Geoclusters

Geoclusters are HA clusters stretched across long distances, such as the following:

- C*ampus cluster*—Nodes across buildings in a campus
- *Metro cluster*—Nodes placed within metro distances (for example, from a few kilometers to 50 km)
- *Regional cluster*—Nodes that are hundreds of kilometers apart
- *Continental cluster*—Nodes that are thousands of kilometers apart.

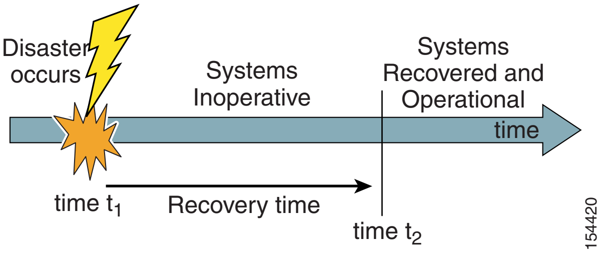
This document refers generically to *metro, regional,* and *continental clusters* as geoclusters.

Geoclusters Overview

The use of geoclusters is very relevant in the context of business continuance as a technology to lower the recovery time objective. Business continuance requirements are measured based on the following:

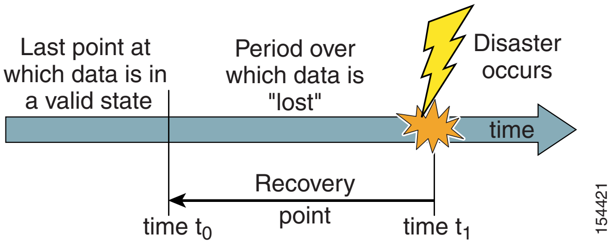
- Recovery time objective (RTO)—How long it takes for the enterprise systems to resume operation after a disaster, as shown in Figure 3-1. RTO is the longest time that your organization can tolerate.

Figure 3-1 Recovery Time and Recovery Time Objective



- Recovery point objective (RPO)—How current or fresh the data is after a disaster, as shown in Figure 3-2. RPO is the maximum data loss after a disaster.

Figure 3-2 Recovery Point and Recovery Point Objective


Last point at which data is in a valid state
Period over which data is "lost"
Disaster occurs
time
Recovery point
time t_0
time t_1
154421

The distance between the data centers and how well applications tolerate network latency determine whether zero RPO is possible. The following recovery technologies support increasingly lower RPO at increasingly higher cost:

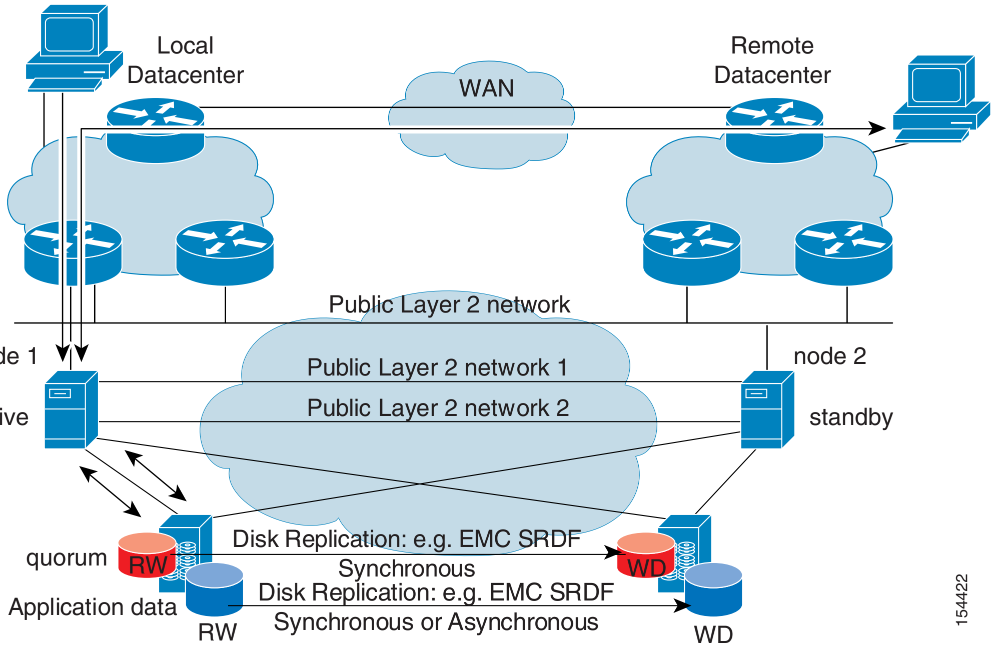
- Tape backup and restore
- Periodic replication and backups
- Asynchronous replication
- Synchronous replication

The design of a geocluster requires investigation and design choices in areas that include the following:

- Assessment of the available fiber, type of fiber, and distance to interconnect the data center
  - Choice of a way to multiplex LAN and SAN onto the fiber, such as using CWDM or DWDM
  - Choice of a service provider Metro Ethernet or SONET offering
  - Choice of the protocols used to share this transport infrastructure
- Choice of a data replication technology, such as disk-based replication or host-based mirroring, which is in turn influenced by the distance and performance required by the application
- Choice of the cluster technology, integration of the cluster software with the transport infrastructure, NIC teaming, and host bus adapter (HBA) multipath input/output (MPIO). Design of the cluster itself in terms of the following:
  - Number of nodes
  - Bandwidth requirements
  - Local failure versus remote failover of the nodes
  - Performance of the application when the cluster is operated from the remote site
- Integration of the server farm switching infrastructure with the cluster, the replication technology and with the client-to-server routing (DNS routing and regular IP routing)

Figure 3-3 shows a high level representation of geoclusters spanning two data centers. Node1 and node2 share a common Layer 2 segment, as is often required by the clustering software. Node1 and node2 also monitor the health of their peer by using multiple private LAN segments, called Private Layer 2 network 1 and Private Layer 2 network 2.

Figure 3-3 Geocluster—High Level Topology



Access to the storage happens on an extended SAN where depending on the specific configuration, both node1 and node2 might be zoned to see storage array1, or simply zoned to see the local storage.

The quorum disk and the disk used for the data are replicated from the first site to the remote site. In Figure 3-3, the replication mechanism for the quorum is synchronous (which is a requirement of Microsoft Cluster Server [MSCS]); and for the data disk, it is asynchronous.

The disks are in read-write mode in the primary site, and in write-disabled mode in the secondary site.

Replication and Mirroring

The main difference between local clusters and geographical clusters is the presence of a disk array at each site. If one site is not available, the application must be restarted at the remote site, requiring an additional disk at the remote site.

For the application data to be available at both sites, you need to choose between two main categories of "replication": *host-based mirroring* (such as Veritas Volume Manager), and *disk-based replication* (such as EMC Symmetrix Remote Data Facility).

Following is a list of commonly-used replication products:

- Veritas Volume Replicator—Performs either synchronous or asynchronous replication. It can replicate over any distance: campus, metro, or global.
- IBM Peer-to-Peer Remote Copy (PPRC)—Remote mirroring hardware-based solution of the IBM Enterprise Storage Server. It can be either synchronous or asynchronous.
- EMC Symmetrix Remote Data Facility (SRDF)—An online host-independent mirrored data solution. It duplicates production site data on one or more physically separate target Symmetrix systems, and can operate in either synchronous or asynchronous mode.
- HP Data Replication Manager (DRM)—Mirrors online data in real time to remote locations via local or extended SANs. It can operate in either synchronous or asynchronous mode.

- Hitachi TrueCopy—Replicates information locally between Hitachi Freedom Storage systems within the data center, or to remote models in distributed centers anywhere in the world, with minimum performance impact. TrueCopy is available for both synchronous and asynchronous replication.

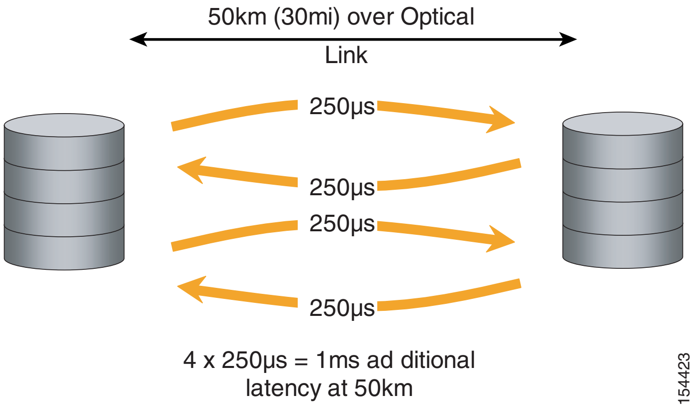
Host-based mirroring is a synchronous mirroring method where the host itself is responsible for duplicating writes to the storage arrays. However, because the two storage arrays are identical copies, reads are performed only against one array.

Disk-based replication uses storage arrays, which can be equipped with data replication software. Replication is performed transparently to the host without additional processing overhead. Each manufacturer has a variety of replication products that can be categorized as either *synchronous* or *asynchronous*. Synchronous replication is a zero data loss (or zero RPO) data replication method. All data is written to the local storage array and the remote array before the I/O is considered complete or acknowledged to the host. Disk I/O service time increases as distance increases because the I/O must be completed at the remote storage array. Higher disk I/O service times negatively impact application performance.

When using an optical network, the additional network latency because of speed of light through fiber is approximately 5 µs per kilometer (8 µs per mile). At two rounds trips per write, the additional service time accumulates at 20 µs per kilometer. For example at 50 km, the additional time is 1000 µs or 1 ms.

Figure 3-4 shows the network latency for synchronous replication.

Figure 3-4 Network Latency for Synchronous Replication



Asynchronous replication is a real-time replication method in which the data is replicated to the remote array after the I/O is acknowledged as complete to the host. This means application performance is not impacted. The enterprise can therefore locate the remote array virtually any distance away from the primary data center without impact. However, because data is replicated at some point after local acknowledgement, the storage arrays are slightly out-of-step; the remote array is behind the local array. If the local array at the primary data center breaks down, some data loss results.

In the case of clusters, some disks may or may not be synchronously or asynchronously replicated. For example, if you are using MSCS with the quorum disk concept, the quorum disk can be replicated only synchronously. This does not necessarily mean that the cluster cannot span more than typical synchronous distances (such as, say ~100 km). The I/O performance required by the quorum disk might be compatible with longer distances as long as the replication is synchronous.

Whether an application can be operated with synchronous or asynchronous replication depends on the read/write (R/W) ratio, the distance between the sites, the software that is used as an interface between the cluster software, the disks, and so forth.

Geocluster Functional Overview

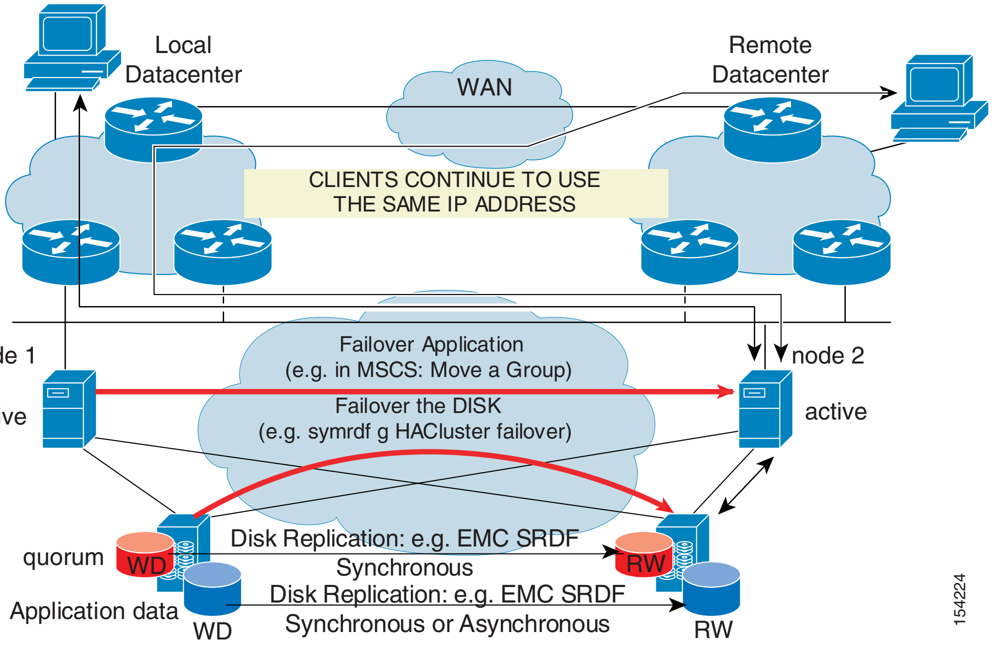
Disk replication and clustering software alone might not be enough to build a geocluster solution. When a failover happens, you may want the disks to failover also to the data center where the server nodes are active. This may require additional software or scripting. Typical products that are used for this purpose include the following:

- EMC SRDF/Cluster Enabler (CE), also known as Geospan
- EMC/Legato Autostart, also known as Automated Availability Manager (AAM)
- HP Continental Clusters
- IBM Geographical Disperse Parallel Sysplex (GDPS)

In the context of this document, the approach of using this software component is generically referred to as "assisted disk failover" to contrast it to the procedure of failing over the disks manually (that is, making the disks at the remote site R/W, pointing the servers at the remote site to the remote disks, and potentially replicating from the remote site to the local site).

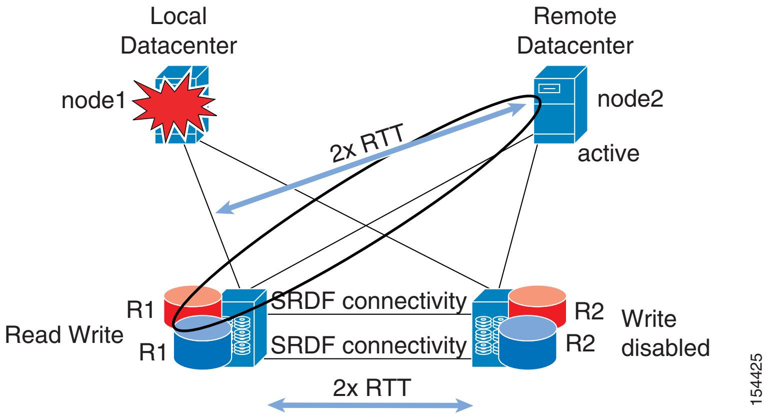
Figure 3-5 shows how a failure scenario can appear with geoclusters present. The administrator of the local data center might *move* a group to the remote data center. The disks should move together with the application; that is, the disks in the remote data center should become R/W, while the disks in the primary site should become write disabled. In the SYMCLI (Symmetrix Command Line Interface) syntax from EMC, this is equivalent to issuing the *failover* command **symrdf –g HAcluster failover** where "HAcluster" simply identifies the disks associated with the application that is being *moved.* After the failover, client requests are taken by node2, which writes and reads to its local disk. Figure 3-5 shows the traffic still entering from the local data center. If the primary site is completely isolated, traffic can enter directly from the remote data center.

Figure 3-5 "Assisted Disk Failover" with Geoclusters



The failover scenario described in Figure 3-5 works when using products such as the *cluster enabler*. If not, the group move appears more like that shown in Figure 3-6. In this case, the node from the remote site continues to read and write to the disks in the primary site with a potential performance issue.

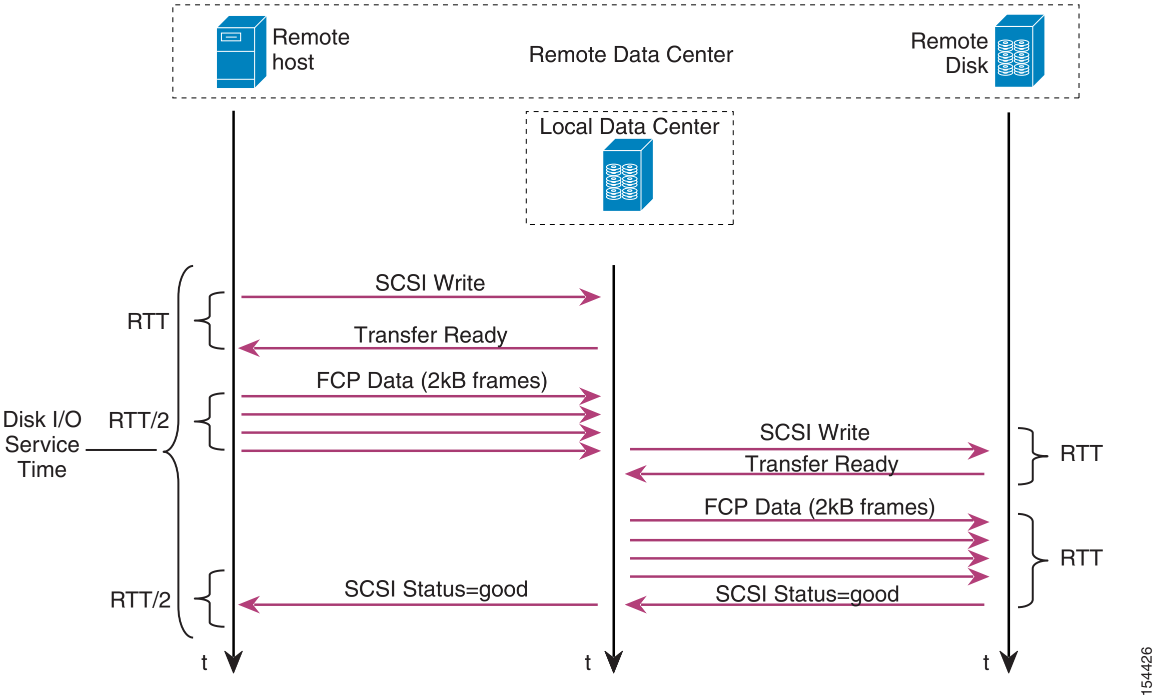
Figure 3-6 Node Failover without "Assisted Disk Failover" (Cluster Enabler and Similar Products)



This clearly poses a problem if the primary site goes down completely, because you must then reconfigure node2, the zoning on the SAN, and the Logical Unit Number (LUN) masking, and restart the application manually.

Figure 3-6 also shows that besides the aspects of the manual intervention, there may be a significant performance implication if the administrator wants to operate the application from site2. Remember that with the SCSI write sequence of operations, as shown in Figure 3-7, you have to consider four round trip times (RTTs) per each write. Whether this is a problem or not depends on the distance between the data centers, the R/W ratio of the application, and the bandwidth available between the sites.

Figure 3-7 Latency Considerations



This demonstrates that for reasons of performance and business continuance, it is very important to consider a solution capable of failing over the disks, whether this is based on software such as the cluster enabler, or it is based on invoking scripts from the clustering software itself. These scripts cause the disks to failover.

Several additional considerations are required to complete the geocluster solution. Depending on the desired distance, you might need to consider various software combinations. For example, the combination of MSCS with Windows 2000 and EMC SRDF/CE works well for metro distances, but cannot be extended to continental distances:

- MSCS requires the quorum disk to be synchronously replicated
- The disk used by the application can be replicated asynchronously
- SRDF/CE requires the disks to be configured for synchronous replication

This means that theoretically, an MSCS cluster with disk-based quorum can be used for continental distances (because the performance required by the quorum disk might be compatible with the latency introduced by synchronous replication), but even so, the SRDF/CE component does not work at such distances.

To build clusters at greater distances, consider the following:

- A majority node set approach instead of a disk-based one for the quorum (assuming that the quorum disk is the limit)
- Scripts to drive the disk failover
- A different clustering/cluster enabler product; for example, EMC/Legato Autostart makes it possible to build clusters with asynchronous disks

After you have found the software that is compatible with the type of cluster that you want to build, you need to consider the best transport mechanism.

Geographic Cluster Performance Considerations

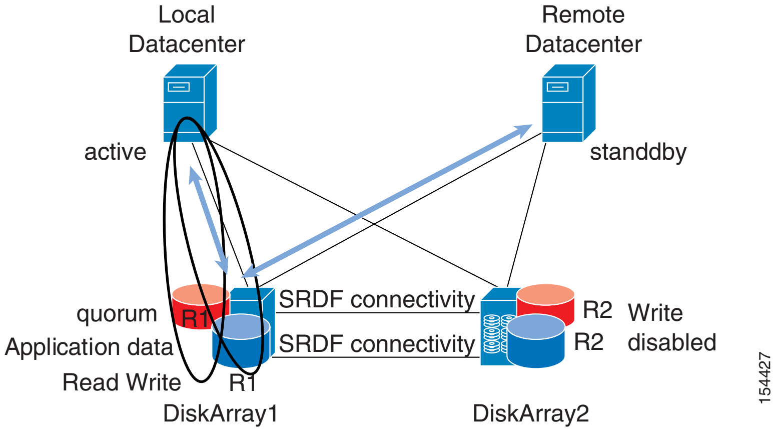
Various factors impact the performance of a geographically-clustered application, including the following:

- Available bandwidth
- R/W ratio of the application
- Record size
- Distance, and as a result, latency
- Synchronous or asynchronous replication on the disk

Tools such as IOmeter (http://www.iometer.org/) provide a method to benchmark the disk access performance of a server to a disk array. Running this tool on the local and remote nodes in the cluster provides useful information on which cluster configuration is viable.

Assume the reference cluster configuration in Figure 3-8, with local and remote node configured to see DiskArray1.

Figure 3-8 Reference Configuration



Assume that DiskArray1 uses synchronous replication. The performance of node1 is affected by the distance between the two sites, because for each write, it has to wait for an explicit acknowledgement from the remote site. The theoretical latency added equals 2xRTTs. The performance of node2 writing to DiskArray1 includes the latency experienced by node1 plus the latency of sending a write across the geographical network, for a total of 4xRTT, as described in Figure 3-7.

