a split brain syndrome would occur if the following happened simultaneously:
  - — All heartbeat networks fail.
  - — All disk links fail. The disk link from data center A to cluster lock disk B fails (see Figure 2-1 on page 72.), and the disk link from data center B to cluster lock disk A fails.

  The chances are slight, however these events happening at the same time would result in split brain syndrome and probable data inconsistency. Planning different physical routes for both network and data connections or adequately protecting the physical routes greatly reduces the possibility of split brain syndrome.
- Software mirroring increases CPU overhead.
- The cluster must be either two or four nodes with cluster lock disks. Larger clusters are not supported due to cluster lock requirements.
- Although it is a low cost solution, it does require some additional cost:
  - — FibreChannel links are required for both local and remote connectivity.
  - — All systems must be connected to multiple copies of the data and to both cluster lock disks.

Two Data Center and Third Location Architectures

A two data center and third location have the following configuration requirements:

NOTE

There is no hard requirement on how far the third location has to be from the two main data centers. The third location can be as close as the room next door with its own power source or can be as far as in another site across town. The distance between all three locations dictates that level of disaster tolerance a cluster can provide.

The third location, also known as the Arbitrator data center, can contain either Arbitrator nodes or a Quorum Server node.

- If Arbitrator nodes are used there must be an equal number of nodes (1-7) in each Primary data center, and the third location can contain 1 or 2 Arbitrator nodes. The Arbitrator nodes are standard Serviceguard nodes configured in the cluster, however, they are not allowed to be connected to the shared disks in either of the Primary data centers. Arbitrator nodes are used as tie breakers to maintain cluster quorum when all communication between the two Primary data centers is lost. The data center containing the Arbitrator Nodes must be located separately from the nodes in the Primary data centers. If the Primary data centers each contain a single node, then only one Arbitrator node is allowed. Cluster lock disks are not supported in this configuration. Arbitrator nodes are not supported if CVM or CFS is used in the cluster.
- The third location can contain a single Serviceguard Quorum Server node (running HP-UX or Red Hat Linux), in place of an Arbitrator node(s), however that node must still be located in a third location separate from the Primary data centers, with a separate power circuit. The Quorum Server does not have to be on the same subnet as the cluster nodes, but network routing must be configured such that all nodes in the cluster can contact the Quorum Server via separate physical routes. Since only one IP address can be configured for a Quorum Server, it is suggested to make the Quorum Server more highly available. This is done by running it in it's own

Serviceguard cluster, or to configure the LAN used for the Quorum Server IP address with at least two LAN interface cards using APA (Automatic Port Aggregation) `LAN_MONITOR` mode to improve the availability if a LAN failure occurs. *Prior to Quorum Server revision A.02.00, it was not supported to run the Quorum Server in a Serviceguard cluster*.
For more information about quorum server, refer to the *Managing Serviceguard* user's guide and the *Serviceguard Quorum Server Release Notes*

- No routing is allowed for the networks between the data centers. Routing is allowed to the third location if a Quorum Server is used in that data center site.
- MirrorDisk/UX mirroring for LVM and VxVM mirroring is supported for clusters of up to 16 nodes.
- CVM 3.5 or CVM 4.1 mirroring is supported for Serviceguard and Extended Clusters for RAC clusters containing 2, 4, 6, or 8 nodes*. In CVM and CFS configurations Arbitrator nodes are not supported, and a Quorum Server node must be used instead.
- MirrorDisk/UX mirroring for Shared LVM volume groups is supported for EC RAC clusters containing 2 nodes.
- If Serviceguard OPS Edition or Serviceguard Extension for RAC is used, then there can only be two or four nodes configured to share OPS/RAC data, as MirrorDisk/UX only supports concurrent volume group activation for up to two nodes. CVM and CFS allows for clusters containing 2, 4, 6 or 8 nodes*.
- There can be separate networking and Fibre Channel links between the data centers, or both networking and Fibre Channel can go over DWDM links between the data centers.
- Fibre Channel Direct Fabric Attach (DFA) is recommended over Fibre Channel Arbitrated loop configurations, due to the superior performance of DFA, especially as the distance increases. Therefore Fibre Channel switches are preferred over Fibre Channel hubs.
- Any combination of the following Fibre Channel capable disk arrays may be used: HP StorageWorks FC10, HP StorageWorks FC60, HP StorageWorks Virtual Arrays, HP StorageWorks Disk Array XP or EMC Symmetrix Disk Arrays.

- Application data must be mirrored between the primary data centers. If MirrorDisk/UX is used, Mirror Write Cache (MWC) must be the Consistency Recovery policy defined for all mirrored logical volumes. This will allow for resynchronization of stale extents after a node crash, rather than requiring a full resynchronization. For SLVM (concurrently activated) volume groups, Mirror Write Cache must not be defined as the Consistency Recovery policy for mirrored logical volumes (*that is, NOMWC must be used*). This means that a full resynchronization may be required for shared volume group mirrors after a node crash, which can have a significant impact on recovery time. To ensure that the mirror copies reside in different data centers, it is recommended to configure physical volume groups for the disk devices in each data center, and to use Group Allocation Policy for all mirrored logical volumes.
- Due to the maximum of 3 images (1 original image plus two mirror copies) allowed in MirrorDisk/UX, if JBODs are used for application data, only one data center can contain JBODs while the other data center must contain disk arrays with hardware mirroring. Note that having three mirror copies will affect performance on disk writes. VxVM and CVM mirroring does not have a limit on the number of mirror copies, so it is possible to have JBODS in both data centers, however increasing the number of mirror copies may adversely affect performance on disk writes.
- No routing is allowed for the networks between data centers. Routing is allowed to the third data center if a Quorum Server is used in that data center.
- Veritas Volume Manager (VxVM) mirroring is supported for distances of up to 100 kilometers for clusters of 16 nodes*. However, VxVM supports up to 10 kilometers for clusters of 16 nodes depending on your HP-UX version*. Ensure that the mirror copies reside in different data centers and the DRL (Dirty Region Logging) feature are used. Raid 5 mirrors are not supported. It is important to note that the data replication links between the data centers VxVM can only perform a full resynchronization (that is, it cannot perform an incremental synchronization) when recovering from the failure of a mirror copy or loss of connectivity to a data center. This can have a significant impact on performance and availability of the cluster if the disk groups are large.

- Veritas CVM mirroring is supported for Serviceguard, Serviceguard OPS Edition, or Serviceguard Extension for RAC clusters for distances up to 10 kilometers for 2, 4, 6, or 8 node clusters, and up to 100 kilometers for 2 node clusters*.

  S*ince CVM 3.5 does not support multiple heartbeats and allows only one heartbeat network to be defined for the cluster, you must make the heartbeat network highly available, using a standby LAN to provide redundancy for the heartbeat network*. The heartbeat subnet should be a dedicated network, to ensure that other network traffic will not saturate the heartbeat network. The CVM Mirror Detachment Policy must be set to "Global".

- For clusters using Veritas CVM 3.5, only a single heartbeat subnet is supported, so it is required to have both Primary and Standby LANs configured for the heartbeat subnet on all nodes. For SGeRAC clusters, it is recommended to have an additional network for Oracle RAC cache fusion traffic. It is acceptable to use a single Standby network to provide backup for both the heartbeat network and the RAC cache fusion network, however it can only provide failover capability for one of these networks at a time.

- If Serviceguard Extension for Faster Failover (SGeFF) is used in a two data center and third location architecture, a two node cluster with multiple heartbeats and a quorum server in the third location are required. For more detailed information on Serviceguard Extension for Faster Failover, refer to the *Serviceguard Extension for Faster Failover Release Notes* and the "*Optimizing Failover Time in a Serviceguard Environment*" white paper.

NOTE

* Refer to Table 1-2 for the maximum supported distances between data centers for Extended Distance Cluster configurations.

For more detailed configuration information on Extended Distance Cluster, refer to the *HP Configuration Guide* (available through your HP representative).

For the most up-to-date support and compatibility information see the *SGeRAC for SLVM, CVM & CFS Matrix and Serviceguard Compatibility and Feature Matrix* on **`http://docs.hp.com`** -> High Availability -> Serviceguard Extension for Real Application Cluster (ServiceGuard OPS Edition) -> Support Matrixes

The following table shows the possible configurations using a three data center architecture.

Table 2-2 Supported System and Data Center Combinations

Data Center A	Data Center B	Data Center C	Serviceguard Version
1	1	1 Arbitrator Node	A.11.13 or later
1	1	Quorum Server System	A.11.13 or later
1	1	Quorum Server System	A.11.16 or later (including SGeFF)
2	1	2 Arbitrator Nodes	A.11.13 or later
1	2	2 Arbitrator Nodes	A.11.13 or later
2	2	1 Arbitrator Node	A.11.13 or later
2	2	2* Arbitrator Nodes	A. 11.13 or later
2	2	Quorum Server System	A. 11.13 or later
3	3	1 Arbitrator Node	A. 11.13 or later
3	3	2* Arbitrator Nodes	A. 11.13 or later
3	3	Quorum Server System	A.11.13 or later
4	4	1 Arbitrator Node	A.11.13 or later

Table 2-2 Supported System and Data Center Combinations (Continued)

Data Center A	Data Center B	Data Center C	Serviceguard Version
4	4	2* Arbitrator Nodes	A.11.13 or later
4	4	Quorum Server System	A.11.13 or later
5	5	1 Arbitrator Node	A.11.13 or later
5	5	2* Arbitrator Nodes	A.11.13 or later
5	5	Quorum Server System	A.11.13 or later
6	6	1 Arbitrator Node	A.11.13 or later
6	6	2* Arbitrator Nodes	A.11.13 or later
6	6	Quorum Server System	A.11.13 or later
7	7	1 Arbitrator Node	A.11.13 or later
7	7	2* Arbitrator Nodes	A.11.13 or later
7	7	Quorum Server System	A.11.13 or later
8	8	Quorum Server System	A.11.13 or later

* Configurations with two arbitrators are preferred because they provide a greater degree of availability, especially in cases when a node is down due to a failure or planned maintenance. It is highly recommended that two arbitrators be configured in Data Center C to allow for planned downtime in Data Centers A and B.