This guide describes the following performance characteristics:

- Maximum throughput—Measured with a record size of 64 KB
- Maximum I/Os per second (IOPS)—Measured with a record size of 512 B
- Response time

Server Performance Considerations

The servers used for the testing had the following characteristics:

- Intel® Pentium 4 2.8 GHz CPU
- 1 GB or PC3200 DDR RAM
- Supermicro P4SC8 motherboard (http://www.supermicro.com/products/motherboard/P4/E7210/P4SC8.cfm)
- Dual integrated Gigabit NICs
- Emulex LP8000 Host Bus Adapters (http://www.emulex.com/products/eol/lp8000.html), full duplex 1 Gbps PCI Fibre Channel (FC) adapter. This adapter driver used in conjunction with Windows 2000 can support up to 150 outstanding I/Os.

Note Windows 2003 increases the number of outstanding I/Os to 254.

Factors that influence the maximum I/O performance measured on a server include the following:

- BUS technology—PCI (32 bits@33 MHz = 133 Mbps; 64 bits@33 MHz = 533 Mbps maximum); PCI-X (64 bits@133 MHz = 1 Gbps); PCI-Express (~250 Mbps per lane; for example, x4, and x8)

- HBA—In this case, a PCI 1 Gbps FC HBA
- Read/Write ratio—Read is faster because data is read from a cache, and also because the read involves only the local storage in the presence of a storage array with replication, whereas the write also involves the remote storage.
- Disk performance—See Disk Performance Considerations, page 3-9.

Note The server and disk array configuration described in this section is *not* a recommendation. The purpose of this section is to give enough data to be able to identify the bottlenecks in the network (LAN and SAN) design.

The following sections highlight design considerations that need to be evaluated on a per-application/customer basis. Use the numbers simply as a relative comparison between the design options. As a reference, with this configuration and no synchronous or asynchronous replication, the maximum performance with 70 percent read and 30 percent write is as follows:

- Maximum throughput—114 MBps (without using MPIO)
- Maximum IOPS—8360 IOPS (without using MPIO) with an average response time of 2 ms

Disk Performance Considerations

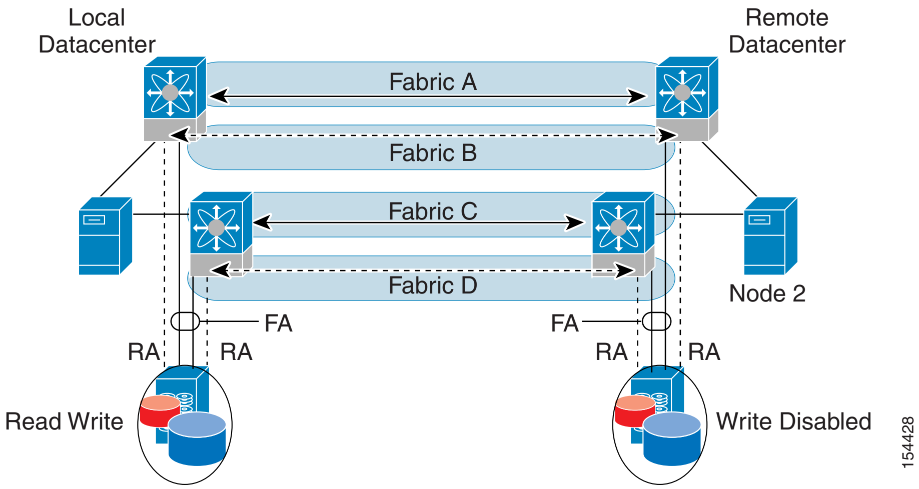
The performance of the disks is a key component in evaluating the performance of clustered servers. Most HA cluster deployments use disk arrays. The storage array performance depends on factors that include the following:

- Disk—Depends on several factors, such as revolutions per minute (RPM), redundant array of independent disks (RAID) type, type of access (*random*, as in the case of a database; *sequential*, as in the case of streaming), and disk connectivity to the SAN. The storage used in these tests consisted of EMC Symmetrix DMX1000 (http://www.emc.com/products/systems/symmetrix/symmetrix_DMX1000/pdf/DMX1000.pdf). Each drive can have a rotational speed of 15,000 RPM, which can yield between 600 Mbps and 900 Mbps.
- The drives can be configured for striping to combine the throughput of the individual drives. The Cisco test bed used RAID5.
- Storage array connectivity—Storage arrays are in turn connected to the SAN with, typically, 2 Gbps Fibre Channel ports. In the configuration used for this test, each storage array connects to two fabrics for client access with 2 Gbps Fibre Channel ports. The storage arrays communicate for the replication traffic across two extended fabrics, and connect to these fabrics with 2 Gbps Fibre Channel ports.
- Synchronous or asynchronous replication—See Asynchronous Versus Synchronous Replication, page 3-19.

Note The server and disk array configuration described in this section is *not* a recommendation. The purpose of this section is to give enough data to be able to identify the bottlenecks in the network (LAN and SAN) design.

Figure 3-9 shows a possible SAN/disk array configuration. The servers can be connected with MPIO with 1 Gbps or 2 Gbps HBAs, and the disk arrays are typically connected with 2 Gbps FC links. The “RA” ports are used for the replication traffic; the two “FA” ports are used for initiator-target communication and provide an aggregate bandwidth of 4 Gbps for the hosts. Typical oversubscription levels in the SAN allow 6:1 or even 12:1 oversubscription between initiator and targets. Fabric A and B are the extended fabrics for the initiator-target communication; Fabric C and D are the extended fabrics for the replication traffic.

Figure 3-9 Redundant Storage Arrays and Extended SAN



Transport Bandwidth Impact on the Application Performance

Assume an application with a R/W ratio of 70/30 and synchronous replication. The maximum throughput performance decreases as shown in Figure 3-10.

Note Note that in the context of this section, the throughput refers to the maximum amount of data per second that the application can write/read to/from the disk. Because the disk is synchronously mirrored, this performance may or may not be bound to the bandwidth available to connect the data centers.

Figure 3-10 Maximum Application Throughput—Fractional OC3 versus GigabitEthernet

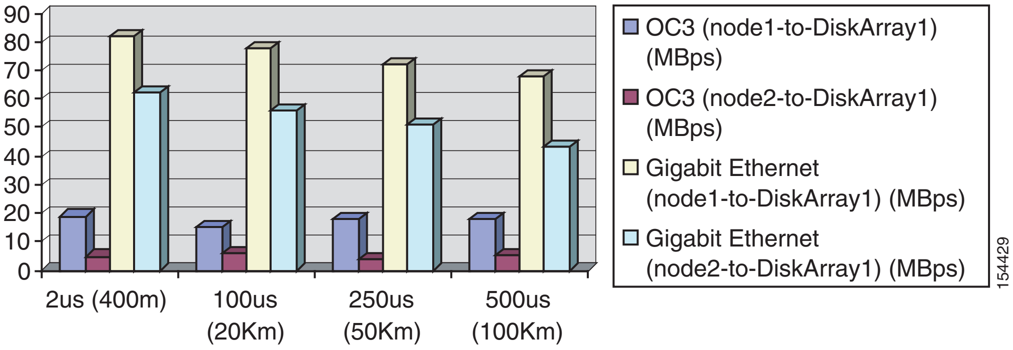


Figure 3-10 shows the variation of the maximum throughput at increasing distances, and contrasts the achievable performance with the following scenarios:

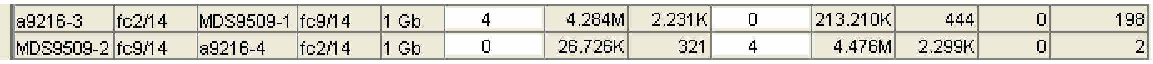
- Writing from node1 to DiskArray1 with an OC-3 link between the sites

- Writing from node2 to DiskArray1 with an OC-3 link between the sites
- Writing from node1 to DiskArray1 with a Gigabit link between the sites
- Writing from node2 to DiskArray1 with a Gigabit link between the sites

The following transport is used for the measurement of Figure 3-10:

- OC-3—With 2 x STS-1s (2 x 51.84 Mbps = ~13 MBps) for FC over SONET, and 1 x STS-1 (51.84 Mbps= ~6.48 MBps) for Ethernet over SONET. As shown by Figure 3-10, the "local" node can fill the maximum throughput (that is, it reaches ~20 MBps) because the bottleneck is really the FC over SONET connectivity. The remote node is constrained by the fact that each disk operation travels on the extended SAN twice: from node2 to DiskArray1, and from DiskArray1 to replicate to DiskArray2. It is evident that the bottleneck in this configuration is not the server but the OC3 link. The sixth and ninth column in Figure 3-11 show the percentage of the FC 1 Gbps link utilization (connected to the OC-3 link). 4 percent indicates the replication traffic is using 4 percent, or in other words, ~40 Mbps.

Figure 3-11 Fibre Channel Link Utilization with OC3 (STS-1) Transport



a9216-3	fc2/14	MDS9509-1	fc9/14	1 Gb	4	4.284M	2.231K	0	213.210K	444	0	198
MDS9509-2	fc9/14	a9216-4	fc2/14	1 Gb	0	26.726K	321	4	4.476M	2.299K	0	2

- GigabitEthernet—This configuration uses a Gigabit Ethernet point-to-point link to transport both Ethernet traffic and Fibre Channel over IP (FCIP) traffic across the sites. Considering that the server platforms used in the first configuration and in this configuration are the same, it is very evident that the performance gain is enormous because the server is no longer constrained by the bandwidth connecting the two sites. There is a ~25 percent performance penalty when operating the cluster from node2 writing to DiskArray1.

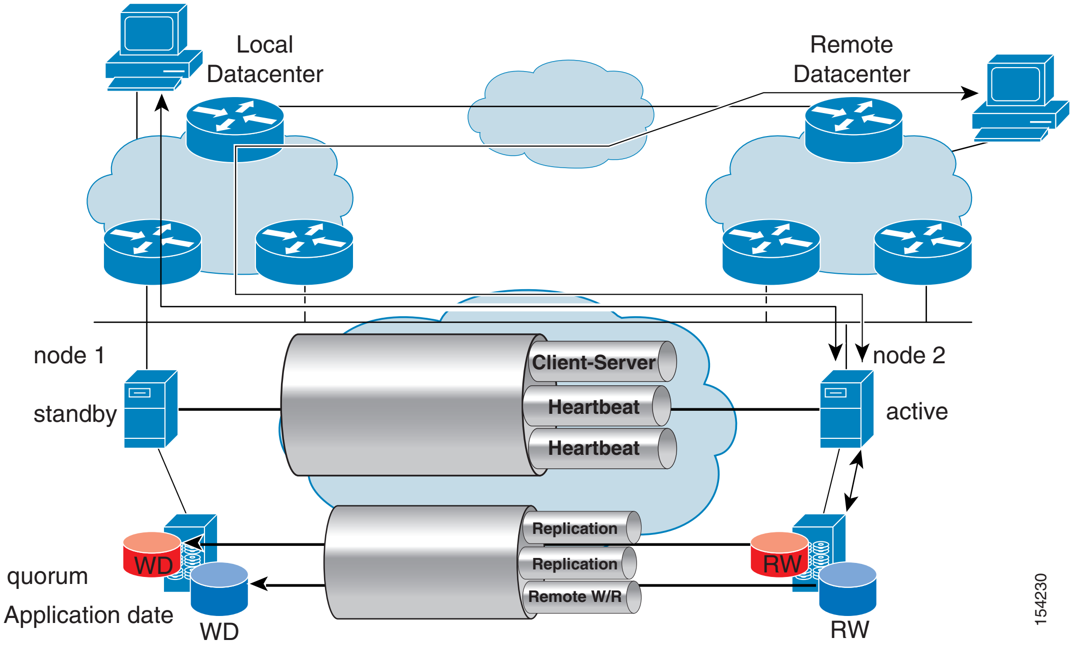
In sum, the test results show the following:

- The SAN extension transport bandwidth affects the maximum I/O throughput for the servers to the local disks (which are synchronously replicated).
- The maximum server I/O bandwidth changes with the distance, for reasons explained in the next section.

From a LAN point of view, the *cluster heartbeats* require a minimum amount of bandwidth because they are mainly used for the servers to monitor each other. The *client-to-server traffic* might instead have bigger needs for bandwidth when the servers are down in the primary site and processing continues from the secondary site. Figure 3-12 shows the components of the traffic on the LAN transport (heartbeats on redundant VLANs and potentially client-server access) and the components of the traffic on the SAN transport (replication traffic on redundant fabrics and potentially write and read access from the remote node).

Also, if a link is not fast enough, outstanding I/Os might accumulate and the disk array might have to switch from asynchronous mode to synchronous mode, which in turn affects the application performance.

Figure 3-12 Bandwidth Requirement Considerations for an HA Cluster

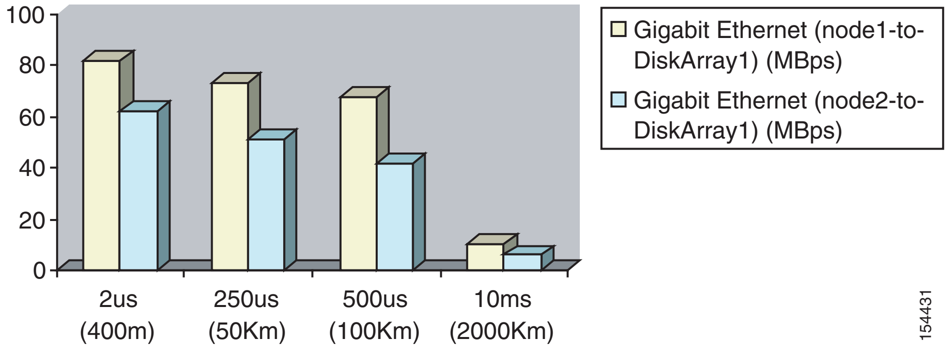


Distance Impact on the Application Throughput

Figure 3-13 shows the variation of the maximum throughput at increasing distances with 70 percent read, 30 percent writes, and the disk configured for synchronous replication. It also contrasts the achievable performance with the following scenarios:

- Writing from node1 to DiskArray1 with a Gigabit link between the sites
- Writing from node2 to DiskArray1 with a Gigabit link between the sites

Figure 3-13 Application Throughput Variation



As Figure 3-13 shows, the performance decreases with increasing distances, even if the total available transport bandwidth does not change. This is because with synchronous replication, the disk cannot acknowledge the write until it is replicated, so the disk array cannot have more than one outstanding I/O. The application in the test issues multiple SCSI writes concurrently (32 in the Cisco test bed) and they are concurrently replicated by the disk array. Depending on the frame size (64 KB), the number of applications outstanding I/Os (32 in this example), and the bandwidth, it may very well be that the throughput decreases with the distance. For example, with a 32 KB record size, 10 ms of latency, and 1 Gbps of transport bandwidth, it takes ~87 outstanding I/Os to keep the link fully utilized. With increasing distances, it takes more time to acknowledge the writes, and as a result the maximum throughput.

Cisco Fibre Channel Write Acceleration (FC-WA) can help increase the application performance. FC-WA increases replication or write I/O throughput and reduces I/O response time in most situations, particularly as the FCIP RTT increases. Each FCIP link can be “filled” with a number of concurrent or outstanding I/Os. Using the previous example, with 10 ms of latency, it takes 45 outstanding I/Os instead of 87 to keep the Gigabit transport fully utilized if FC-WA is enabled.

Note For more information, see *Designing FCIP SAN Extension for Cisco SAN Environments* at the following URL:
http://www.cisco.com/en/US/solutions/ns340/ns517/ns224/ns378/net_design_guidance0900aecd800ed145.pdf.

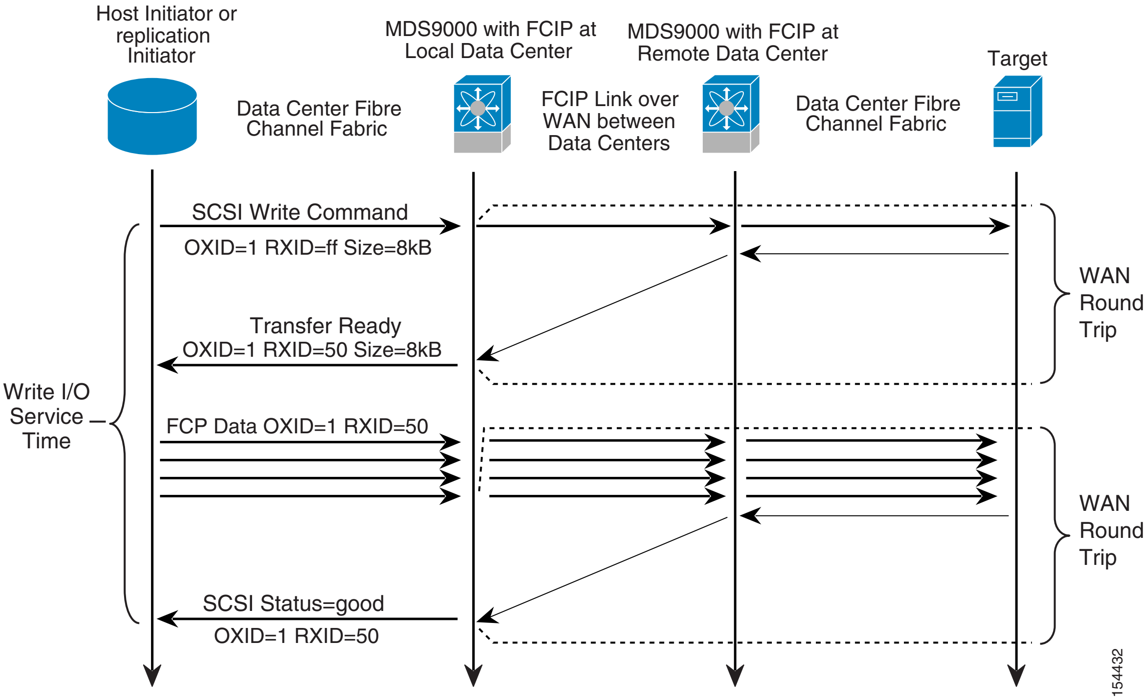
Benefits of Cisco FC-WA

Cisco FC-WA is a configurable feature introduced in Cisco SAN-OS 1.3 that you can use for FCIP SAN extension with the IP Storage Services Module. It is a SCSI protocol spoofing mechanism designed to improve application performance by reducing the overall service time for SCSI write I/Os and replicated write I/Os over distance.

FC-WA can help optimize the application throughput as described in Distance Impact on the Application Throughput, page 3-12, and it can be very beneficial in the presence of a host initiator to remote pooled storage when a node accesses a disk array in another site or data center, such as the case of campus and some metro clusters.

FC-WA reduces the number of FCIP WAN round trips per SCSI Fibre Channel Protocol (FCP) write I/O to one. Most SCSI FCP Write I/O exchanges consist of two or more round trips between the host initiator and the target array or tape. An example showing an 8 KB SCSI FCP write exchange, which is typical of an online transaction processing database application, is shown in Figure 3-14.

Figure 3-14 Short SCSI Write Exchange Without FC-WA (Two WAN Round Trips)

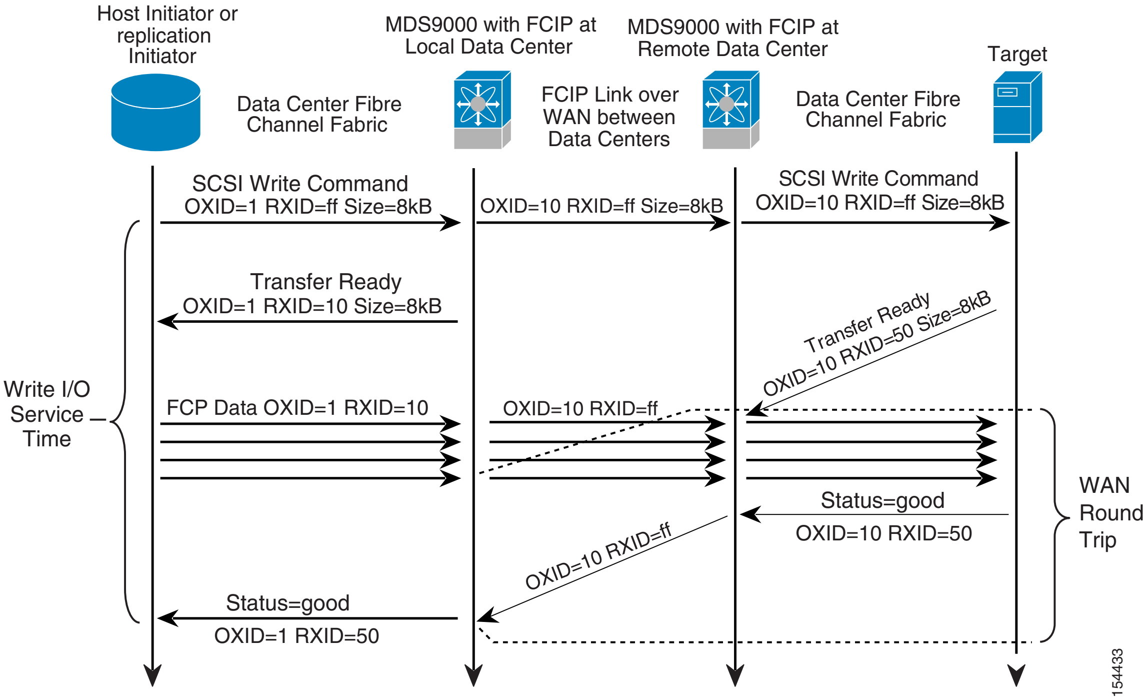

Host Initiator or replication Initiator
MDS9000 with FCIP at Local Data Center
MDS9000 with FCIP at Remote Data Center
Target
Data Center Fibre Channel Fabric
FCIP Link over WAN between Data Centers
Data Center Fibre Channel Fabric
SCSI Write Command
OXID=1 RXID=ff Size=8kB
Transfer Ready
OXID=1 RXID=50 Size=8kB
FCP Data OXID=1 RXID=50
SCSI Status=good
OXID=1 RXID=50
Write I/O Service Time
WAN Round Trip
WAN Round Trip
154432

The protocol for a normal SCSI FCP Write without FC-WA is as follows:

1. Host initiator issues a SCSI write command, which includes the total size of the write (8 KB, in this example), and also issues an origin exchange identifier (OXID) and a receiver exchange identifier (RXID).
2. The target responds with an FCP Transfer Ready message. This tell the initiator how much data the target is willing to receive in the next write sequence, and also tells the initiator the value the target has assigned for the RXID (50, in this example).
3. The initiator sends FCP data frames up to the amount specified in the previous Transfer Ready message.
4. The target responds with a SCSI status = good frame if the I/O completed successfully.

An example of short SCSI write exchange using FC-WA is shown in Figure 3-15.

Figure 3-15 Short SCSI Write Exchange Using FC-WA (Single WAN Round Trip)



The protocol for FC-WA differs as follows:

1. After the initiator issues a SCSI FCP Write, a Transfer Ready message is immediately returned to the initiator by the Cisco MDS 9000. This Transfer Ready contains a locally-allocated RXID.
2. At the remote end, the target, which has no knowledge of FC-WA, responds with a Transfer Ready message. The RXID of this is retained in a local table.
3. When the FCP data frames arrive at the remote MDS 9000 from the initiator, the RXID values in each frame are replaced according to the local table.

The RXID for the SCSI status = good frame is replaced at the local MDS 9000 with the “made up” value assigned in Step 1.

The expected performance gains when using FC-WA with synchronous applications with a single outstanding I/O is shown in Figure 3-16, Figure 3-17, Figure 3-18, and Figure 3-19. Each pair of graphs shows the following:

- *I/O Response* (in milliseconds) with FC-WA on (enabled) and off (disabled)—The IOPS is inversely proportional to the I/O response (IOPS = 1000 / I/O response). Throughput is calculated by multiplying the write size by IOPS.

Note This is per data stream sharing the FCIP link.

- *FC-WA Ratio*—This is the ratio of I/O response (inverse) and IOPS with FC-WA enabled versus disabled. This is graphed against RTT with write sizes of 4 KB, 8 KB, 16 KB, and 32 KB.

Figure 3-16 FC-WA I/O Response and Ratio at 45 Mbps (Single Outstanding I/O)—Approximate

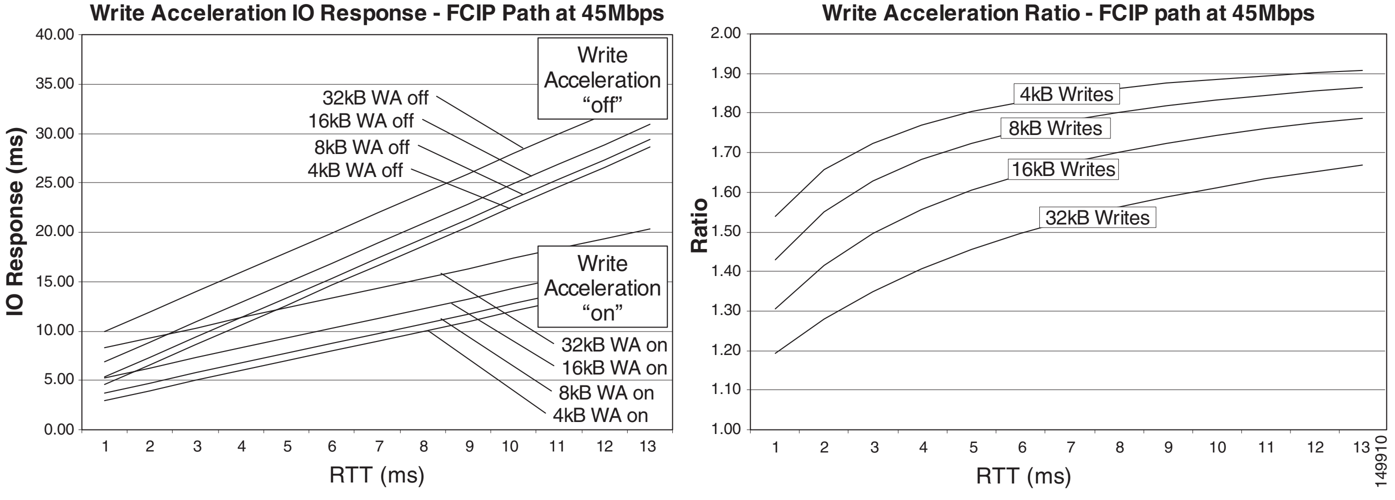


Figure 3-17 FC-WA I/O Response and Ratio at 155 Mbps (Single Outstanding I/O)—Approximate

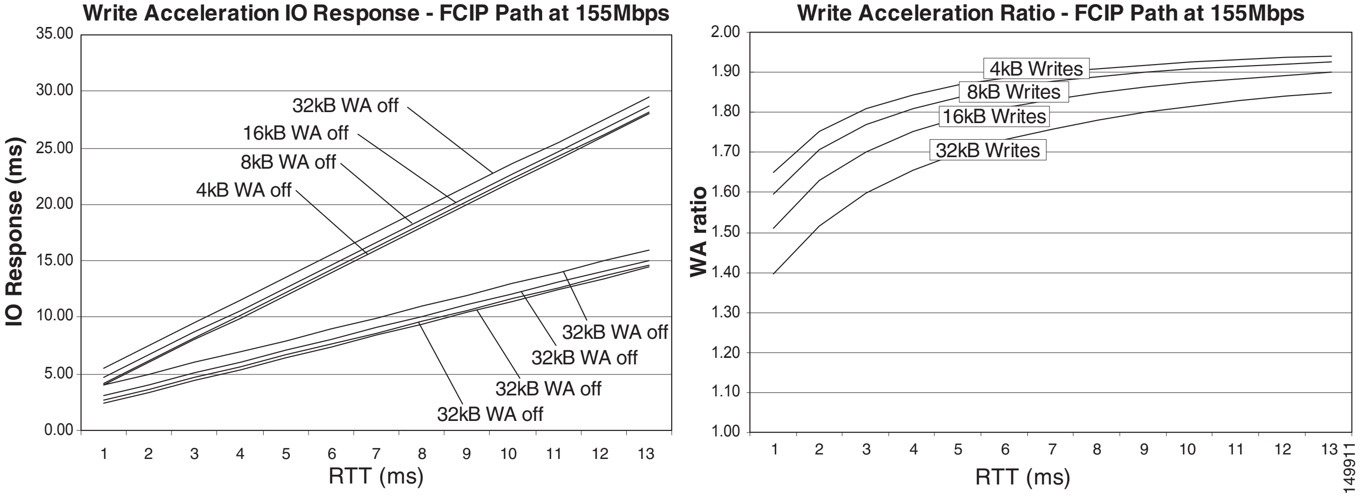


Figure 3-18 FC-WA I/O Response and Ratio at 622 Mbps (Single Outstanding I/O)—Approximate

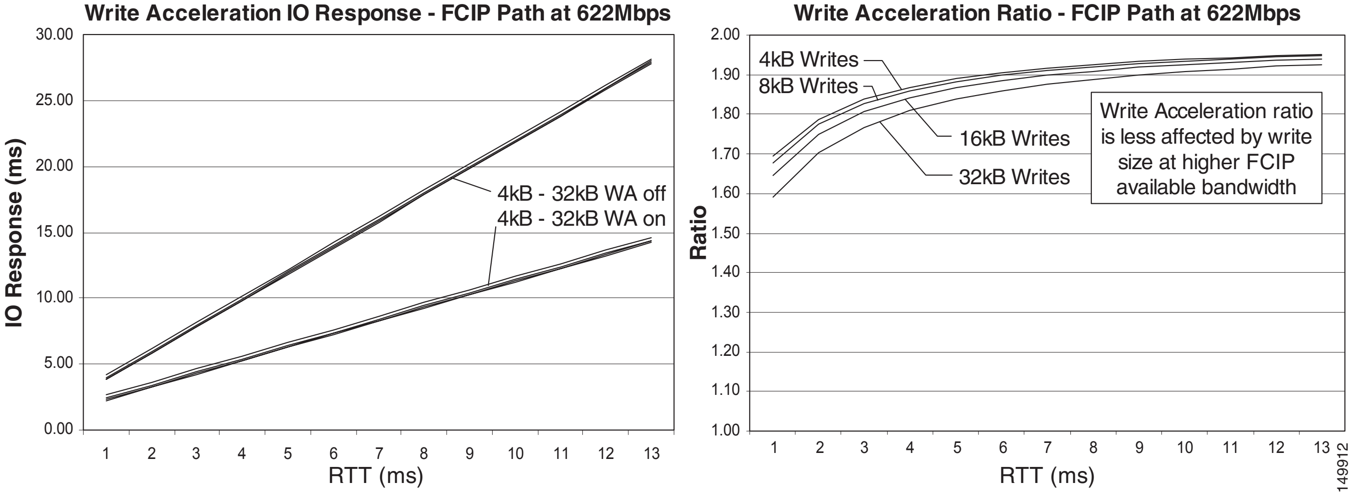
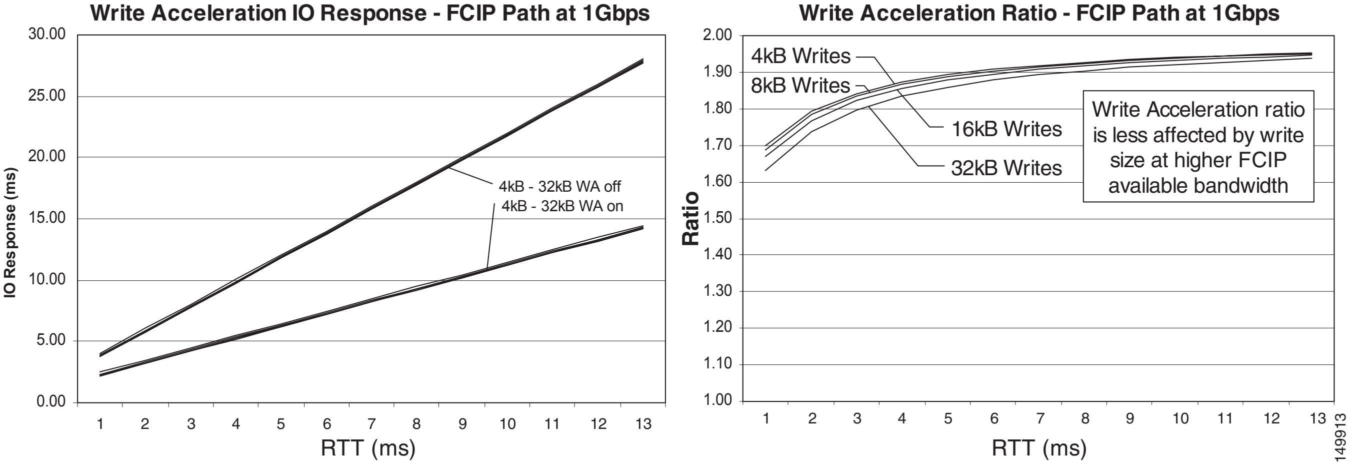


Figure 3-19 FC-WA I/O Response and Ratio at 1 Gbps (Single Outstanding I/O)—Approximate



To enable FC-WA simply apply the following configuration to both ends of the FCIP tunnel:

```
int fcip 1
   write-accelerator
```

Distance Impact on the Application IOPS

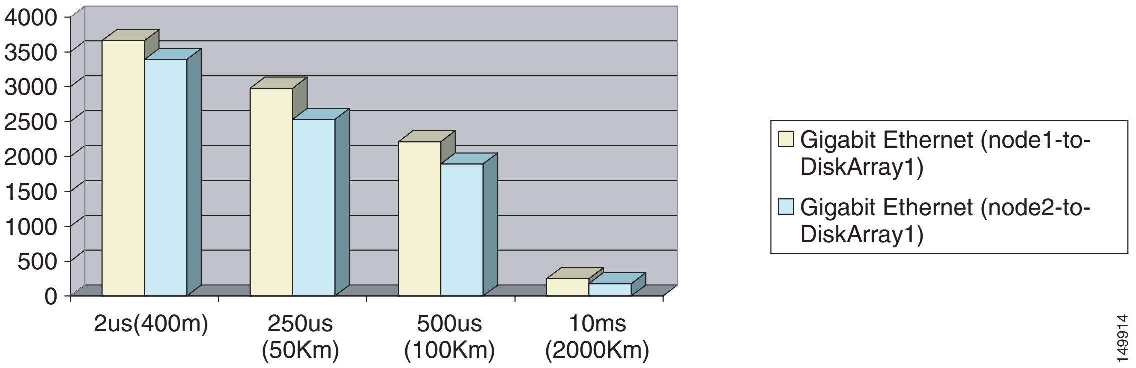
The distance between data centers decides what type of replication can be implemented on the disk arrays, but also constrains the performance of the local server to a certain level of response time (and IOPS as a result). Based on the distance, you may not be able to operate the application from the remote node when this node writes to the disk in the primary data center. The main factor that impacts the maximum IOPS performance is the response time, which for the most part is a function of 2 x RTT (4 x latency).

Figure 3-20 shows the variation of the maximum throughput at increasing distances with a R/W ratio of 70 percent read and 30 percent write, record size of 512 B, and disks configured for synchronous replication.

The achievable performance is contrasted with the following scenarios:

- Writing from node1 to DiskArray1 with a Gigabit link between the sites
- Writing from node2 to DiskArray1 with a Gigabit link between the sites

Figure 3-20 Variation of Maximum IOPS with the Distance (Synchronous Replication)



Consider that without any replication, the configuration can yield ~8000 IOPS. With replication in place and a Gigabit interconnect between the sites, you can achieve ~3500 IOPS at 400 m distance. This goes down to ~2200 IOPS at 100 km, which is a 37 percent performance penalty.

Note Note that the IOPS performance largely depends on the write IOPS performance (because the writes need to be replicated in a synchronous replication configuration). The write IOPS performance is proportional to 1/(write response time) where the response time in turn depends on 2 x RTT (4 x latency). At 500 us (100 km), a write response time of ~11 ms was measured, so theoretically the write IOPS should reach its maximum at ~90 write IOPS. Considering that the application is generating 32 outstanding I/Os and that the writes contribute to 30 percent of the combined performance, this yields ~864 write IOPS. As the graphs show, the total IOPS measured were 2200, of which ~730 were write IOPS.

The performance approaches zero for distances of thousands of kilometers. For specific cases, it still may be possible to operate some disks at thousands of kilometers if the application can tolerate a few 10s of IOPS. For example, this may be the case for the quorum disk.

Note Be sure to verify with your clustering software vendor and the storage vendor whether this configuration is supported.

Figure 3-20 also shows that it is possible to operate an application from a remote node writing to the disk in the main site at a decreased performance. The performance penalty is around ~10 percent.

This section shows in summary the following:

- The local and remote server maximum IOPS performance depends on the distance between the sites; that is, on the response time.
- Operating the application from a remote node writing to the disk in the primary site may be feasible, depending on the distance and the acceptable performance penalty.

- FC-WA can help increasing the application IOPS.

Asynchronous Versus Synchronous Replication

When data centers are further apart than 100 km, using synchronous replication causes a significant performance penalty on the local servers, as shown in Figure 3-13. Asynchronous replication addresses this problem, with the well-known drawback that a disk failure in the primary site can cause loss of data, which may not be compatible with the RPO.

For example, compare the performance of node1 writing to DiskArray1 with 70 percent read and 30 percent writes, when the distance between the two data centers is 2000 km:

- Synchronous replication—66 maximum IOPS (@512 B), average response time 299 ms
- Asynchronous replication—5984 IOPS, average response time 3 ms

Considering the performance improvement on the local node, you might wonder whether it would be feasible to operate the application from the remote node while still writing to the local disk. The performance of node2 writing to DiskArray1 with 70 percent read and 30 percent writes with a distance of 2000 km is 142 Maximum IOPS (@512 B), average response time 140 ms.

Note Write IOPS performance is proportional to 1/(write response time) where the response time in turn depends on 2 x RTT. At 2000 km, Cisco measured a write response time of ~299 ms, so theoretically the write IOPS should reach its maximum at ~3 write IOPS. Considering that the application is generating 32 outstanding I/Os, and that the write contributes to 30 percent of the total IOPS performance, this gives 28 write IOPS. The total IOPS measured were ~66, of which ~21 were write IOPS.

From a disk configuration point of view, it is important to monitor the status of the “RDF” groups (in EMC nomenclature) to verify that the replication mechanism is compatible with the distance and the performance requirements of the application. The following configuration samples show what to look for, and Cisco highly recommends that you talk to your storage vendor to verify your design choices.

If the disks are configured for synchronous replication, you should see the following output where the RDF pair shows as *Synchronized*, the field “MDA” shows an S for synchronous, and there are no *Invalid* tracks. Also notice that the disks in site1 are in *RW* (read-write) status, while the disks in site2 are in *WD* (write disabled) status.

```
C:\>symrdf -g HA1 query


Device Group (DG) Name             : HA1
DG's Type                          : RDF1
DG's Symmetrix ID                  : 000187431320


       Source (R1) View                 Target (R2) View     MODES
--------------------------------    ------------------------ ----- ------------
             ST                  LI      ST
Standard      A                   N       A
Logical       T  R1 Inv   R2 Inv  K       T  R1 Inv   R2 Inv       RDF Pair
Device  Dev   E  Tracks   Tracks  S Dev   E  Tracks   Tracks MDA   STATE
-------------------------------- -- ------------------------ ----- ------------

DEV001  00B9 RW       0        0 RW 00B9 WD       0        0 S..   Synchronized

Total          -------- --------           -------- --------
  MB(s)             0.0      0.0                0.0      0.0
```

```
Legend for MODES:

 M(ode of Operation): A = Async, S = Sync, E = Semi-sync, C = Adaptive Copy
 D(omino)           : X = Enabled, . = Disabled
 A(daptive Copy)    : D = Disk Mode, W = WP Mode, . = ACp off
```

If the disks are configured for asynchronous replication, you should see the following output where the RDF pair shows as *Consistent*, the field “MDA” shows an A for asynchronous, and there are no *Invalid* tracks. Also note that the disks in site1 are in RW (read-write) status, while the disks in site2 are in WD (write disabled) status.

```
C:\>symrdf -g HA1 query


Device Group (DG) Name             : HA1
DG's Type                          : RDF1
DG's Symmetrix ID                  : 000187431320


       Source (R1) View                 Target (R2) View     MODES
--------------------------------    ------------------------ ----- ------------
             ST                  LI      ST
Standard      A                   N       A
Logical       T  R1 Inv   R2 Inv  K       T  R1 Inv   R2 Inv       RDF Pair
Device  Dev   E  Tracks   Tracks  S Dev   E  Tracks   Tracks MDA   STATE
-------------------------------- -- ------------------------ ----- ------------

DEV001  00B9 RW       0        0 RW 00B9 WD       0        0 A..   Consistent

Total          -------- --------           -------- --------
  MB(s)             0.0      0.0                0.0      0.0
```

If for any reason the transport pipe is not fast enough for the rate at which the application writes to the local disk, you see that the links connecting the two sites are being used even if the application is not writing to the disk, because the disks buffer the writes and keep replicating to the remote site until the tracks are all consistent.

If for any reason there are Invalid tracks, you can force the disks to synchronize by issuing the command **symrdf –g** <*group name*> **establish**. This initiates an RDF “Incremental Establish” whose state can be monitored via the command **symrdf –g HA1 query**. A “SynchInProg” status message then appears.

```
Device Group (DG) Name             : HA1
DG's Type                          : RDF1
DG's Symmetrix ID                  : 000187431320


       Source (R1) View                 Target (R2) View     MODES
--------------------------------    ------------------------ ----- ------------
             ST                  LI      ST
Standard      A                   N       A
Logical       T  R1 Inv   R2 Inv  K       T  R1 Inv   R2 Inv       RDF Pair
Device  Dev   E  Tracks   Tracks  S Dev   E  Tracks   Tracks MDA   STATE
-------------------------------- -- ------------------------ ----- ------------

DEV001  00B9 RW       0     1483 RW 00B9 WD       0        0 A..   SyncInProg

Total          -------- --------           -------- --------
  MB(s)             0.0      0.0                0.0      0.0
```

Read/Write Ratio

The performance of the application depends not only on the distance between the sites but on the Read/Write (R/W) ratio characteristics of the application itself. The following shows the difference in I/O throughput measured on the host when the disks are configured for synchronous replication on an OC-3 transport, with a distance between ~100 and 200 km:

- Maximum throughput (@64 KB record size)—30 percent write and 70 percent read is 15 Mbps; 70 percent write and 30 percent read is 11 Mbps
- Maximum IOPS (@512 KB record size)—30 percent write and 70 percent read is ~1040 IOPS; 70 percent write and 30 percent read is ~544 IOPS.

Note that with 70 percent write and 30 percent read, the write throughput is ~5 Mbps, which is the same maximum write throughput as with the 30 percent write and 70 percent read; however, the combined throughput with 30 percent write and 70 percent read is higher. This indicates that it is likely that the OC3 connection between the sites is using a single STS-1 (~51 Mbps).

As for the maximum, write IOPS is ~360 for the 30 percent write and 70 percent read configuration, and ~380 for the 70 percent write and 30 percent read configuration. This shows that with 70 percent write and 30 percent read, the write IOPS goes up as expected, because the maximum write IOPS performance is proportional to 1/(response time), where the response time is in turn proportional to 2 x RTT.

Transport Topologies

Two Sites

The simplest topologies involve only two data centers and resemble Figure 3-21. One of the main design criteria is high availability to reduce the possibility that the LAN or the SAN becomes segmented.