NOTE Serviceguard Extension for RAC clusters are limited to 2, 4, 6, or 8 nodes.

The following is a list of recommended arbitration methods for Metrocluster solutions in order of preference:

- 2 arbitrator nodes, where supported
- 1 arbitrator node, where supported
- Quorum Server running in a Serviceguard cluster
- Quorum Server with APA
- Quorum Server

For more information on Quorum Server, refer to the *Serviceguard Quorum Server Version A.01.00 Release Notes for HP-UX*.

Figure 2-4 is an example of a two data center and third location configuration using DWDM, with arbitrator nodes in the third location.

Figure 2-4 Two Data Centers and Third Location with DWDM and Arbitrators

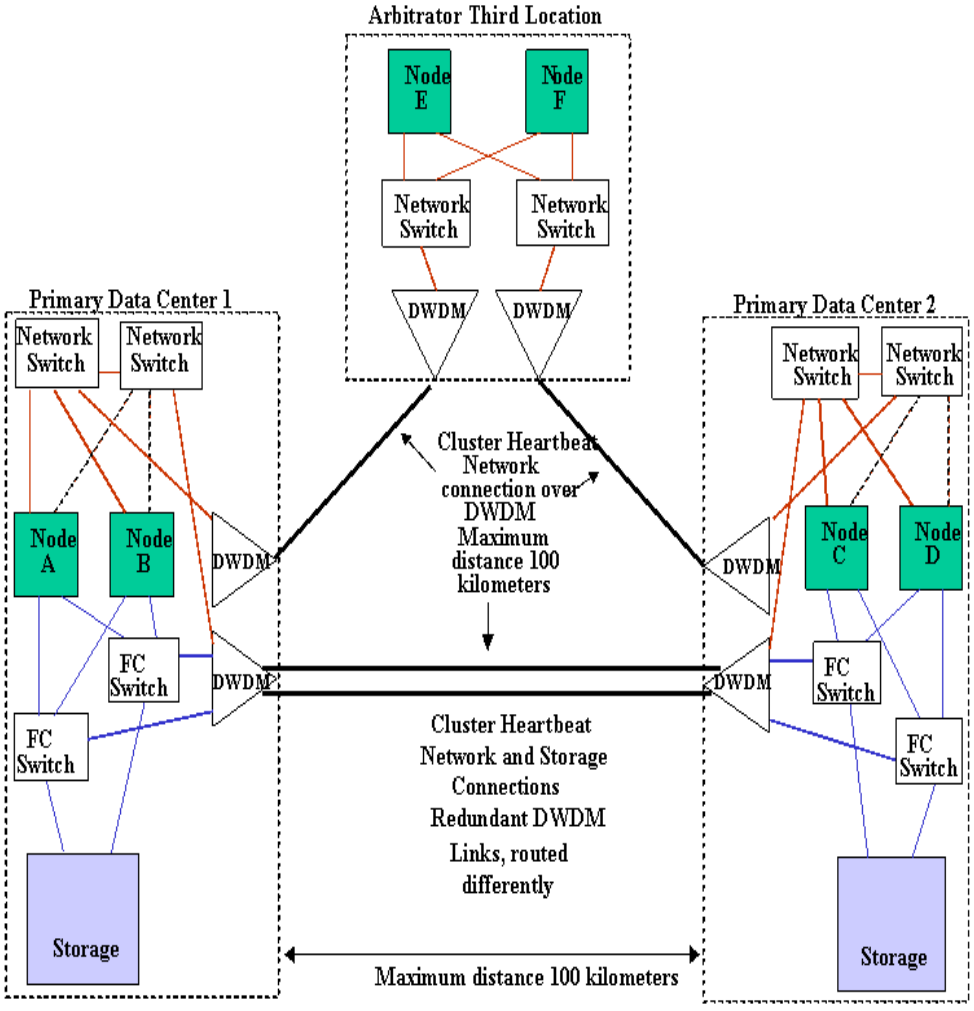


Figure 2-5 Two Data Centers and Third Location with DWDM and Quorum Server

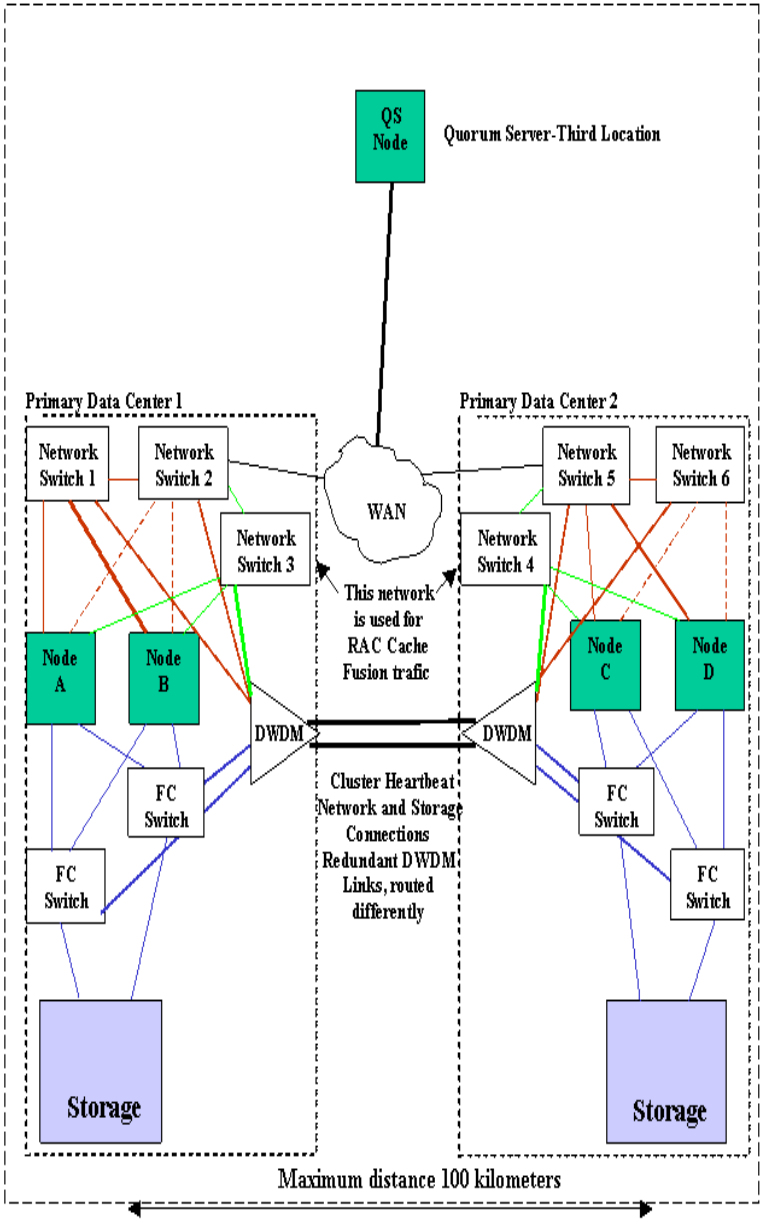


Figure 2-5 is an example of a two data center and third location configuration using DWDM, with a quorum server node on the third site and is specifically for a SGeRAC cluster. The DWDM boxes connected between the two Primary Data Centers are configured with redundant dark fibre links and the standby fibre feature has been enabled.

Note that there is a separate network (indicated by the lines to switches #3 and #4) being used for the RAC Cache Fusion traffic to ensure good RAC performance. Switches #2 and #5 are used for the Standby network, which can provide local LAN failover for both the Primary Heartbeat network and the Primary RAC Cache Fusion network. However it must be noted that the Standby network can only provide local failover capability for one of the Primary networks at a time. For that reason, it is preferable to have a separate Standby network for the Heartbeat network and for the RAC Cache Fusion network.

There are no requirements for the distance between the Quorum Server Data center and the Primary Data Centers, however it is necessary to ensure that the Quorum Server can be contacted within a reasonable amount of time (should be within the NODE_TIMEOUT period). Cluster lock disks are not allowed in this configuration. There can be 2, 4, 6, or 8 nodes in this cluster if CVM 3.5 is used and the distance is 10 kilometers or less. However, there can be only 2 nodes in this cluster if CVM is used, the distance is 100 kilometers and if shared LVM is used.

Since there are 4 nodes shown in this example cluster, this means that this cluster can only use CVM as the volume manager, and the distance between the Primary data centers cannot exceed 10 kilometers.

Rules for Separate Network and Data Links

- The network interfaces used must support DLPI (link level).
- There must be less than 200 milliseconds of latency in the network between the data centers.
- No routing is allowed for the networks between the data centers.
- Routing is allowed to the third data center if a Quorum Server is used in that data center.
- The maximum distance between the data centers for this type of configuration is currently limited by the maximum distance supported for the networking type or Fibre Channel link type being used, whichever is shorter.
- Currently, FDDI offers a total ring length of 100 kilometers. This will allow up to 50 kilometers between data centers for both two and three data center solutions.
- There can be a maximum of 500 meters between the Fibre Channel hubs or switches in the two data centers if Short-wave ports are used. This distance can be increased to 10 kilometers by using a Long-wave Fibre Channel port on the hubs or switches. The distance can be increased to 80 kilometers if Finisar (Long-haul) GBICs are used in Fibre Channel switches (hubs are not supported for this distance). If DWDM links are used, the maximum distance between the data centers is 100 kilometers.
- There must be at least two alternately routed networking links between each primary data center to prevent the “backhoe problem”. The “backhoe problem” can occur when all cables are routed through a single trench and a tractor on a construction job severs all cables and disables all communications between the data centers. It is allowable to have only a single network link routed from each primary data center to the third location, however in order to survive the loss of the network link between a primary data center and the arbitrator data center, the network routing should be configured so that a primary data center can also reach the arbitrator via a route which passes through the other primary data center.