Aggregating or Separating SAN and LAN Transport

You can carry LAN and SAN on different paths, or on the same path but different lambdas or STSs, or you can carry them on the same IP path (by using FCIP). Each approach has pros and cons, as follows:

- SAN and LAN on different paths—This approach may be the most expensive one, but it has the advantage that a failure on the LAN connectivity does not cause a split-brain scenario because the nodes can still arbitrate the ownership of the quorum disk (if the quorum disk approach is used for arbitration). Operations then continue from the site where the node owns the quorum. A failure on the SAN connectivity prevents a failover, but operations can continue while Invalid tracks accumulate.
- SAN and LAN on the same transport, but on different lambdas/STSs/pseudowires—With this type of design, LAN and SAN are using the same physical transport but on different lambdas. This means that, for example, a broadcast storm or a spanning tree loop on the LAN does not affect the SAN traffic. On the other hand, there is the unlikely possibility that both LAN and SAN become partitioned. To the cluster software, this appears to be a complete failure of the other site. The normal policy is to not bring any new resource online, and optionally you can configure the cluster to also stop the resources that were previously online. Consider that from a routing point of view, you can ensure that the traffic goes where the resources are supposed to be online (that is, where the quorum is normally met). Note also that it is very unlikely that the two sites become completely partitioned if the optical/SONET design provides some protection mechanism.

- SAN and LAN on the same transport/pseudowire/lambda—This may be the most cost-efficient option, which has the intrinsic disadvantage, as compared with the other options, that LAN and SAN are more likely become partitioned at the same time. It is also possible that a broadcast storm or spanning tree reconvergence on the LAN could affect the SAN. This is still a valid design when redundancy is correctly implemented.

Common Topologies

For most current HA clusters, you need to provision a Layer 2 path between the data centers. HA and Layer 2 means that you need spanning tree to keep the topology free from loops. Depending on the technology used to interconnect the sites, you may be able to create an EtherChannel between the sites, which allows Layer 2 extension without the risk of Layer 2 loops. The topologies to connect two sites with Layer 2 extension can be built starting from the following three basic modes:

- Square spanning tree—This topology can have one or two switches at each site; all the switches used for cluster extension can be part of the same spanning tree domain.
- EtherChannel—Each switch connects to the remote switch via an EtherChannel across lambdas/circuits/pseudowires (notice that the EtherChannel is end-to-end between A1 and A2).
- Cross-stack EtherChannel—Two switches at each site are clustered to look like a single switch and in turn they connect via the pseudowires to the remote site with a cross-stack EtherChannel.

CWDM and DWDM Topologies

Figure 3-21 shows the Layer 2 extension topologies with CWDM.

Figure 3-21 Layer 2 Extension Topologies with CWDM

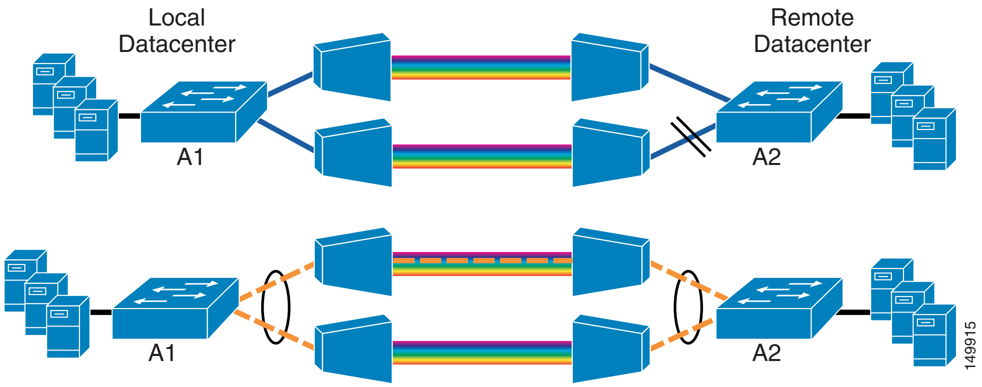
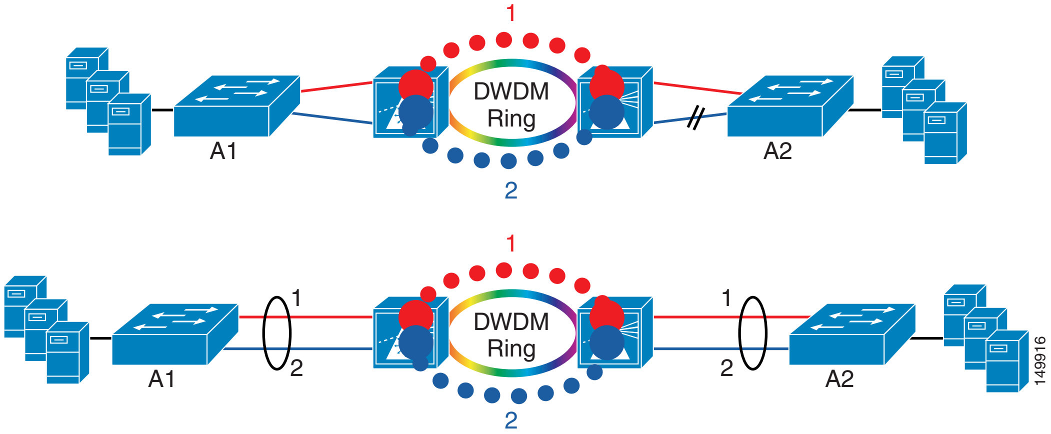


Figure 3-22 shows the Layer 2 extension topologies with DWDM. The lambdas for each port in the channel can be configured to use a different physical route on the ring. You can choose between using client protection (that is, using the port channel protection) or adding a further level of protection, such as splitter protection or Y-cable protection. Note that the EtherChannels are end-to-end, which provides verification of the Layer 1 path.

Figure 3-22 Layer 2 Extension Topologies with DWDM

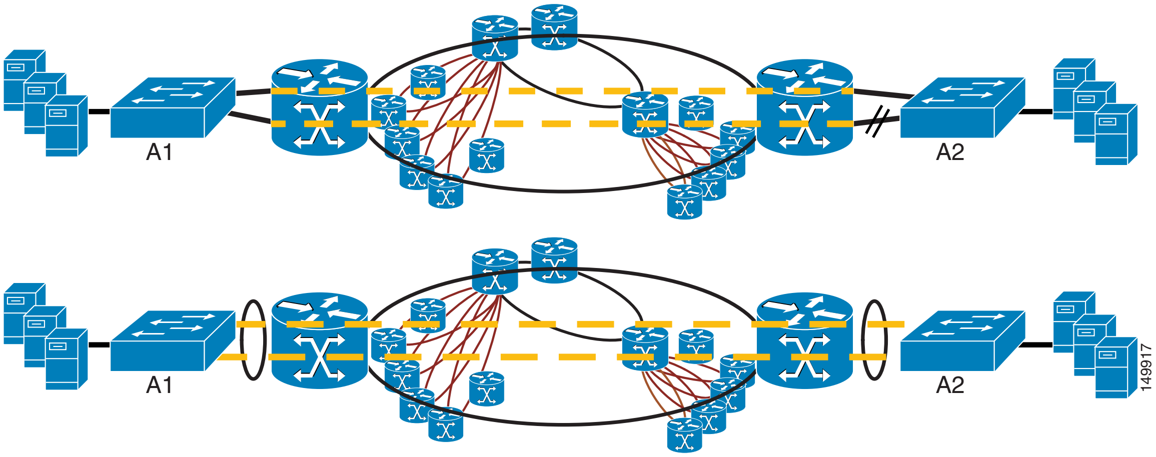


SONET Topologies

Figure 3-23 shows the Layer 2 extension topology with SONET. Note that with SONET, you can use the Layer1 option (that is, the G-series card) where each circuit is a Layer 1 link (SONET-protected or -unprotected) with client-side protection (port channel or spanning tree on the access switch). The EtherChannels are end-to-end and provide verification of the channel.

Note The SONET topologies in Figure 3-23 show one SONET ring provided by a service provider and, at regional or continental distances, it is very likely that many rings are present between the two sites. Eventually, this does not matter, because all you are buying is two circuits, protected or unprotected.

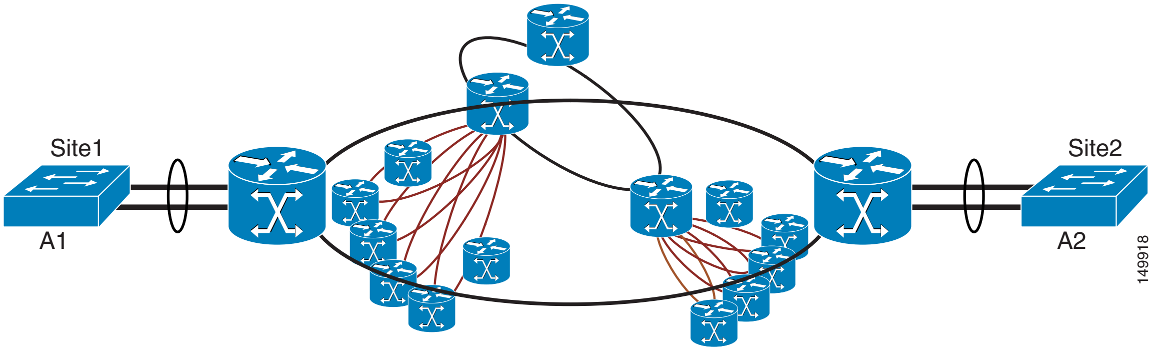
Figure 3-23 Layer 2 Extension Topologies with SONET (Layer 1 Circuits)



With SONET, you can also terminate the EtherChannel on the SONET gear itself and have a shared Layer 2 ring free from loops by using Spatial Reuse Protocol/Resilient Packet Ring (SRP/RPR). Figure 3-24 shows this option.

Note The SRP/RPR topology shows a SONET ring. SRP/RPR is really a property of the line card in the SONET gear that connects the customer to the SONET infrastructure, not a property of the SONET infrastructure itself. Also, the ring is not a physical SONET ring, but a logical ring of circuits that connect the customer over a SONET infrastructure.

Figure 3-24 Layer 2 Extension Topologies with SONET (SRP/RPR)



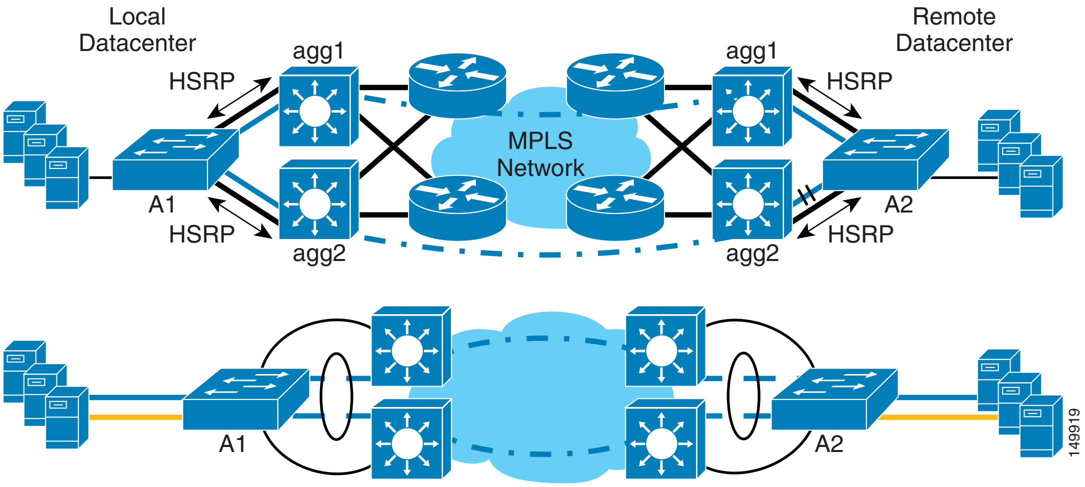
Multiprotocol Label Switching Topologies

If you have a Multiprotocol Label Switching (MPLS) infrastructure, you can also build Ethernet over MPLS (EoMPLS) tunnels to carry Layer 2 traffic on top of a Layer 3 network. Figure 3-25 shows the spanning tree and EtherChannel models applied to an MPLS network. Port-based cross-connect allows running port channeling end-to-end, which provides verification of the path. MPLS ensures fast convergence and high availability of the EoMPLS pseudowires.

When deploying an MPLS-based solution, realize that local switching on the MPLS "PE" device may not be possible for the interface that is tunneled, which is why Figure 3-25 displays an access switch connected to the PE switch (agg1 and agg2). If VLAN-based cross-connect is supported, local switching may be possible; if port-based cross-connect is used, you need to provide an access switch to support this function. Figure 3-25 shows a port-based cross-connect.

Note For more information about LAN extension over MPLS, see Chapter 4, "FCIP over IP/MPLS Core."

Figure 3-25 Layer 2 Extension Topologies with SONET (SRP/RPR)



SAN extension in the presence of an MPLS core can leverage FCIP. For more information, see Chapter 4, "FCIP over IP/MPLS Core."

When buying a metro Ethernet offering from a service provider, the underlying technology can belong to any of these scenarios: a lambda, a circuit, or an EoMPLS tunnel.

Three or More Sites

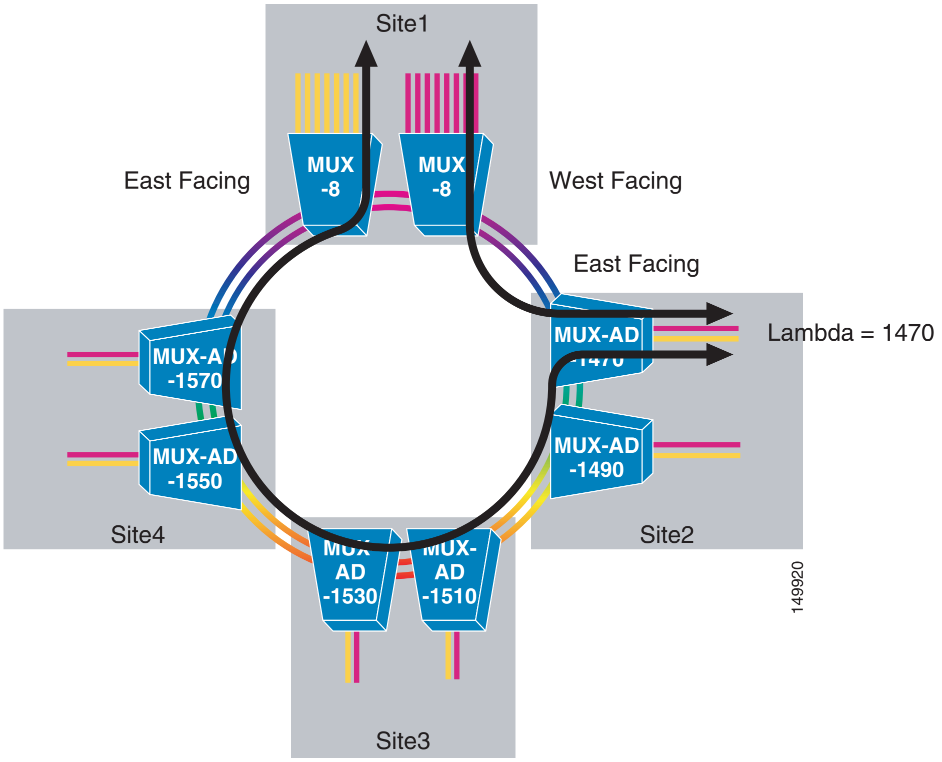
With three or more sites, the main topologies are hub-and-spoke and ring. The ring topology may match a physical topology, or may be just the topology of a specific lambda/circuit/pseudowire. The hub-and-spoke is typically a "logic" topology carried on top of one or several rings, or any multi-hop routed topology. SONET with SRP/RPR offers an interesting alternative by means of a shared RPR ring, which works as a bus between all the sites without the need to carry spanning tree or EtherChanneling across sites.

Hub-and-Spoke and Ring Topologies with CWDM

Figure 3-26 shows a CWDM ring topology.

Note For distances and further configuration details, consult an optical specialist.

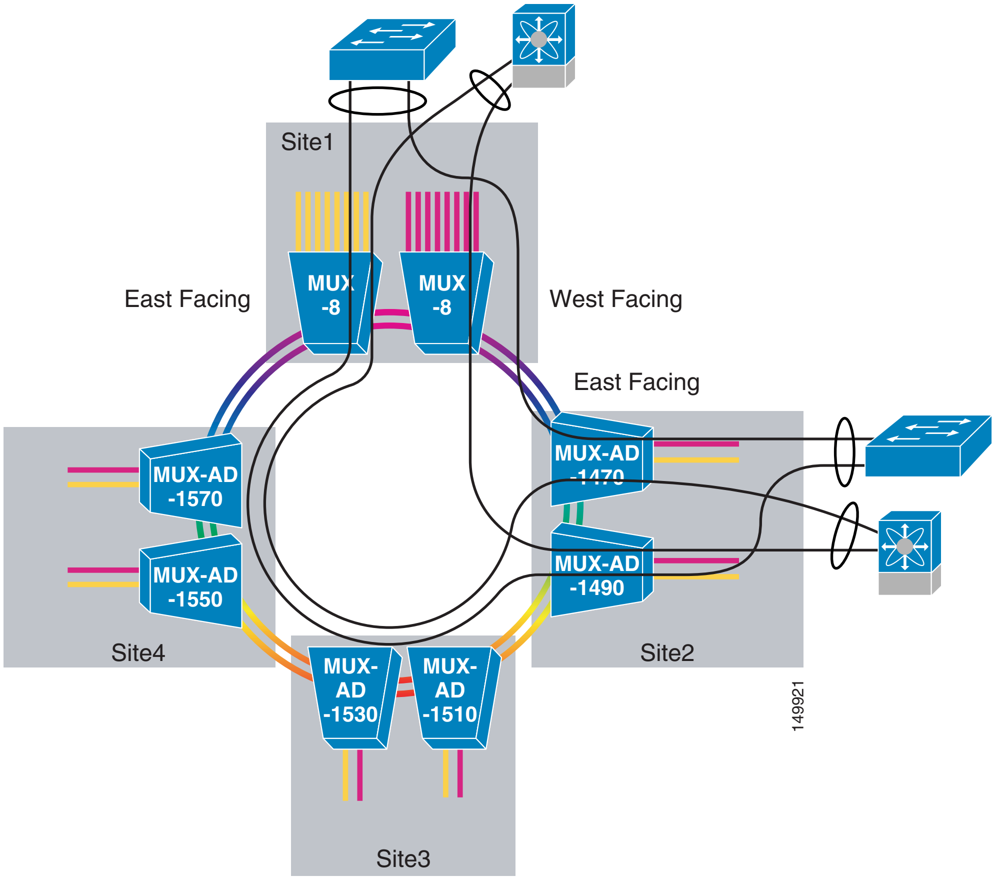
Figure 3-26 CWDM Ring Topology for Hub-and-Spoke Deployments



This topology provides redundant multiplexers (muxes) at each site (1 lambda 2 channels), and each mux is in turn connected to both the muxes in the primary site. Figure 3-26 shows that the 1470 lambda enters the primary site on the West Facing mux, and is pulled from the ring at Site2. The same lambda is re-used from Site2 to Site1 via the West Facing path that terminates on the East Facing mux at Site1. This basically provides two "physical" links for an access switch to Site1.

The 1470-MUX at site2 is a single point of failure. For additional redundancy, there is a 1490-MUX that pulls out lambda 1470 for a point-to-point "physical" link between Site2 and Site1 along the east path, and re-uses the 1470 lambda to connect Site2 to Site1 along the west path. This creates a fully redundant hub-and-spoke topology that can be used in several ways. One of them is shown in Figure 3-27, which shows a possible use of the protected CWDM ring. Each switch at each site connects to the primary site via two paths, a west and an east path, which are part of the same port channel. Note that each switch connects to both muxes, so each port in the port channel uses a different lambda.

Figure 3-27 CWDM Client-protected Ring to Transport SAN and LAN



The resulting topology is a hub-and-spoke topology (see Figure 3-28). The primary site aggregates traffic from all the sites with multiple port channels from the central switch to each site switch. The central switch is a single point of failure. The main benefit of this topology is that there is no need for spanning tree, so all links are forwarding.