- There must be at least two alternately routed Fibre Channel Data Replication links between each data center. If a third location is used for arbitrator nodes, no Fibre Channel Data Replication links are required for third location.
- Fibre Channel hubs are only supported for distances up to 10 kilometers. For distances longer than 10 kilometers, Fibre Channel switches are required. Fibre Channel switches are always the preferred solution, since they offer much better performance than Fibre Channel hubs, which only support Fibre Channel Arbitrated loop mode.
- Refer to the *HP Configuration Guide* (available through your HP representative) for a list of supported FibreChannel hardware.

Guidelines on DWDM Links for Network and Data

- The network interfaces used must support DLPI (link level).
- There must be less than 200 milliseconds of latency in the network between the data centers.
- No routing is allowed for the networks between the data centers.
- Routing is allowed to the third data center if a Quorum Server is used in that data center.
- The maximum distance supported between the data centers for DWDM configurations is 100 kilometers.
- Both the networking and Fibre Channel Data Replication can go through the same DWDM box - separate DWDM boxes are not required.
- Since DWDM converters are typically designed to be fault tolerant, it is acceptable to use only one DWDM box (in each data center) for the links between each data center. However, for the highest availability, it is recommended to use two separate DWDM boxes (in each data center) for the links between each data center. If using a single DWDM box for the links between each data center the redundant standby fibre link feature of the DWDM box must be configured. If the DWDM box supports multiple active DWDM links, that feature can be used instead of the redundant standby feature.
- At least two dark fibre optic links are required between each Primary data center, each fibre link routed differently to prevent the “backhoe problem.” It is allowable to have only a single fibre link routed from each Primary data center to the third location, however in order to survive the loss of a link between a Primary data center and the third data center, the network routing should be configured so that a Primary data center can also reach the Arbitrator via a route passing through the other Primary data center.
- The network switches in the configuration must support DLPI (link level) packets. The network switch can be 100BaseT (TX or FX), 1000BaseT (TX or FX) or FDDI. The connection between the network switch and the DWDM box must be fibre optic.

- FibreChannel switches must be used in a DWDM configuration; FibreChannel hubs are not supported. Direct Fabric Attach mode must be used for the ports connected to the DWDM link.

Refer to the *HP Configuration Guide*, available through your HP representative, for more information on supported devices.

Additional Disaster Tolerant Solutions Information

For information on how to build, configure, and manage disaster tolerant cluster solutions using *Metrocluster Continuous Access XP*, *Metrocluster Continuous Access EVA*, *Metrocluster EMC SRDF*, *Continentalclusters* and *Three Data Center Architecture* refer to the following guide:

- *Designing Disaster Tolerant HA Clusters Using Metrocluster and Continentalclusters* user's guide.

On-line versions of the above document and other HA documentation are available at **`http://docs.hp.com`** -> High Availability -> Metrocluster or Continentalclusters.

Glossary

A

application restart Starting an application, usually on another node, after a failure. Application can be restarted manually, which may be necessary if data must be restarted before the application can run (example: Business Recovery Services work like this.) Applications can by restarted by an operator using a script, which can reduce human error. Or applications can be started on the local or remote site automatically after detecting the failure of the primary site.

arbitrator Nodes in a disaster tolerant architecture that act as tie-breakers in case all of the nodes in a data center go down at the same time. These nodes are full members of the Serviceguard cluster and must conform to the minimum requirements. The arbitrator must be located in a third data center to ensure that the failure of an entire data center does not bring the entire cluster down. See also **quorum server**.

asynchronous data replication Local I/O will complete without waiting for the replicated I/O to complete; however, it is expected that asynchronous data replication will process the I/Os in the original order.

asymmetrical cluster A cluster that has more nodes at one site than at another. For example, an asymmetrical metropolitan cluster may have two nodes in one building, and three nodes in another building. Asymmetrical clusters are not supported in all disaster tolerant architectures.

automatic failover Failover directed by automation scripts or software (such as Serviceguard) and requiring no human intervention. In a ContinentalClusters environment, the start-up of package recovery groups on the Recovery Cluster without intervention. See also **application restart**.

B

bi-directional configuration A continental cluster configuration in which each cluster serves the roles of primary and recovery cluster for different recovery groups. Also known as a **mutual recovery configuration**.

BC (Business Copy) A PVOL or SVOL in an HP StorageWorks XP series disk array that can be split from or merged into a normal PVOL or SVOL. It is often used to create a snapshot of the data taken at a known point in time. Although this copy, when split, is often consistent, it is not usually current.