Figure 3-28 Hub-and-Spoke Topology

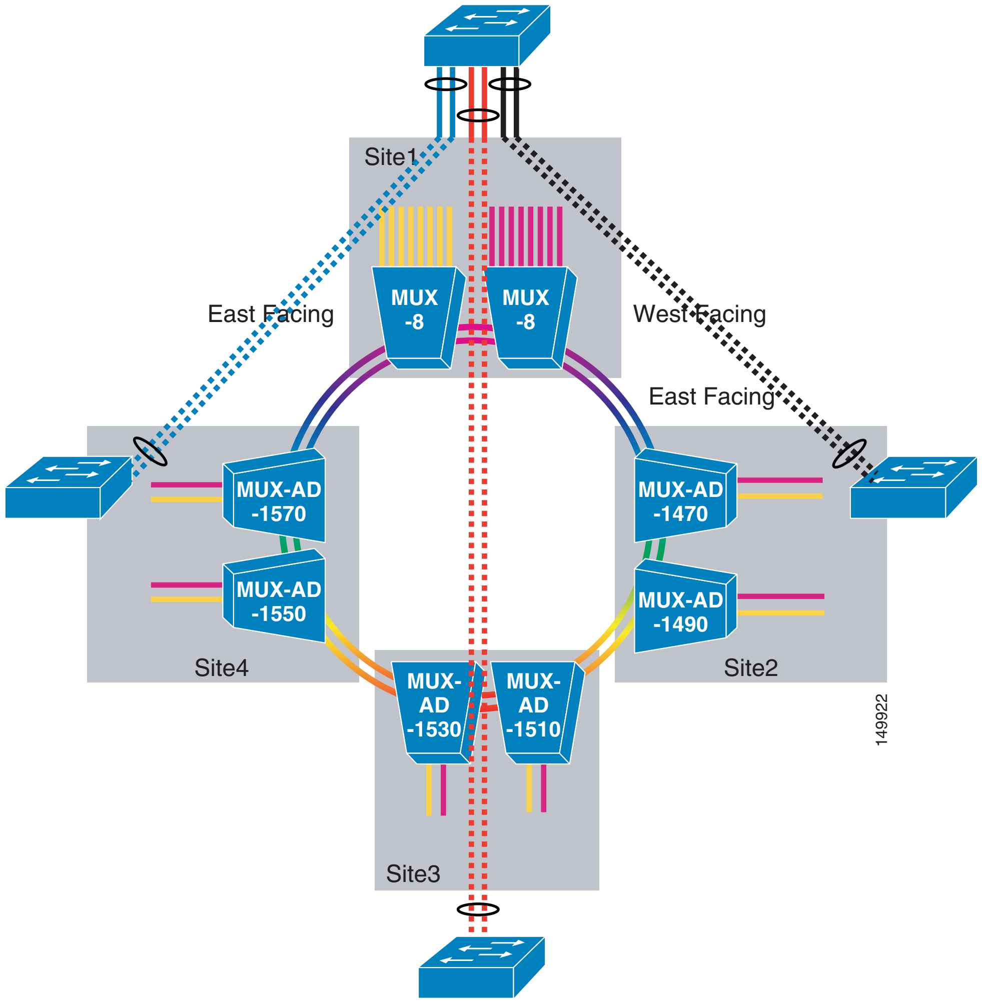
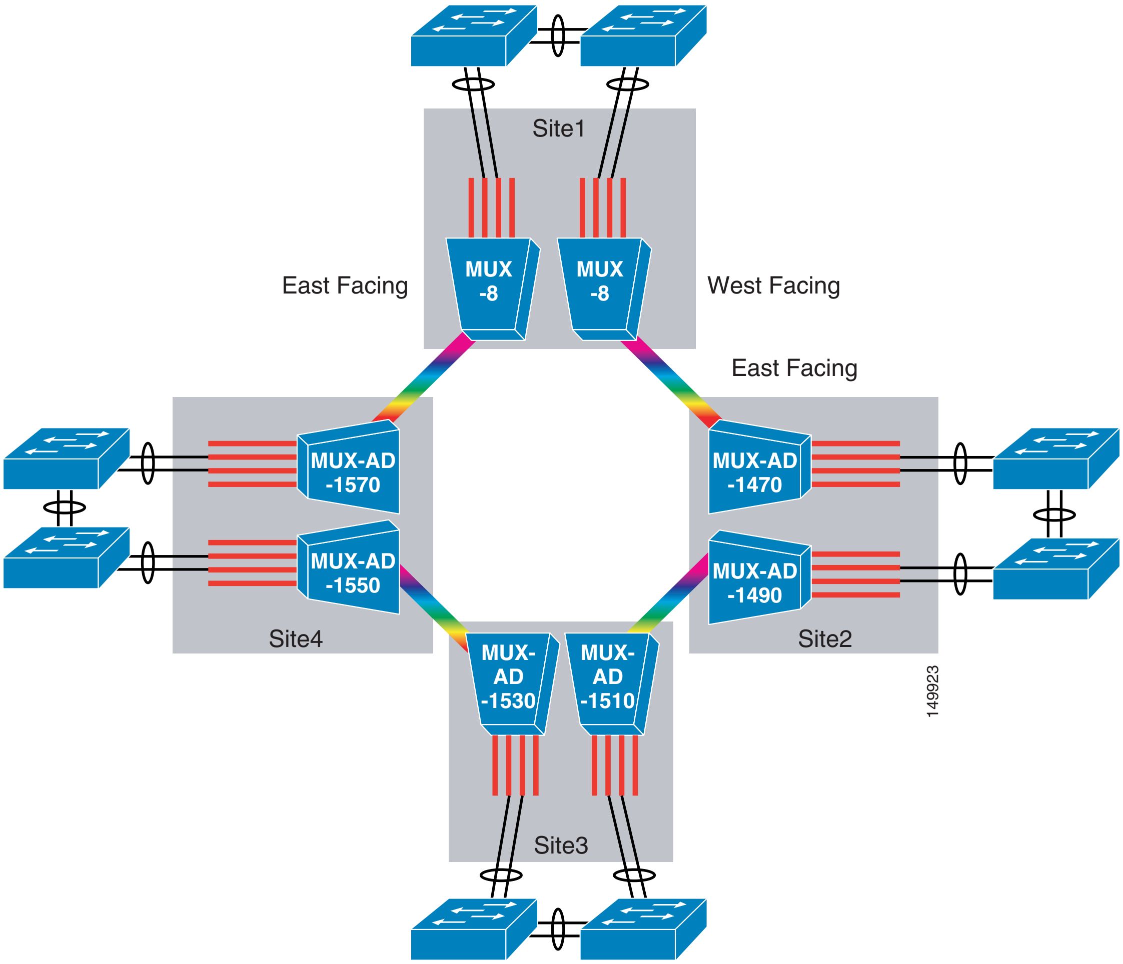


Figure 3-29 shows a topology with no single point of failure, which relies on spanning tree to provide a redundant Layer 2 path when one of the links fails. This topology is simply made of point-to-point links between the sites; these links are multiplexed with CWDM muxes.

Figure 3-29 Layer 2 Ring Topology



Hub-and-Spoke and Ring Topologies with DWDM

You can provision DWDM rings to provide a hub-and-spoke topology, such as the one shown in Figure 3-30.

Figure 3-30 DWDM Hub-and-Spoke Topology

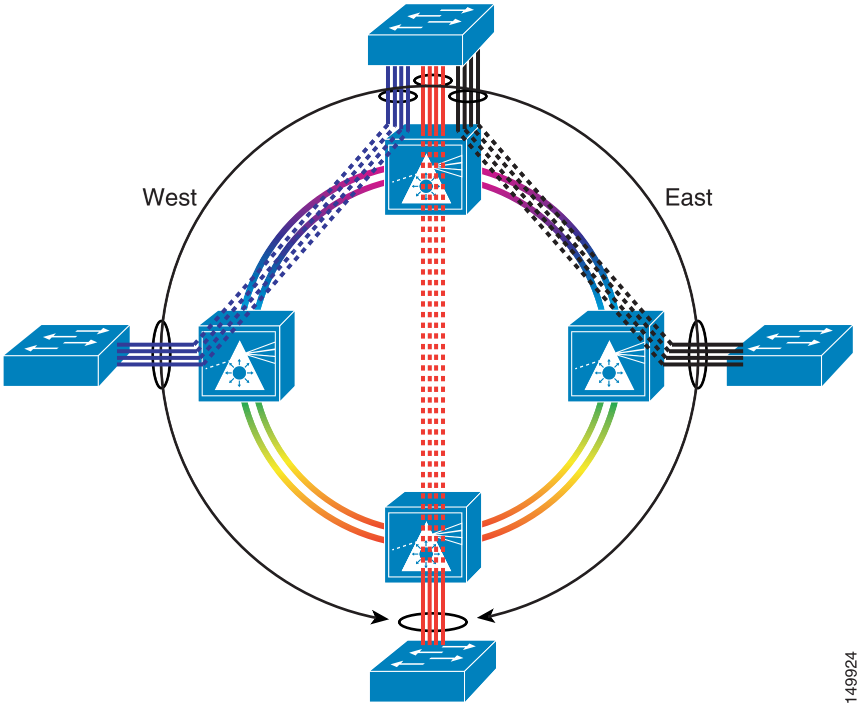
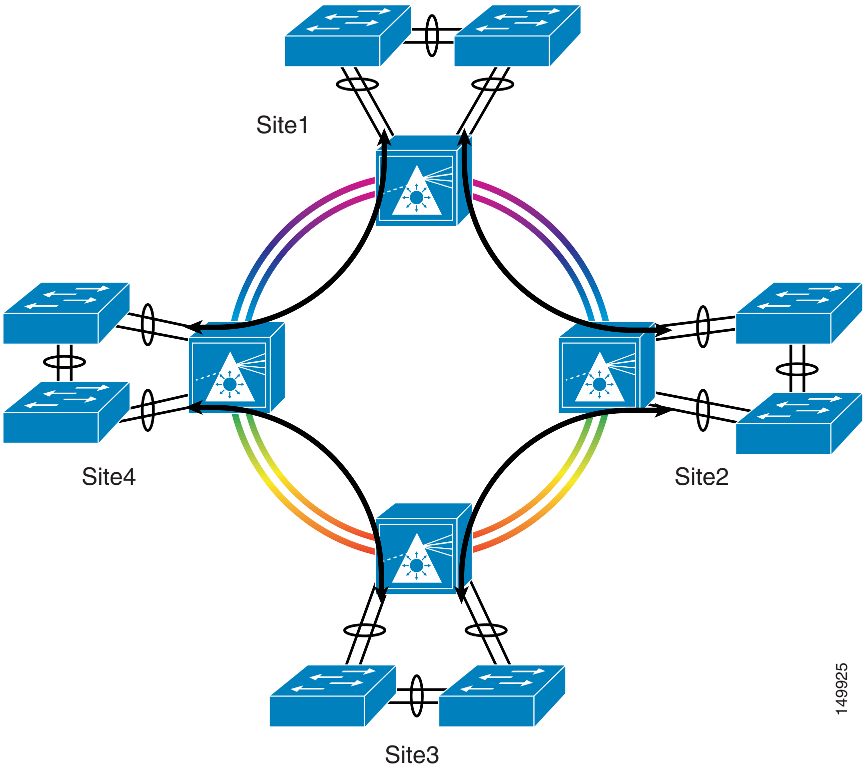


Figure 3-31 shows a ring topology with DWDM.

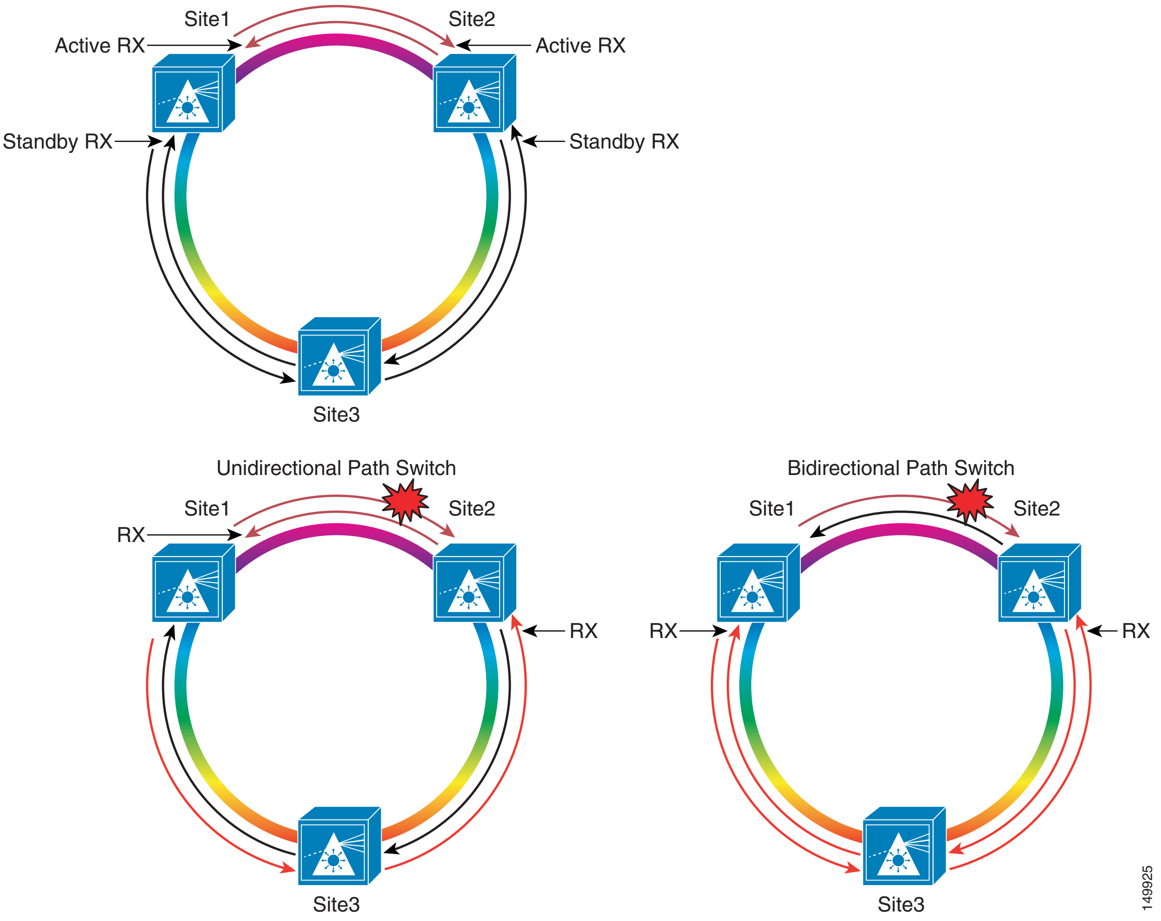
Differently from the CWDM topology, a DWDM ring can provide additional protection mechanisms than client protection.

Figure 3-31 DWDM Ring Topology



DWDM provides several HA features, such as splitter protection and Y-cable protection. In addition, failures can be recovered with unidirectional path switch or bidirectional path switch, as shown in Figure 3-32. It is out of the scope of this document to provide all the details of the optical design, but note that several options exist and the paths can be of different length. This may or may not be an issue to the “client” (that is, the Fibre Channel switch and the Ethernet switch), depending on the design details.

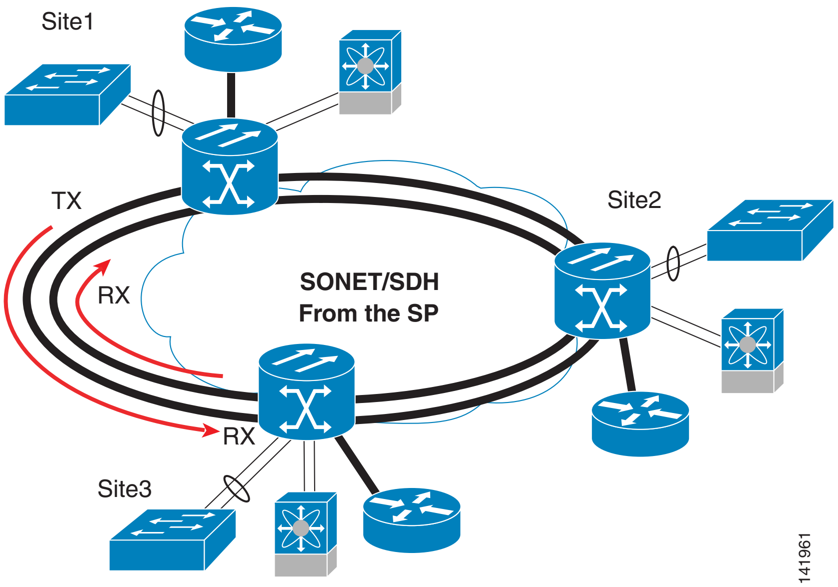
Figure 3-32 DWDM Protection Mechanisms



Shared Ring with SRP/RPR

Figure 3-33 shows the communication between Site1and Site3.

Figure 3-33 Use of SONET to Build an RPR Ring



This communication in traditional ring technologies involves the full ring. With SRP, bandwidth utilization is more efficient, because the destination strips off the frame from the ring (only multicast frames are stripped from the source). By using this mechanism, DPT rings provide packet-by-packet spatial reuse in which multiple segments can concurrently exchange traffic at full ring bandwidth without interference.

Another important aspect of the RPR operation is how the ring is selected. Site1 sends out an Address Resolution Protocol (ARP) request to a ring that is chosen based on a hash. Site3 responds to the ARP by examining the topology and choosing the ring with the shortest path. Site1 then uses the opposite ring to communicate with Site3. This ensures that the communication path is the shortest.

The SRP Fairness Algorithm (SRP-fa) ensures that both global fairness and local bandwidth optimization are delivered on all segments of the ring.

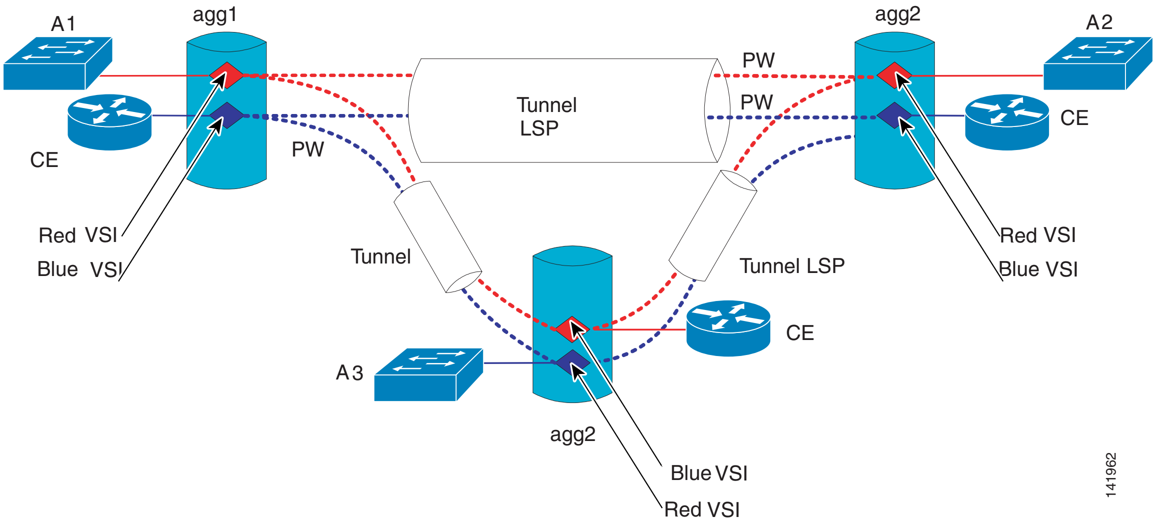
With SRP/RPR, the EtherChannels can be terminated locally on the Cisco ONS 15454 device. The ring appears to the Ethernet switches as a Layer 2 shared link. Neither spanning tree nor EtherChannel need to be carried across sites.

Virtual Private LAN Service

Virtual Private LAN Service (VPLS) is an architecture that provides Ethernet multipoint connectivity across geographically-dispersed locations using MPLS as a transport. VPLS is often used by SPs to provide Ethernet Multipoint Services (EMS), and can be used by enterprises on a self-managed MPLS-based metropolitan area network (MAN) to provide high-speed any-to-any forwarding at Layer 2 without the need to rely on spanning tree to keep the topology free from loops. The MPLS core uses a full mesh of pseudowires and “split-horizon” forwarding to avoid loops.

Figure 3-34 shows the use of virtual switch instances (VSIs) using MPLS pseudowires to form an “emulated” Ethernet switch.

Figure 3-34 Use of VPLS to Connect Multiple Data Centers at Layer 2



Geocluster Design Models

Campus Cluster

Figure 3-35 and Figure 3-36 show two of several possible topologies for an extended cluster within a campus. These illustrations show a topology that uses CWDM on dark fiber for the SAN and LAN extension. On the SAN side, both the initiator disk access VSANs (referred to as FA VSANs) and the VSANs used for replication (RA VSANs) are extended on top of a port channel, which is also a TE port.

On the LAN side, the EtherChannel connects the two root switches for each server farm. Bridge Protocol Data Unit (BPDU)-filtering or another suitable technique to avoid mixing the VLAN Trunk Protocol (VTP) domains can be used. This LAN design allows the network administrator to pick the public and private VLANs from each server farm and connect them, regardless of which access switch they are on (the root switch sees all of them). If aggregation1 disappears, the LAN communication between the two nodes is lost; the quorum disk prevents a split-brain scenario. The routing configuration ensures that the traffic enters the server farm where the servers normally own the quorum disk.

An optimization to this design consists in connecting agg2s and changing the VLAN trunking configuration between the two server farms as follows:

- For example, Agg1s trunk VLAN 10 and 20
- Agg2s trunk VLAN 30

By doing this, you do not introduce additional loops between the two sites (which also allows you to do BPDU filtering between the sites), and in case aggregation1 is lost while the public LAN is segmented, the private LAN communication on VLAN 30 is not lost, and the routing configuration ensures that traffic enters into server farm1, where under normal conditions node1 owns the quorum.