BCV (Business Continuity Volume) An EMC Symmetrix term that refers to a logical device on the EMC Symmetrix that may be merged into or split from a regular R1 or R2 logical device. It is often used to create a snapshot of the data taken at a known point in time. Although this copy, when split, is often consistent, it is not usually current.

Business Recovery Service Service provided by a vendor to host the backup systems needed to run mission critical applications following a disaster.

C

campus cluster A single cluster that is geographically dispersed within the confines of an area owned or leased by the organization such that it has the right to run cables above or below ground between buildings in the campus. Campus clusters are usually spread out in different rooms in a single building, or in different adjacent or nearby buildings. See also *Extended Distance Cluster*.

cascading failover Cascading failover is the ability of an application to fail from a primary to a secondary location, and then to fail to a recovery location on a different site. The primary location contains a metropolitan cluster built with Metrocluster EMC SRDF, and the recovery location has a standard Serviceguard cluster.

client reconnect Users access to the backup site after failover. Client reconnect can be transparent, where the user is automatically connected to the application running on the remote site, or manual, where the user selects a site to connect to.

cluster An Serviceguard cluster is a networked grouping of HP 9000 and/or HP Integrity Servers series 800 servers (host systems known as nodes) having sufficient redundancy of software and hardware that a single failure will not significantly disrupt service. Serviceguard software monitors the health of nodes, networks, application services, EMS resources, and makes failover decisions based on where the application is able to run successfully.

cluster alarm Time at which a message is sent indicating that the Primary Cluster is probably in need of recovery. The `cmrecovercl` command is enabled at this time.

cluster alert Time at which a message is sent indicating a problem with the cluster.

cluster event A cluster condition that occurs when the cluster goes down or enters an `UNKNOWN` state, or when the monitor software returns an error. This event may cause an alert messages to be sent out, or it may cause an alarm condition to be set, which allows the administrator on the Recovery Cluster to issue the `cmrecovercl` command. The return of the cluster to the `UP` state results in a cancellation of the event, which may be accompanied by a cancel event notice. In addition, the cancellation disables the use of the `cmrecovercl` command.

cluster quorum A dynamically calculated majority used to determine whether any grouping of nodes is sufficient to start or run the cluster. Cluster quorums prevent split-brain syndrome which can lead to data corruption or inconsistency. Currently at least 50% of the nodes plus a tie-breaker are required for a quorum. If no tie-breaker is configured, then greater than 50% of the nodes is required to start and run a cluster.

command device A disk area in the HP StorageWorks XP series disk array used for internal system communication. You create two command devices on each array, each with alternate links (PV links).

consistency group A set of Symmetrix RDF devices that are configured to act in unison to maintain the integrity of a database. Consistency groups allow you to configure R1/R2 devices on multiple Symmetrix frames in Metrocluster with EMC SRDF.

continental cluster A group of clusters that use routed networks and/or common carrier networks for data replication and cluster communication to support package failover between separate clusters in different data centers. Continental clusters are often located in different cities or different countries and can span 100s or 1000s of kilometers.

Continuous Access A facility provided by the Continuos Access software option available with the HP StorageWorks E Disk Array XP series. This facility enables physical data replication between XP series disk arrays.

D

data center A physically proximate collection of nodes and disks, usually all in one room.

data consistency Whether data are logically correct and immediately usable; the validity of the data after the last write. Inconsistent data, if not recoverable to a consistent state, is corrupt.

data currency Whether the data contain the most recent transactions, and/or whether the replica database has all of the committed transactions that the primary database contains; speed of data replication may cause the replica to lag behind the primary copy, and compromise data currency.

data loss The inability to take action to recover data. Data loss can be the result of transactions being copied that were lost when a failure occurred, non-committed transactions that were rolled back as pat of a recovery process, data in the process of being replicated that never made it to the replica because of a failure, transactions that were committed after the last tape backup when a failure occurred that required a reload from the last tape backup. **transaction processing monitors** (TPM), message queuing software, and synchronous data replication are measures that can protect against data loss.

data mirroring *See See **mirroring**.*

data recoverability The ability to take action that results in data consistency, for example database rollback/roll forward recovery.

data replication The scheme by which data is copied from one site to another for disaster tolerance. Data replication can be either physical (see **physical data replication**) or logical (see **logical data replication**). In a ContinentalClusters environment, the process by which data that is used by the Primary Cluster packages is transferred to the Recovery Cluster and made available for use on the Recovery Cluster in the event of a recovery.

database replication A software-based logical data replication scheme that is offered by most database vendors.

disaster An event causing the failure of multiple components or entire data centers that render unavailable all services at a single location; these include natural disasters such as earthquake, fire, or flood, acts of terrorism or sabotage, large-scale power outages.

disaster protection (Don't use this term?) Processes, tools, hardware, and software that provide protection in the event of an extreme occurrence that causes application downtime such that the application can be restarted at a different location within a fixed period of time.

disaster recovery The process of restoring access to applications and data after a disaster. Disaster recovery can be manual, meaning human intervention is required, or it can be automated, requiring little or no human intervention.

disaster recovery services Services and products offered by companies that provide the hardware, software, processes, and people necessary to recover from a disaster.

disaster tolerant The characteristic of being able to recover quickly from a disaster. Components of disaster tolerance include redundant hardware, data replication, geographic dispersion, partial or complete recovery automation, and well-defined recovery procedures.

disaster tolerant architecture A cluster architecture that protects against multiple points of failure or a single catastrophic failure that affects many components by locating parts of the cluster at a remote site and by providing data replication to the remote site. Other components of disaster tolerant architecture include redundant links, either for networking or data replication, that are installed along different routes, and automation of most or all of the recovery process.

E, F

ESCON Enterprise Storage Connect. A type of fiber-optic channel used for inter-frame communication between EMC Symmetrix frames using EMC SRDF or between HP StorageWorks E XP series disk array units using Continuous Access XP.

event log The default location (`/var/adm/cmconcl/eventlog`) where events are logged on the monitoring ContinentalClusters system. All events are written to this log, as well as all notifications that are sent elsewhere.

Extended Distance Cluster A cluster with alternate nodes located in different data centers separated by some distance. Formerly known as *campus cluster*.

failback Failing back from a backup node, which may or may not be remote, to the primary node that the application normally runs on.

failover The transfer of control of an application or service from one node to another node after a failure. Failover can be manual, requiring human intervention, or automated, requiring little or no human intervention.

filesystem replication The process of replicating filesystem changes from one node to another.

G

gatekeeper A small EMC Symmetrix device configured to function as a lock during certain state change operations.

H, I

heartbeat network A network that provides reliable communication among nodes in a cluster, including the transmission of heartbeat messages, signals from each functioning node, which are central to the operation of the cluster, and which determine the health of the nodes in the cluster.

high availability A combination of technology, processes, and support partnerships that provide greater application or system availability.

J, K, L

local cluster A cluster located in a single data center. This type of cluster is not disaster tolerant.

local failover Failover on the same node; this most often applied to hardware failover, for example local LAN failover is switching to the secondary LAN card on the same node after the primary LAN card has failed.

logical data replication A type of on-line data replication that replicates logical transactions that change either the filesystem or the database. Complex transactions may result in the modification of many diverse physical blocks on the disk.

LUN (Logical Unit Number) A SCSI term that refers to a logical disk device composed of one or more physical disk mechanisms, typically configured into a RAID level.

M

M by N A type of Symmetrix grouping in which up to two Symmetrix frames may be configured on either side of a data replication link in a Metrocluster with EMC SRDF configuration. M by N configurations include 1 by 2, 2 by 1, and 2 by 2.

manual failover Failover requiring human intervention to start an application or service on another node.

Metrocluster A Hewlett-Packard product that allows a customer to configure an Serviceguard cluster as a disaster tolerant metropolitan cluster.

metropolitan cluster A cluster that is geographically dispersed within the confines of a metropolitan area requiring right-of-way to lay cable for redundant network and data replication components.

mirrored data Data that is copied using mirroring.

mirroring Disk mirroring hardware or software, such as MirrorDisk/UX. Some mirroring methods may allow splitting and merging.

mission critical application Hardware, software, processes and support services that must meet the uptime requirements of an organization. Examples of mission critical application that must be able to survive regional disasters include financial trading services, e-business operations, 911 phone service, and patient record databases.

mission critical solution The architecture and processes that provide the required uptime for mission critical applications.

multiple points of failure (MPOF) More than one point of failure that can bring down an Serviceguard cluster.

multiple system high availability

Cluster technology and architecture that increases the level of availability by grouping systems into a cooperative failover design.

mutual recovery configuration A continental cluster configuration in which each cluster serves the roles of primary and recovery cluster for different recovery groups. Also known as a **bi-directional configuration**.

N

network failover The ability to restore a network connection after a failure in network hardware when there are redundant network links to the same IP subnet.

notification A message that is sent following a cluster or package event.

O

off-line data replication. Data replication by storing data off-line, usually a backup tape or disk stored in a safe location; this method is best for applications that can accept a 24-hour recovery time.

on-line data replication Data replication by copying to another location that is immediately accessible. On-line data replication is usually done by transmitting data over a link in real time or with a slight delay to a remote site; this method is best for applications requiring quick recovery (within a few hours or minutes).