Figure 3-35 Cluster Extension within a Campus (Built on Top of a Looped Access)

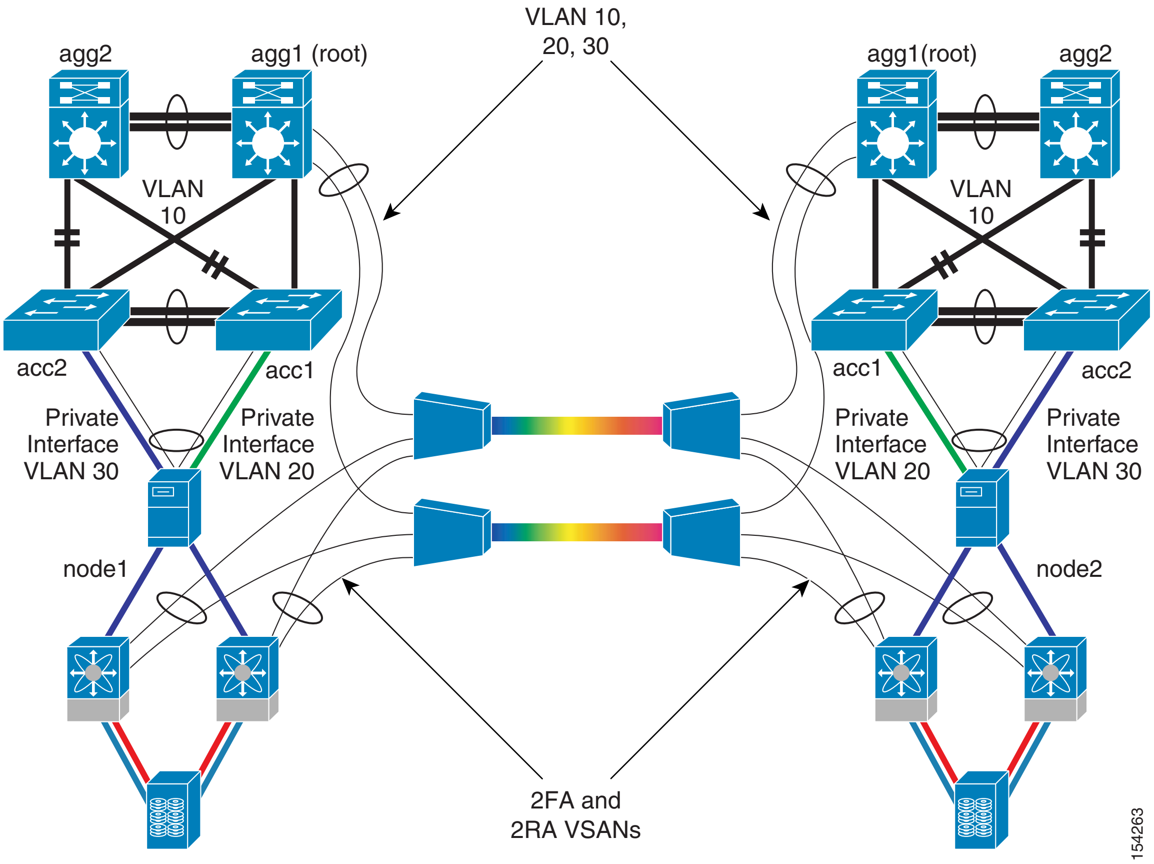
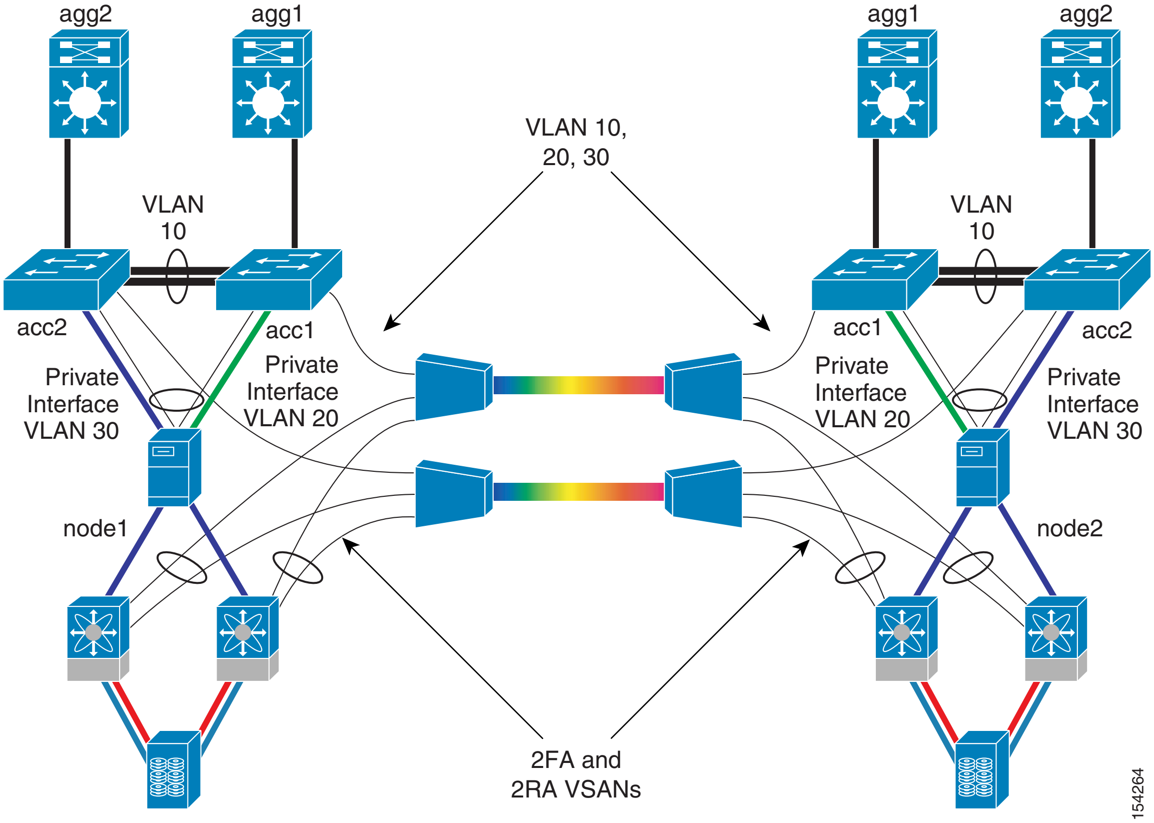


Figure 3-36 shows the cluster extension in the presence of a looped access.

Figure 3-36 Cluster Extension within a Campus (Built on Top of a Loop-Free Access)



In this case, it makes sense to connect the access switches between the sites directly via the CWDM. Depending on how you trunk the VLANs, you may have a completely loop-free solution or a squared loop between the two sites. BPDU filtering is not required in this case. Optionally, if the access switches support cross-stack EtherChannels, you would have a completely loop-free topology even in the presence of redundant paths.

In sum, there are at least two VLAN trunking options:

- Access1-site1 to access-1-site2 trunking VLAN 10, 20, and 30; and access2-site1 to access2-site2 carrying VLAN 10, 20, and 30 (squared loop topology between the two sites)
- Access1-site1 to access-1-site2 trunking VLAN 10 and 20; and access2-site1 to access2-site2 carrying VLAN 30. This topology is loop-free, but if access1 goes down, the public VLAN is segmented. This is not a problem because the two nodes can still communicate on VLAN 30, and the routing configuration ensures that traffic enters into serverfarm1 where under normal conditions, node1 owns the quorum.

In a campus environment, the distance between the sites is typically limited to multiples of 100 m, so the performance of operating the cluster from node2 using the disk in site1 is not significantly different.

Synchronous replication is used; disk failover may not be used with the caveat that a complete failure of site1 (including the disk) requires manual intervention to restart the application from site2. The failure of node1, or its network components, recovers automatically from site2 (node2).

The following is an example of campus cluster configuration:

- LAN extension via CWDM
- SAN extension via CWDM

- Assisted disk failover with software similar to EMC SRDF/CE, or simply manual disk failover (longer RTO)
- Disks configured for synchronous replication (both quorum and application data)

This case study showed two topologies using CWDM as a transport. Other typical transport choices include the following:

- DWDM (in a campus, this provides an enormous increment in bandwidth)
- EoMPLS (in the presence of an MPLS core)

Metro Cluster

A metro cluster involves data center distances up to ~100 km apart. Typical transport technologies that can be used for this type of deployment include the following:

- DWDM—Provides point-to-point “logical connections” on top of a physical ring
- CWDM—Provides point-to-point “logical connections” on top of a physical ring
- Metro Ethernet—Can in turn rely on DWDM, SONET, and MPLS technology to provide point-to-point pseudowire services
- EoMPLS—Provides a pseudowire on top of an MPLS core by tunneling Layer 2 traffic

At metro distances, you use synchronous replication, which causes the performance on node1 to decrease. For example, Cisco tests showed that the maximum IOPS achievable on node1 goes down by 37 percent at 100 km with 70 percent read and 30 percent write. A common rule of thumb is to expect a 50 percent performance decrease every 150 km.

A design choice needs to be made as to whether only servers can failover automatically, or the disks should failover together with the servers. The difference is as follows

- Node failover—If node1 fails, node2 writes and reads from DiskArray1 (as in the campus cluster example). The performance penalty of operating the application from node2 is ~10 percent compared to operating the application from node1. If site1 fails, the application needs to be manually restarted from site2.
- Node failover with software-assisted disk failover—By using appropriate software (such as EMC SRDF Cluster Enabler), you can failover the disks together with the node, so that node2 reads and writes from DiskArray2. By doing this, there is no performance penalty (besides the penalty because of the use of synchronous replication).

In the first case, if you need to keep using node2 for an extended period of time, you may want to failover the disk manually and perform the necessary configuration operations of LUN masking and zoning to use DiskArray2 instead of DiskArray1 (that is, making the disks on DiskArray2 RW and making the disks on DiskArray1 WD). The failover command is as follows:

```
C:\>symrdf -g HA1 failover
An RDF 'Failover' operation execution is in progress for device group 'HA1'. Please
wait...
Write Disable device(s) on SA at source (R1)..............Done.
Suspend RDFlink(s)..........Done.
Read/Write Enable device(s) on RA at target (R2)..........Done.
The RDF 'Failover' operation successfully executed for device group 'HA1'.
```

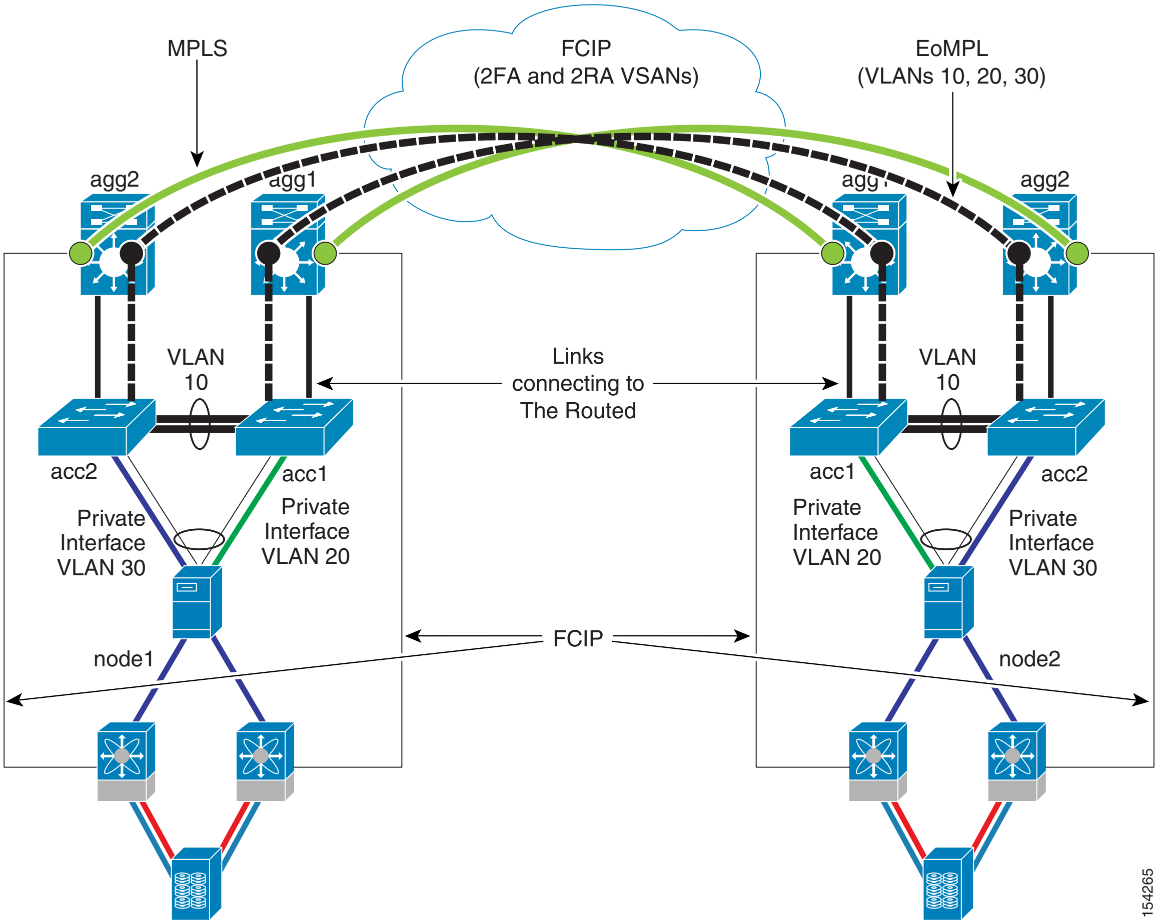
Note Note that by just failing over the disk, there is no automatic synchronous replication from site2 to site1 unless you also change the disk role of DiskArray2 from R2 to R1

If you use a solution similar to EMC SRDF/CE, it is equivalent to having the clustering software issue the failover command **symrdf –g HA1 failover**. The special software that provides the communication between the cluster software and the disks gives you control over whether you also want to change the disks in site2 from R2 to R1 on failover.

Note Note that the two configurations (without disk failover and with automatic disk failover) require a storage configuration that is somewhat different in terms of zoning, LUN masking, and disk groups. This topic is further expanded in Manual Disk Failover and Failback, page 3-43 and Software-Assisted Disk Failover, page 3-47.

Figure 3-37 shows an example of a metro cluster built on top of an existing MPLS infrastructure used by the customer for its MAN.

Figure 3-37 Metro Cluster with EoMPLS



The cluster private and public VLANs are extended by creating an EoMPLS tunnel (which is supported in hardware on the sup720) from the aggregation switches at each site. FCIP is used for SAN extension over the routed infrastructure as documented in Chapter 4, "FCIP over IP/MPLS Core."

The following is an example of metro cluster configuration:

- LAN extension via EoMPLS for a self-managed MPLS-based MAN (or a pseudowire solution from a service provider)

- SAN extension via FCIP over MPLS
- Assisted disk failover with software similar to EMC SRDF/CE
- Disks configured for synchronous replication (both quorum and application data)
- Optionally, two nodes in the primary site and one node at the remote site with “lateral failover” configured; that is, if the primary node fails, the application is recovered from the second node at the primary site, without the need to failover the application to the secondary site.

Note You can potentially carry FCIP inside the EoMPLS pseudowire. This may not be the best option for availability because you may have both LAN and SAN disconnected. A better option is to separate LAN extension and SAN extension failures.

Many other design combinations are possible. For example, you can use DWDM for SAN extension and EoMPLS for LAN extension, or you can build 10 Gbps connectivity with DWDM and run LAN extension and FCIP SAN extension on the 10 GigE transport.

Regional Cluster

At regional distances, it may be possible to use synchronous replication, but there may be the need to use asynchronous replication. Assuming that the distances do not exceed ~100 or 150 km, it may be a wise choice to use synchronous replication and to integrate the cluster software with a product such as EMC SRDF/CE to control the disk failover in conjunction with the node failover (this is what is called assisted disk failover).

Note Note that cluster and disk software that provide assisted disk failover may not be compatible with asynchronous disks. Consult your storage vendor to verify your design assumptions.

A regional cluster typically leverages the following technologies:

- SONET—Protected or unprotected SONET circuits, which can be connected together to create an SRP/RPR ring
- DWDM combined (depending on the distances) with EDFA amplification—DWDM transponders allow unamplified distances of more than 300 km and can be combined with the distance extension feature to spoof buffer-to-buffer credits (BB_credits) or can leverage MDS extended BB_credits for distances of up to 7000 km.
- Metro Ethernet—Can in turn rely on DWDM, SONET, and MPLS technology to provide point-to-point pseudowire services.
- EoMPLS—Provides a pseudowire on top of an MPLS core by tunneling Layer 2 traffic.

Note that a cluster using the quorum disk approach may require that the disk be synchronously replicated. It is very possible to have a quorum disk synchronously replicated and a data disk asynchronously replicated. The performance degradation experienced by the quorum disk may not be a problem (verify with the cluster vendor), because the figures of IOPS may not be that important for the quorum disk.

Besides the disk replication mechanism, you need to consider the BB_credit design to be able to achieve the distance between the sites. At the maximum Fibre Channel frame size of 2148 bytes, one BB_credit is consumed every two kilometers at 1 Gbps, and one BB_credit per kilometer at 2 Gbps. Given an average Fibre Channel frame size for replication traffic between 1600–1900 bytes, a general guide for allocating BB_credits to interfaces is as follows:

- 1.25 BB_credits for every two kilometers at 1 Gbps, which equals the 255 BB_credits of the Cisco MDS line cards ~400 km
- 1.25 BB_credits for every 1 kilometer at 2 Gbps, which equals ~200 km with the BB_credits of the MDS line cards.

Note The BB_credits depend on the Cisco MDS module type and the port type. The 16-port modules use 16 BB_credits on FX ports and 255 on (T)E ports. The BB_credits can be increased to 255. Host-optimized modules use 12 BB_credits regardless of the port type, and this value cannot be changed.

You can further extend the distances by using the following technologies:

- Extended BB_credits—The MDS also offers the capability to configure up to 3500 BB_credits per port with a license. This feature is available on the MPS 14/2 card (this is the same card that provides FCIP functionalities).
- BB_credits spoofing on the DWDM cards—The 2.5 G and 10 G datamux cards on the Cisco ONS 15454 provide BB_credit spoofing, allowing for distance extension up to 1600 km for 1 Gbps FC and 800 km for 2Gbps FC. For more information, see the following URL: http://www.cisco.com/univercd/cc/td/doc/product/ong/15400/r70docs/r70dwdmr/d70cdref.htm#wp907905)
- BB_credits spoofing for FC over SONET—This feature enables SAN extension over long distances through BB_credit spoofing by negotiating 255 credits with the FC switch and providing 1200 BB_credits between SL cards: 2300 km for 1 Gbps ports and 1150 km for 2 Gbps ports (longer distances supported with lesser throughput). For more information, see the following URL: http://www.cisco.com/univercd/cc/td/doc/product/ong/15400/r70docs/r70refmn/r70sancd.htm)

An example of regional cluster configuration is as follows:

- LAN extension via DWDM
- SAN extension via FC over DWDM with BB_credits distance extension
- Assisted disk failover with a software similar to EMC SRDF/CE
- Disks configured for synchronous replication (both quorum and application data)
- Optionally, two nodes in the primary site and one node at the remote site with “lateral failover” configured; that is, if the primary node fails, the application is recovered from the second node at the primary site without the need to failover the application to the secondary site.

Continental Cluster

The transport used at continental distances most likely belongs to one of the following categories:

- SONET circuit—Carrying Ethernet and FCP on different circuits, or a single pair of Ethernet circuits for Layer 2 extension and FCIP.
- Generic Ethernet circuit from a service provider (based on SONET/DWDM lambdas/MPLS pseudowires)

Note SPs may not offer FC over SONET services. If your service provider offers only Ethernet Over SONET, consider using FCIP.

At continental distances, disk replication is based on asynchronous replication for at least two reasons:

- The application performance with synchronous replication at 1000 km is typically unacceptable because the number of IOPS goes down enormously. In the Cisco test bed. for example, a local node can perform ~3664 IOPS with synchronous replication with 400m distance between the data centers, and ~251 IOPS with 2000 km distance between the sites. By using asynchronous replication, you can go up to more than 3000 IOPS.
- With synchronous replication, per every write you need to wait a response ready before sending the data, so tuning the TCP window for FCIP to achieve the maximum throughput offered by the data center transport does not help. With asynchronous replication, it is possible to send multiple unacknowledged writes, thus taking full advantage of the bandwidth between the data centers.

The throughput requirements of the application (and as a result, of the storage replication) can be addressed by taking advantage of the following technologies:

- Extended BB_credits—The Cisco MDS also offers the capability to configure up to 3500 BB_credits per port with a license. The feature works on the MPS 14/2 module.
- BB_credits spoofing on the DWDM cards—The 2.5 G and 10 G datamux cards on the ONS15454 provide BB_credit spoofing, allowing for distance extension up to 1600 km for 1 Gbps FC and 800 km for 2 Gbps FC. For more information, see the following URL: http://www.cisco.com/univercd/cc/td/doc/product/ong/15400/r70docs/r70dwdmr/d70cdref.htm#wp907905)
- BB_credits spoofing for FC over SONET—This feature enables SAN extension over long distances through BB_credit spoofing by negotiating 255 credits with the FC switch and providing 1200 BB_credits between SL cards: 2300 km for 1 Gbps ports and 1150 km for 2 Gbps ports (longer distances supported with lesser throughput). For more information, see the following URL: (http://www.cisco.com/univercd/cc/td/doc/product/ong/15400/r70docs/r70refmn/r70sancd.htm)

Some cluster implementations require that the quorum disk be synchronously replicated. This should not prevent building a continental cluster for the following reasons:

- The quorum disk can be configured for synchronous replication at 1000 km because the performance in terms of IOPS may be small but still enough for the quorum purposes.
- If the quorum disk approach is not working, you can use other quorum mechanisms, such as the majority node set, in which case you configure two nodes in the first site and one node at the remote site, for example.

Configuring two nodes at the primary site and one node at the remote site is desirable in any case, because with continental clusters, you may want a server failure to be recovered locally. Disk failover may not be possible, depending on the storage vendor and on what software-assisted disk failover is available from the storage vendor. For example, EMC SRDF/CE does not support asynchronous replication. In the case of EMC, you may want to consider the EMC/Legato Autostart or Automated Availability Manager (AAM) solution.