P

package alert Time at which a message is sent indicating a problem with a package.

package event A package condition such as a failure that causes a notification message to be sent. Package events can be accompanied by alerts, but not alarms. Messages are for information only; the `cmrecovercl` command is not enabled for a package event.

package recovery group A set of one or more packages with a mapping between their instances on the Primary Cluster and their instances on the Recovery Cluster.

physical data replication An on-line data replication method that duplicates I/O writes to another disk on a physical block basis. Physical replication can be hardware-based where data is replicated between disks over a dedicated link (for

example EMC's Symmetrix Remote Data Facility or the HP StorageWorks E Disk Array XP Series Continuous Access), or software-based where data is replicated on multiple disks using dedicated software on the primary node (for example, MirrorDisk/UX).

planned downtime An anticipated period of time when nodes are taken down for hardware maintenance, software maintenance (OS and application), backup, reorganization, upgrades (software or hardware), etc.

PowerPath A host-based software product from Symmetrix that delivers intelligent I/O path management. PowerPath is required for M by N Symmetrix configurations using Metrocluster with EMC SRDF.

Primary Cluster A cluster in production that has packages protected by the HP ContinentalClusters product.

primary package The package that normally runs on the Primary Cluster in a production environment.

pushbutton failover Use of the `cmrecovercl` command to allow all package recovery groups to start up on the Recovery Cluster following a significant cluster event on the Primary Cluster.

PV links A method of LVM configuration that allows you to provide redundant disk interfaces and buses to disk arrays, thereby protecting against single points of failure in disk cards and cables.

PVOL A primary volume configured in an XP series disk array that uses Continuous Access. PVOLs are the primary copies in physical data replication with Continuos Access on the XP.

Q

quorum *See See **cluster quorum***.

quorum server A cluster node that acts as a tie-breaker in a disaster tolerant architecture in case all of the nodes in a data center go down at the same time. See also **arbitrator**.

R

R1 The Symmetrix term indicating the data copy that is the primary copy.

R2 The Symmetrix term indicating the remote data copy that is the secondary copy. It is normally read-only by the nodes at the remote site.

Recovery Cluster A cluster on which recovery of a package takes place following a failure on the Primary Cluster.

recovery group failover A failover of a package recovery group from one cluster to another.

recovery package The package that takes over on the Recovery Cluster in the event of a failure on the Primary Cluster.

regional disaster A disaster, such as an earthquake or hurricane, that affects a large region. Local, campus, and proximate metropolitan clusters are less likely to protect from regional disasters.

remote failover Failover to a node at another data center or remote location.

resynchronization The process of making the data between two sites consistent and current once systems are restored following a failure. Also called data resynchronization.

rolling disaster A second disaster that occurs before recovering from a previous disaster, for example, while data is being synchronized between two data centers after a disaster, one of the data centers fails, interrupting the data synchronization process. Rolling disasters may result in data corruption that requires a reload from tape backups.

S

single point of failure (SPOF) A component of a cluster or node that, if it fails, affects access to applications or services. See also **multiple points of failure**.

single system high availability

Hardware design that results in a single system that has availability higher than normal. Hardware design examples are:

- n+1 fans
- n+1 power supplies
- multiple power cords
- on-line addition or replacement of I/O cards, memory, etc.

special device file The device file name that the HP-UX operating system gives to a single connection to a node, in the format `/dev/devtype/`*`filename`*.

split-brain syndrome When a cluster reforms with equal numbers of nodes at each site, and each half of the cluster thinks it is the authority and starts up the same set of applications, and tries to modify the same data, resulting in data corruption. Serviceguard architecture prevents split-brain syndrome in all cases unless dual cluster locks are used.

SRDF (Symmetrix Remote Data Facility) A level 1-3 protocol used for physical data replication between EMC Symmetrix disk arrays.

SVOL A secondary volume configured in an XP series disk array that uses Continuous Access. SVOLs are the secondary copies in physical data replication with Continuos Access on the XP.

SymCLI The Symmetrix command line interface used to configure and manage EMC Symmetrix disk arrays.

Symmetrix device number The unique device number that identifies an EMC logical volume.

synchronous data replication Each data replication I/O waits for the preceding I/O to complete before beginning another

replication. Minimizes the chance of inconsistent or corrupt data in the event of a rolling disaster.

T

transaction processing monitor (TPM)

Software that allows you to modify an application to store in-flight transactions in an external location until that transaction has been committed to all possible copies of the database or filesystem, thus ensuring completion of all copied transactions. A TPM protects against data loss at the expense of the CPU overhead involved in applying the transaction in each database replica.

Software that provides a reliable mechanism to ensure that all transactions are successfully committed. A TPM may also provide load balancing among nodes.

transparent failover A client application that automatically reconnects to a new server without the user taking any action.

transparent IP failover Moving the IP address from one network interface card (NIC), in the same node or another node, to another NIC that is attached to the same IP subnet so that users or applications may always specify the same IP name/address whenever they connect, even after a failure.

U-Z

volume group In LVM, a set of physical volumes such that logical volumes can be defined within the volume group for user access. A volume group can be activated by only one node at a time unless you are using Serviceguard OPS Edition. Serviceguard can activate a volume group when it starts a package. A given disk can belong to only one volume group. A logical volume can belong to only one volume group.

WAN data replication solutions Data replication that functions over leased or switched lines. See also **continental cluster**.

Index

Index