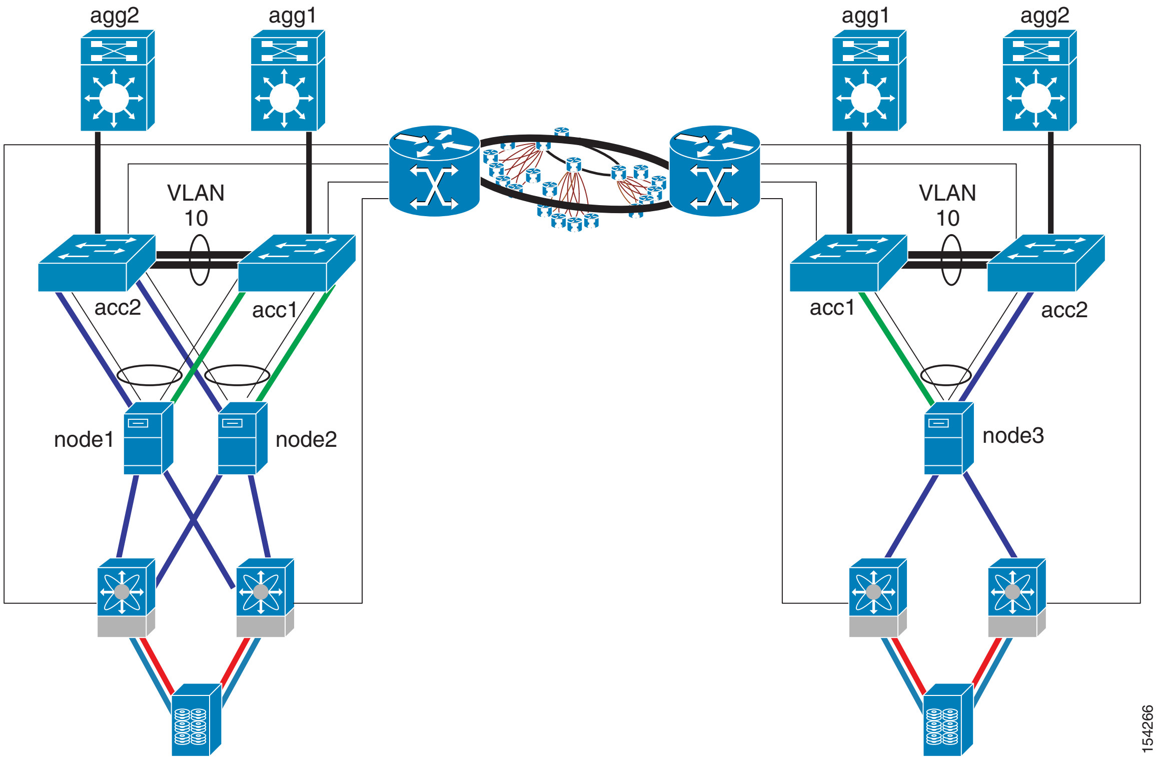
The following is an example of continental cluster configuration:

- LAN extension via Ethernet over SONET
- SAN extension via FCIP over SONET
- Scripted disk failover
- Disks configured for asynchronous replication (application data disk)
- Majority node set quorum—Note that this is *network-based* and relies on a network share containing a replica of the quorum data. This majority approach uses server message bloc (SMB) to mount the disks across servers (the servers use the local disks to manage the quorum information), which in turn requires Layer 2 extension (provided via Ethernet over SONET).

- Optionally, two nodes in the primary site and one node at the remote site with "lateral failover" configured; that is, if the primary node fails, the application is recovered from the second node at the primary site, without the need to failover the application to the secondary site). This can be achieved by configuring the clustering software "preferred owner list".

Figure 3-38 shows a continental cluster built on top of a SONET transport (SONET circuits provided by a service provider.

Figure 3-38 Continental Cluster over SONET



The connectivity to the SONET cloud can be configured as follows:

- One (or two for redundancy) Ethernet over SONET circuit for LAN extension (an STS-1 might be just enough for the heartbeats, but you may need more if you plan to have client-to-server traffic traversing the WAN); one (or two for redundancy) circuit for SAN replication with FC over SONET. If the circuits are "Layer 1", you can run an end-to-end port channel.
- One (or two for redundancy) Ethernet over SONET circuit for LAN extension (an STS-1 might be just enough for the heartbeats, but you may need more if you plan to have client-to-server traffic traversing the WAN); one (or two for redundancy) circuit for SAN replication with FC over SONET. For LAN extension, you can terminate the port channel locally on an ML-series card and use SRP/RPR to manage the geographical ring.
- One (or two for redundancy) Ethernet over SONET circuit for LAN extension, another one (or two for redundancy) Ethernet over SONET circuit for SAN extension via FCIP. This option can in turn be configured with end-to-end port channels as well as local port channel termination on an ML-card and SRP/RPR to manage the geographical ring.

Potentially, you can carry SAN extension and LAN extension on the same circuit, but with Layer 2 LAN extension, you may want to keep the two on different circuits so that a broadcast storm or a Layer 2 loop caused by a misconfiguration does not affect the SAN traffic.

Note EMC/Legato AAM provides a cluster solution that does not require Layer 2 extension, in which case you can route the heartbeat traffic over the SONET link. Managed IP addresses (virtual IP address) can be on different subnets in each site, and DNS ensures rerouting of the client traffic to the site where the application is available.

Storage Design Considerations

This section analyzes the disk failover design by comparing “manual” disk failover versus “software-assisted” disk failover.

Manual disk failover refers to the design where the cluster software on all the nodes performs read and writes on the same disk array, regardless of which node is active. With this design, you may have node2 in data center 2 read and writing on DiskArray1 in data center 1. When data center 1 is completely unavailable, you need to perform a “manual” failover of the disk and restart the application.

Software-assisted disk failover refers to the design where a node failure may cause a disk failover. For example, if node1 in data center 1 fails, node2 in data center may become active and the disk fails over from DiskArray1 to DiskArray2. This behavior requires some software as an interface between the clustering software and the disk array, or simply some scripts that the cluster invokes when a node fails over. With software-assisted failover, it may be a good practice to deploy two nodes in the primary site and one node in the remote site, so that a node failure can be recovered locally without having to perform a disk failover.

In either design, a complete site failure does not bring up the remote resources automatically. With software-assisted disk failover, the administrator can restart the application at the remote site by using a GUI. The reason why this is not automatic is because a complete site failure is indistinguishable from the loss of LAN and SAN connectivity between the sites.

Manual Disk Failover and Failback

A cluster design that relies on manual disk failover consists of two or more cluster nodes zoned to see the same storage (for example, DiskArray1 in data center 1). The disk array in the remote site (for example, DiskArray2) is used for replication but not physically visible to the remote node. Failover of the nodes is automatic, but failover of the disks is not. This is only a problem in the following two cases:

- Complete failure of the primary site (because the local disks would be configured for RAID, a disk failure does not result in a failover across disk arrays)—The RTO is longer than in the case of software-assisted disk failover.
- Performance of node2 at long distances between the sites (for example, regional or continental distances)

With this design, the disks may be grouped. For example, the quorum and the application data disks may be part of the same disk group, so as to failover together (this is not the case with the software-assisted disk failover).

When you need to failover the disks to the secondary site, you need to perform the following operations:

- Failover the disks (the command with EMC is **symrdf –g HACluster failover**)
- Reconfigure the zoning on the VSANs so that the remote node (node2) sees the disk array at the remote site (DiskArray2).

- Reconfigure the LUN mapping so that the LUNs on the remote disk array are mapped to be seen by node1 and node2.

The following example shows the failover of the disks from the primary site to the remote site:

```
C:\>symrdf -g HACluster failover

Execute an RDF 'Failover' operation for device
group 'HACluster' (y/[n]) ? y

An RDF 'Failover' operation execution is
in progress for device group 'HACluster'. Please wait...

    Write Disable device(s) on SA at source (R1)..............Done.
    Suspend RDF link(s).......................................Done.
    Read/Write Enable device(s) on RA at target (R2)..........Done.

The RDF 'Failover' operation successfully executed for
device group 'HACluster'.
```

The following example shows the change of the LUN mapping on the remote disk arrays to present the disks (devs 0029 and 002A) to node2 (WWN 10000000c92c0f2e) via the disk array ports (-dir 3A –p 0):

```
symmask -sid 1291 -wwn 10000000c92c0f2e add devs 0029,002A -dir 3a -p 0

symmask -sid 1291 refresh

C:\Documents and Settings\Administrator>symmaskdb list database -sid 1291

Symmetrix ID            : 000187431291

Database Type           : Type5
Last updated at         : 08:02:37 PM on Fri Nov 11,2005

Director Identification : FA-3A
Director Port           : 0

                               User-generated
Identifier        Type   Node Name        Port Name         Devices
----------------  -----  ---------------------------------  ---------
10000000c92c142e  Fibre  10000000c92c142e 10000000c92c142e  0029:002A
10000000c92c0f2e  Fibre  10000000c92c0f2e 10000000c92c0f2e  0029:002A

Director Identification : FA-4A
Director Port           : 0
```

After the failover and the LUN mapping configuration, you can verify that the disks are RW (before the failover they were configured as WD) as follows:

```
C:\Documents and Settings\Administrator.EXAMPLE>sympd list

Symmetrix ID: 000187431291

        Device Name           Directors                  Device
-------------------------- ------------- -------------------------------------
                                                                           Cap
Physical               Sym  SA :P DA :IT  Config        Attribute    Sts   (MB)
-------------------------- ------------- -------------------------------------

\\.\PHYSICALDRIVE2     0029 03A:0 01C:C2  RDF2+Mir      N/Grp'd       RW    8714
\\.\PHYSICALDRIVE3     002A 03A:0 16B:C2  RDF2+Mir      N/Grp'd  (M) RW   43570
```

Note that there is no expectation at this point that the writes to the disks on DiskArray2 are replicated to DiskArray1 unless you perform a swapping of the R1 and R2 roles. Writing to DiskArray2 increments the number of Invalid tracks:

```
C:\Documents and Settings\Administrator>symrdf -g HACluster query


Device Group (DG) Name             : HACluster
DG's Type                          : RDF1
DG's Symmetrix ID                  : 000187431320


       Source (R1) View                 Target (R2) View     MODES
--------------------------------    ------------------------ ----- ------------
             ST                  LI      ST
Standard      A                   N       A
Logical       T  R1 Inv   R2 Inv  K       T  R1 Inv   R2 Inv       RDF Pair
Device  Dev   E  Tracks   Tracks  S Dev   E  Tracks   Tracks MDA   STATE
-------------------------------- -- ------------------------ ----- ------------

DEV001  0029 WD       0        0 NR 0029 RW      38        0 S..   Failed Over
DEV002  002A WD       0        0 NR 002A RW   13998        0 S..   Failed Over

Total          -------- --------           -------- --------
  Track(s)            0        0              14036        0
  MB(s)             0.0      0.0              438.6      0.0

Legend for MODES:

 M(ode of Operation): A = Async, S = Sync, E = Semi-sync, C = Adaptive Copy
 D(omino)           : X = Enabled, . = Disabled
 A(daptive Copy)    : D = Disk Mode, W = WP Mode, . = ACp off
```

The invalid tracks are synchronized back to DiskArray1 when you perform a “restore” or a “failback”.

```
C:\Documents and Settings\Administrator>symrdf -g HACluster failback.

Execute an RDF 'Failback' operation for device
group 'HACluster' (y/[n]) ? y

An RDF 'Failback' operation execution is
in progress for device group 'HACluster'. Please wait...

    Write Disable device(s) on RA at target (R2)..............Done.
    Suspend RDF link(s).......................................Done.
    Merge device track tables between source and target.......Started.
    Devices: 0029-002E ...................................... Merged.
    Merge device track tables between source and target.......Done.
    Resume RDF link(s)........................................Started.
    Resume RDF link(s)........................................Done.
    Read/Write Enable device(s) on SA at source (R1)..........Done.

The RDF 'Failback' operation successfully executed for
device group 'HACluster'.
```

After the “failback”, the number of invalid tracks slowly returns to zero:

```
C:\Documents and Settings\Administrator>symrdf -g HACluster query


Device Group (DG) Name             : HACluster
DG's Type                          : RDF1
DG's Symmetrix ID                  : 000187431320
```

```
       Source (R1) View                 Target (R2) View     MODES
--------------------------------    ------------------------ ----- ------------
             ST                  LI      ST
Standard      A                  N       A
Logical       T  R1 Inv   R2 Inv  K       T  R1 Inv   R2 Inv       RDF Pair
Device  Dev   E  Tracks   Tracks  S Dev   E  Tracks   Tracks MDA   STATE
-------------------------------- -- ------------------------ ----- ------------

DEV001  0029 RW      33        0 RW 0029 WD      33        0 S..   SyncInProg
DEV002  002A RW    9914        0 RW 002A WD    7672        0 S..   SyncInProg

Total          -------- --------           -------- --------
  Track(s)         9947        0               7705        0
  MB(s)           310.8      0.0              240.8      0.0

Legend for MODES:

 M(ode of Operation): A = Async, S = Sync, E = Semi-sync, C = Adaptive Copy
 D(omino)           : X = Enabled, . = Disabled
 A(daptive Copy)    : D = Disk Mode, W = WP Mode, . = ACp off


C:\Documents and Settings\Administrator>symrdf -g HACluster query


Device Group (DG) Name             : HACluster
DG's Type                          : RDF1
DG's Symmetrix ID                  : 000187431320


       Source (R1) View                 Target (R2) View     MODES
--------------------------------    ------------------------ ----- ------------
             ST                  LI      ST
Standard      A                  N       A
Logical       T  R1 Inv   R2 Inv  K       T  R1 Inv   R2 Inv       RDF Pair
Device  Dev   E  Tracks   Tracks  S Dev   E  Tracks   Tracks MDA   STATE
-------------------------------- -- ------------------------ ----- ------------

DEV001  0029 RW       0        0 RW 0029 WD       0        0 S..   Synchronized
DEV002  002A RW       0        0 RW 002A WD       0        0 S..   Synchronized

Total          -------- --------           -------- --------
  Track(s)            0        0                  0        0
  MB(s)             0.0      0.0                0.0      0.0

Legend for MODES:

 M(ode of Operation): A = Async, S = Sync, E = Semi-sync, C = Adaptive Copy
 D(omino)           : X = Enabled, . = Disabled
 A(daptive Copy)    : D = Disk Mode, W = WP Mode, . = ACp off
```

In sum, when designing a solution for manual disk failover, consider the following factors:

- The main need for Layer 2 extension is driven by the cluster software. If the cluster software does not require Layer 2 extension, you may be able to build a routed infrastructure to interconnect the two sites.
- Tune the routing to route traffic preferably to the primary site, because this is the site where the node is normally active. This is also the site whose disk array is used by both local and remote nodes.
- SAN zoning needs to remember that both nodes need to see only the storage in the primary site, so node1 in data center 1 and node2 in data center 2 need to be zoned to see DiskArray 1.

- LUN mapping on the disk array follows a similar configuration as the zoning, in that the LUNs in DiskArray1 need to be presented to node1 and node 2.

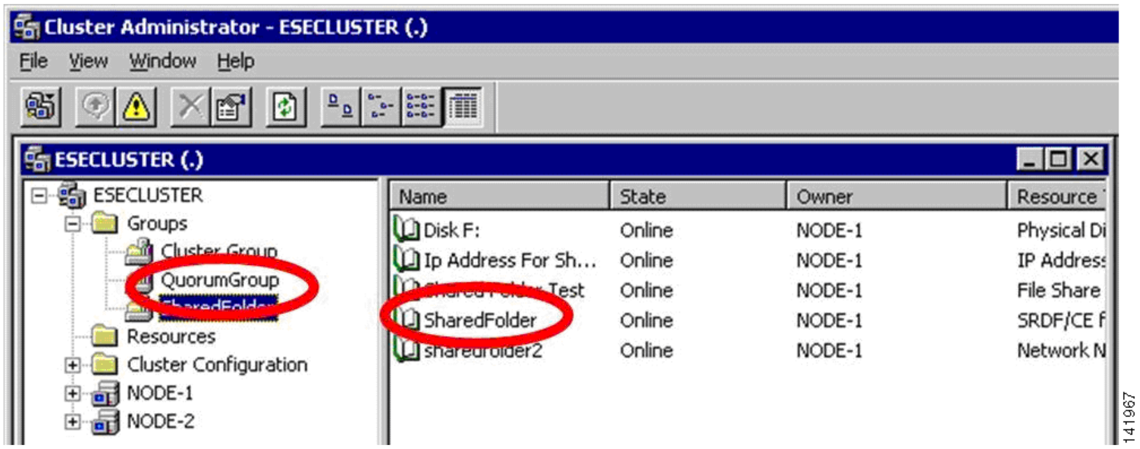
Software-Assisted Disk Failover

With software-assisted disk failover, each node is zoned to the disk array local to the node. In the remote site, the disks are in write disabled mode, which is why the special software is required, to control the disks and synchronize the operations with the cluster software. The cluster tries to access the quorum disk from the remote node, which is not possible if the quorum disk is write disabled.

Differently from the manual failover configuration, the software-assisted disk failover configuration has each disk zoned to the respective node, and the LUN mapping configured accordingly: node1 is zoned to see DiskArray1, and node2 is zoned to see DiskArray2. LUN mapping on DiskArray1 presents the LUNs to node1, and LUN mapping on DiskArray2 presents the LUNs to node2.

Differently from the manual failover configuration, each disk is its own group. The reason is that node1 may own the quorum disk on DiskArray1 and node2 may own the application data disk on DiskArray2. Figure 3-39 shows the *quorum group* and the *shared folder* group. These are disk groups of a single disk, and they are managed by the special software that interfaces the cluster with the disks (EMC SRDF/CE in the Cisco test environment).

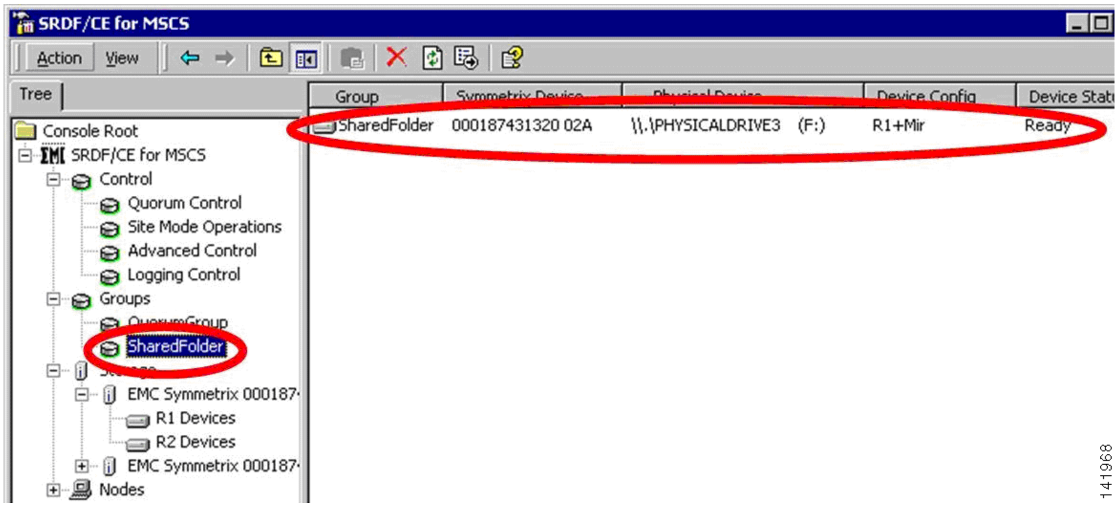
Figure 3-39 Cluster Configuration Showing Resources Managed by SRDF Cluster Enabler



Name	State	Owner	Resource
Disk F:	Online	NODE-1	Physical Di
Ip Address For Sh...	Online	NODE-1	IP Address
SharedFolder Test	Online	NODE-1	File Share
SharedFolder	Online	NODE-1	SRDF/CE f
sharedfolder2	Online	NODE-1	Network N

If the public NIC of the primary node fails, the associated application disk (shared folder) can be failed over via the SRDF/CE software while the quorum group may still be owned by the primary node. Figure 3-40 shows the shared folder disk group from the cluster enabler GUI.

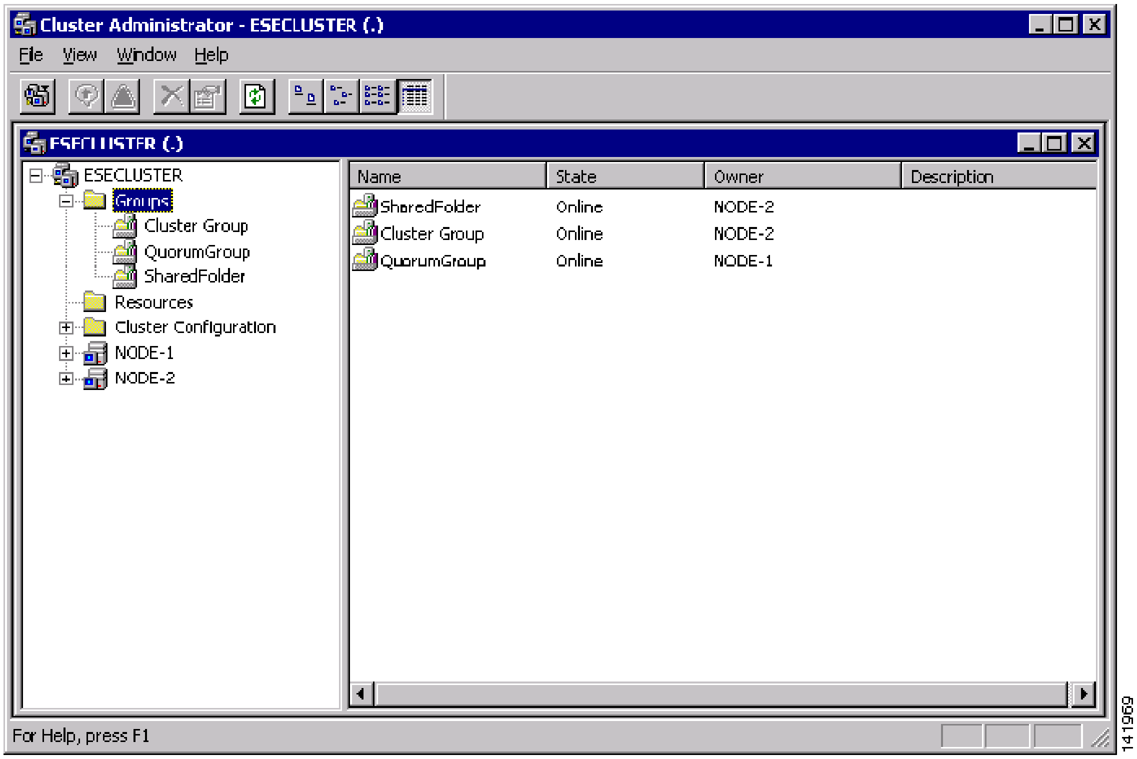
Figure 3-40 Cluster Enabler View of the Storage Resources



Group	Symmetrix Device	Physical Device	Device Config	Device State
SharedFolder	000187431320 02A	\\.\PHYSICALDRIVE3 (F:)	R1+Mir	Ready

Figure 3-41 shows the failover of the Shared Folder disk when the NIC on the primary node fails, while the quorum group is still owned by the primary node. The owner for the shared folder disk is node2, while the owner for the quorum is still node1.

Figure 3-41 Shared Folder Failover when the NIC on the Primary Node Fails

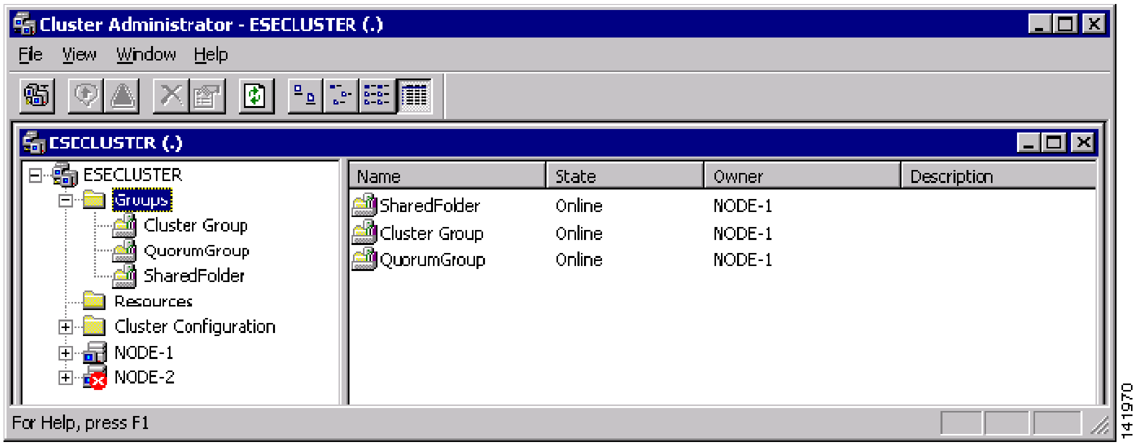


Name	State	Owner	Description
SharedFolder	Online	NODE-2	
Cluster Group	Online	NODE-2	
QuorumGroup	Online	NODE-1	

For Help, press F1

If the two data center sites are disconnected on the extended LAN segment, and node2 owns the application disk, node1 takes back the application disk, as shown in Figure 3-42. The application processing continues as long as the routing is configured to advertise the path to data center 1 with a better cost than data center 2. That is, you need to configure the routing cost to match the site where the preferred owner for the quorum is located.

Figure 3-42 Application Disk Failback to Node1 when LAN Connectivity Fails



Name	State	Owner	Description
SharedFolder	Online	NODE-1	
Cluster Group	Online	NODE-1	
QuorumGroup	Online	NODE-1	

If all communication (LAN extension and SAN extension) between the sites is lost, no failover happens. This type of failure can be caused by the following two scenarios:

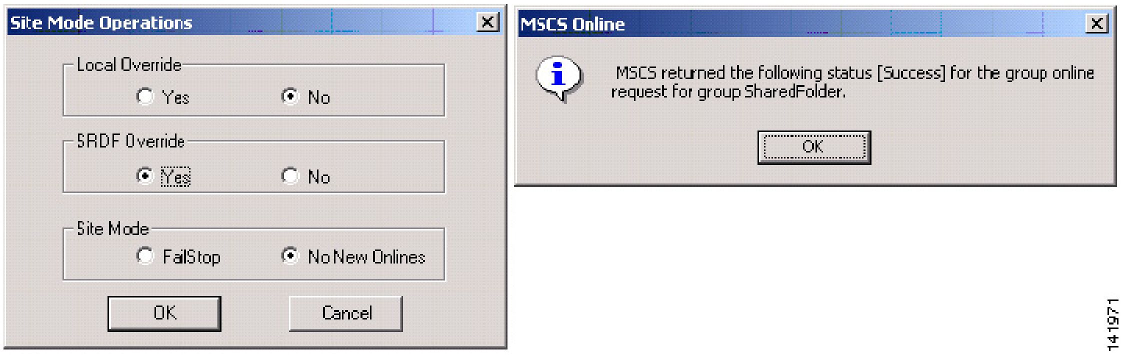
- Lost connectivity between the sites (in which case failover is not necessary)
- Complete failure of the primary site

If the SRDF communication is lost and you want to failover to the secondary site, you need to take the following steps:

1. From the cluster enabler GUI, select the option "SRDF override".
2. From the cluster enabler GUI, failover the quorum and the application disk.
3. From the cluster software GUI, bring the quorum and the application group online.

Figure 3-43 shows the cluster enabler configuration to restart the operations at the secondary site and the cluster (MSCS) message when the application group is brought online successfully.

Figure 3-43 Cluster Enabler Configuration



Local Override: ○ Yes ● No

SRDF Override: ● Yes ○ No

Site Mode: ○ FailStop ● No New Onlines

MSCS returned the following status [Success] for the group online request for group SharedFolder.

When the primary site is back online, you need to follow a specific procedure to restore the communication between the cluster and cluster enabler software running on the cluster nodes, and to ensure that the disks are synchronized. This is out of the scope of this document.

In sum, when designing a solution with software-assisted disk failover, consider the following factors:

- There may need to be Layer 2 communication between the cluster enabler software components (for example, EMC SRDF/CE uses a local multicast address 127.0.0.x with TTL=1).

- Tune the routing to match the quorum disk preferred owner configuration. For example, if the preferred owner is node1 in data center1, make sure that the cost of the route to data center 1 for the cluster subnets is preferred to the path to data center2.
- SAN zoning needs to remember that each node needs to see only the storage local to the node, so node1 in data center 1 needs to be zoned to see DiskArray 1 and node2 in data center 2 needs to be zoned to see DiskArray2.
- LUN mapping on the disk array follows a similar configuration as the zoning, in that the LUNs in DiskArray1 need to be presented to node1 and the LUNs in DiskArray2 need to be presented to node2.
- If disk groups need to be configured, make sure that they can be failed over independently because the cluster enabler software may have one disk active in data center 1 and another disk active in data center 2.

Network Design Considerations

This section focuses on the LAN extension part of the cluster design.

As previously mentioned, clusters often require Layer 2 connectivity between the sites. From a routing and switching point of view, this is not the best practice; however, besides a few cluster products, it is currently often a requirement to provide an extended Layer 2 LAN segment.

The network design needs to consider the following factors:

- Provide as much availability as possible, which means providing redundant network components.
- Avoid as much as possible losing connectivity on both the LAN extension and SAN extension (keeping SAN connectivity may be more important than keeping LAN connectivity, but losing both looks like a complete site failure to the cluster software).
- LAN communication between the nodes typically consists of heartbeats (small UDP datagrams). These datagrams can be unicast (normally, if there are only two nodes involved in the cluster) or multicast. Very often this multicast traffic is local multicast with TTL=1; that is, it is not routable. It is also common to have SMB traffic. This means that Layer 2 extension is often required.
- On the LAN side, provide multiple paths for the public and the private cluster segments. You may give up some redundancy on either LAN segment (which means you may be able to avoid including spanning tree in the design) as long as you can disassociate failures of the two segments. That is, as long as the nodes can talk on either the public or the private segment, the cluster is still usable.
- Tune the routing such that when the public LAN segment is disconnected, the application can still continue to operate, and no user is routed to the site where the node is in standby.

LAN Extension and Redundancy

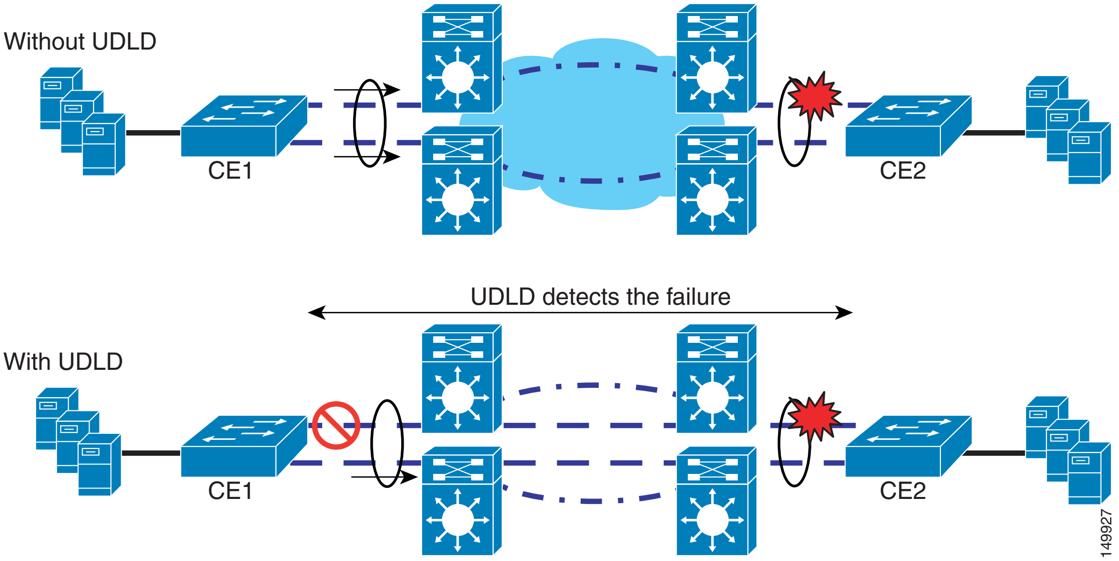
Layer 2 LAN extension means that either spanning tree or EtherChannels need to be used. Running spanning tree across sites works fine, but many customers prefer to design loop-free topologies from the very beginning to avoid dealing with Layer 2 loops caused by misconfigurations. Port channeling can be used to implement loop-free and redundant designs.

EtherChannels and Spanning Tree

There are no special design considerations when EtherChannels are implemented as a client-protection mechanism on CWDM or DWDM extended LANs, or even on Layer 1 Ethernet over SONET circuits (with G-series cards, for example). If the remote port fails, there is a link down on the local node and port channeling uses the remaining links to send the traffic.

When deploying EtherChannel across pseudowires, such as EoMPLS tunnels, you need to use some mechanism to detect far-end failures. UniDirectional Link Detection (UDLD) can be used for this purpose. Figure 3-44 shows the use of UDLD for these type of failures.

Figure 3-44 Remote Failure Detection with UDLD



Without UDLD, CE1 still sends traffic to both pseudowires, regardless of the status of the remote port. With UDLD, the remote failure is detected:

```
%UDLD-4-UDLD_PORT_DISABLED: UDLD disabled interface Gi1/0/1, aggressive mode failure
detected
%PM-4-ERR_DISABLE: udld error detected on Gi1/0/1, putting Gi1/0/1 in err-disable state
%LINEPROTO-5-UPDOWN: Line protocol on Interface GigabitEthernet1/0/1, changed state to
down
%LINK-3-UPDOWN: Interface GigabitEthernet1/0/1, changed state to down
```

Although the UDLD message time can be tuned to be ~1s on some switching platforms, the ports would change status continuously between advertisement and discovery. You need to make sure that the ports are in a bidirectional state from the UDLD point of view. For example, with a pseudowire over 20 km, a safe configuration is with UDLD message time 4s. With this configuration, the far-end failure is detected within ~11–15s.

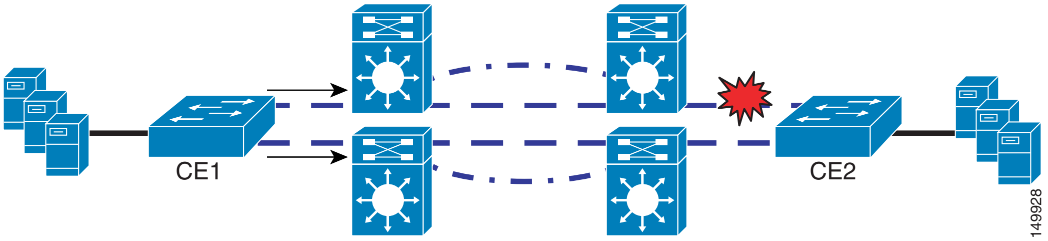
Note Recovery from a far-end failure is not automatic. When the link is recovered, you need to manually ensure that both ports on the local and remote site are shut/unshut to restore complete connectivity.

Deeper understanding of clustering software may allow loop free designs that do not rely on EtherChannel or spanning tree.

Public and Private Links

When using MSCS, the nodes can communicate either via the public LAN segment or the private LAN segment. By leveraging this capability you can design the network in a way that, for example, the public LAN segment takes one pseudowire (or lambda or circuit) and the private LAN segment takes a different pseudowire (or lambda or circuit), as shown in Figure 3-45.

Figure 3-45 Decoupling Public and Private Link on Different Pseudowires



This topology has “no redundancy” for the public LAN segment, but in reality the MPLS network already provides fast convergence and re-routing for the pseudowire. In case the remote port goes down, the heartbeat mechanism on the cluster node detects the problem and the communication between the nodes continues on the private segment.

Under normal circumstances, the client traffic has no need to be routed to the remote site via the public LAN segment, so unless there are double failures, losing the public LAN segment connectivity may be acceptable.

In conclusion, check with your cluster vendor on the cluster capabilities and consider this option as a way to extend the Layer 2 LAN segment without introducing loops by leveraging the cluster software monitoring capabilities.

Also note that the “split subnet” is not a problem if routing is designed correctly, as described in Routing Design, page 3-52.

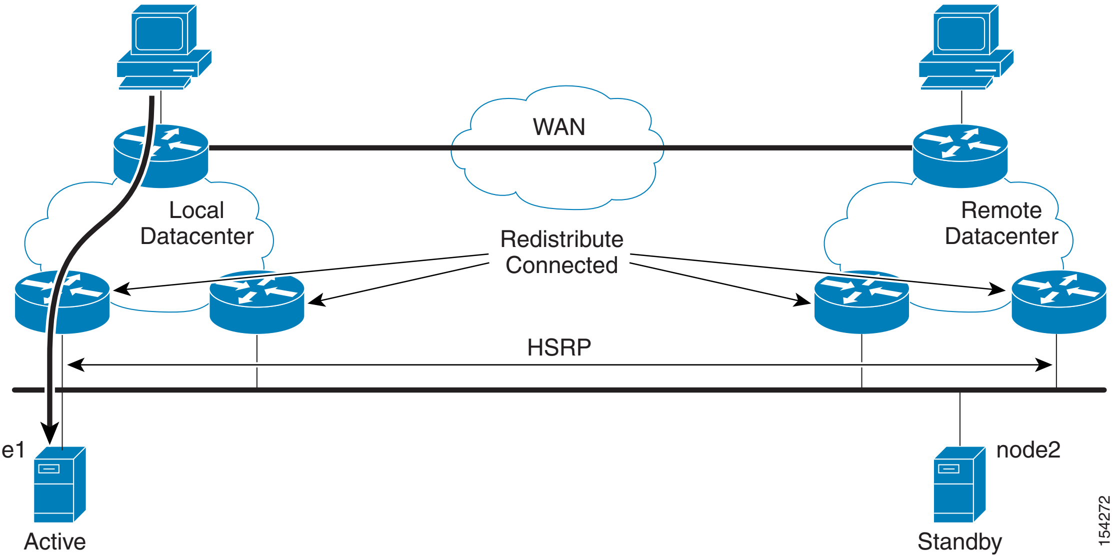
Routing Design

One of the common reasons of concerns for extended Layer 2 LAN segments is the presence of the same subnet in data centers whose subnets are usually summarized according to routing best practices.

Although having the same subnet in two data centers is not really a best practice, this can be supported for the purpose of deploying HA clusters. Advertise the “cluster” subnet as an external route by using **redistribute connected** and by filtering all subnets except the cluster subnet.

While redistributing, you can also control the cost, making the primary data center preferred to the remote one until the primary data center disappears completely. Figure 3-46 shows this concept.

Figure 3-46 Routing Design for Stretched Clusters with Layer 2 LAN Extension



Also note that the routers in the two locations may be participating in the same Hot Standby Routing Protocol (HSRP) group, so you may want to configure the routers in site 1 to have priority over the routers in site 2. This means that traffic going to node2 is going to enter and exit from site1, which, while certainly not optimized for proximity, ensures symmetric traffic paths, which is a highly desirable property, especially if firewalls are present in the path.

The following configurations help explain the design:

```
Aggregation1 (site1)
====================
interface Vlan20
 ip address 11.20.40.2 255.255.255.0
 ip helper-address 10.20.10.151
 standby 1 ip 11.20.40.1
 standby 1 priority 110
 standby 1 preempt
!
router ospf 1
 redistribute connected metric 100 subnets route-map filter-routes  network 1.1.1.0
0.0.0.255
 area 0  network 10.1.0.0 0.0.255.255 area 0
!
! Redistribute only the subnet where the HA cluster is located
!
access-list 10 permit 11.20.40.0 0.0.0.255
!
route-map filter-routes permit 10
 match ip address 10
!

Aggregation2 (site1)
====================
interface Vlan20
 ip address 11.20.40.2 255.255.255.0
 ip helper-address 10.20.10.151
 standby 1 ip 11.20.40.1
 standby 1 priority 100
 standby 1 preempt
!
```

```
router ospf 1
 redistribute connected metric 100 subnets route-map filter-routes  network 1.1.1.0
0.0.0.255
 area 0  network 10.1.0.0 0.0.255.255 area 0
!
! Redistribute only the subnet where the HA cluster is located
!
access-list 10 permit 11.20.40.0 0.0.0.255
!
route-map filter-routes permit 10
 match ip address 10
!

Aggregation3 (site2)
====================
interface Vlan20
 ip address 11.20.40.3 255.255.255.0
 ip helper-address 10.20.10.151
 standby 1 ip 11.20.40.1
 standby 1 priority 90
 standby 1 preempt
!
router ospf 1
 redistribute connected metric 110 subnets route-map filter-routes  network 1.1.1.0
0.0.0.255
 area 0  network 10.1.0.0 0.0.255.255 area 0
!
access-list 10 permit 11.20.40.0 0.0.0.255
!
! Redistribute only the subnet where the HA cluster is located
!
route-map filter-routes permit 10
 match ip address 10
!

Aggregation4 (site2)
====================
interface Vlan20
 ip address 11.20.40.3 255.255.255.0
 ip helper-address 10.20.10.151
 standby 1 ip 11.20.40.1
 standby 1 priority 80
 standby 1 preempt
!
router ospf 1
 log-adjacency-changes
 redistribute connected metric 110 subnets route-map filter-routes  network 1.1.1.0
0.0.0.255
 area 0  network 10.1.0.0 0.0.255.255 area 0
!
! Redistribute only the subnet where the HA cluster is located
!
access-list 10 permit 11.20.40.0 0.0.0.255
!
route-map filter-routes permit 10
 match ip address 10
```

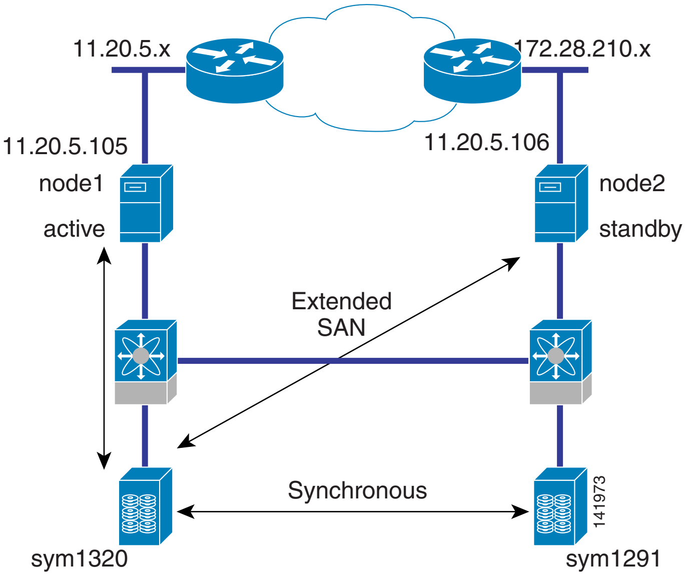
Local Area Mobility

It is very common for customers who are deploying an HA cluster solution to ask whether the network can provide a Layer 2 solution on top of a Layer 3 network. The use of EoMPLS tunnels effectively provides a Layer 2 extension solution on top of a Layer 3 network, but not every enterprise builds an MPLS core in their network. An alternative technology is local area mobility (LAM), which relies on proxy ARP and host routes.

Note Not all cluster vendors support this type of solution. Some cluster products explicitly require the user to not use proxy ARP.

Figure 3-47 shows the use of LAM for a campus cluster.

Figure 3-47 Design with Local Area Mobility



LAM allows the two nodes to unicast on the same subnet even if they are physically placed on two different subnets. This solution is applicable for two nodes only, assuming that the clustering software uses unicast for the heartbeats (this is the case with MSCS with two nodes only).

When using LAM, you can place the nodes in different subnets. The router sends an ARP request to the VIP of the cluster (configured via an ACL), and populates the routing table with a /32 route for the VIP if the node local to the router answers the ARP request for the VIP. Proxy ARP ensures that the nodes can use ARP to discover the address of each other, even if they are not locally adjacent. For example, LAM introduces host routes from site2 for the 11.20.5.106 address, and the cluster virtual address, if it moves to site2.

Routing is “optimal” in that traffic goes to the node that is advertising the /32 route. LAM monitors the nodes by periodically sending ARP requests to the VIP address and the node address.

The following configuration shows this functionality:

```
Local node
==========
int Vlan5
ip proxy-arp
ip address 11.20.5.1 255.255.255.0
```

```
Remote Node
===========
interface Vlan172
 ip address 172.28.210.1 255.255.255.0
 ip mobile arp timers 5 20 access-group MOBILE-DEVICES-ALLOWED
 ip proxy-arp
```

The **ip mobile arp** command sends an ARP to the device specified in the access list to determine whether it is available, and if it is, it adds the route in the routing table. For example, the remote node monitors 11.20.5.106 (node2) and 11.20.5.110 (the VIP address). Under the **router ospf** configuration, you need to add **redistribute mobile** to propagate the route into OSPF.

The access list specifies the addresses that need to be monitored:

```
ip access-list standard MOBILE-DEVICES-ALLOWED
 permit 11.20.5.106
 permit 11.20.5.110
```

LAM is a valid option for limited HA cluster deployments with two nodes only, when the cluster software is compatible with proxy ARP and no other software component generates local multicast traffic.