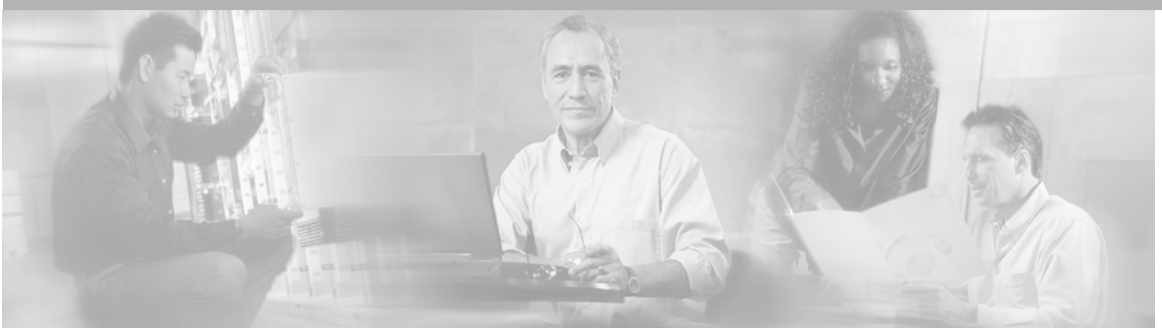
Cisco Systems



Data Center High Availability Clusters Design Guide

Corporate Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
http://www.cisco.com
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 526-4100

Customer Order Number:
Text Part Number: OL-12518-01

Data Center High Availability Clusters Design Guide

Preface

Document Purpose

Data Center High Availability Clusters Design Guide describes how to design and deploy high availability (HA) clusters to provide uninterrupted access to data, even if a server loses network or storage connectivity, or fails completely, or if the application running on the server fails.

Intended Audience

This guide is intended for system engineers who support enterprise customers that are responsible for designing, planning, managing, and implementing local and distributed data center IP infrastructures.

Document Organization

This guide contains the chapters in the following table.

Section	Description
Chapter 1, "Data Center High Availability Clusters."	Provides high-level overview of the use of HA clusters, including design basics and network design recommendations for local clusters.
Chapter 2, "Data Center Transport Technologies."	Describes the transport options for interconnecting the data centers.
Chapter 3, "Geoclusters."	Describes the use and design of geoclusters in the context of business continuance as a technology to lower the recovery time objective.
Chapter 4, "FCIP over IP/MPLS Core."	Describes the transport of Fibre Channel over IP (FCIP) over IP/Multiprotocol Label Switching (MPLS) networks and addresses the network requirements from a service provider (SP) perspective.
Chapter 5, "Extended Ethernet Segments over the WAN/MAN using EoMPLS."	Describes the various options available to extend a Layer 2 network using Ethernet over Multiprotocol Label Switching (EoMPLS) on the Cisco Sup720-3B.
Chapter 6, "Metro Ethernet Services."	Describes the functional characteristics of Metro Ethernet services.
Appendix A "Configurations for Layer 2 Extension with EoMPLS."	Describes the lab and test setups.

Data Center High Availability Clusters

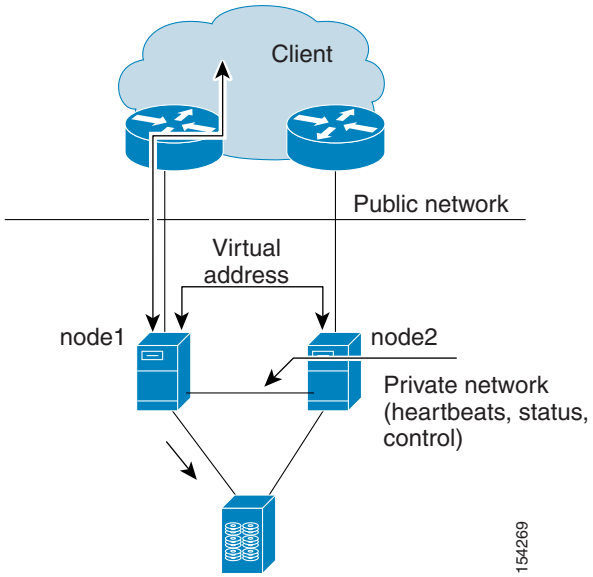
High Availability Clusters Overview

Clusters define a collection of servers that operate as if they were a single machine. The primary purpose of high availability (HA) clusters is to provide uninterrupted access to data, even if a server loses network or storage connectivity, or fails completely, or if the application running on the server fails.

HA clusters are mainly used for e-mail and database servers, and for file sharing. In their most basic implementation, HA clusters consist of two server machines (referred to as "nodes") that "share" common storage. Data is saved to this storage, and if one node cannot provide access to it, the other node can take client requests. Figure 1-1 shows a typical two node HA cluster with the servers connected to a shared storage (a disk array). During normal operation, only one server is processing client requests and has access to the storage; this may vary with different vendors, depending on the implementation of clustering.

HA clusters can be deployed in a server farm in a single physical facility, in different facilities at various distances for added resiliency. The latter type of cluster is often referred to as a *geocluster*.

Figure 1-1 Basic HA Cluster



Geoclusters are becoming very popular as a tool to implement business continuance. Geoclusters improve the time that it takes for an application to be brought online after the servers in the primary site become unavailable. In business continuance terminology, geoclusters combine with disk-based replication to offer better recovery time objective (RTO) than tape restore or manual migration.

HA clusters can be categorized according to various parameters, such as the following:

- How hardware is shared (shared nothing, shared disk, shared everything)
- At which level the system is clustered (OS level clustering, application level clustering)
- Applications that can be clustered
- Quorum approach
- Interconnect required

One of the most relevant ways to categorize HA clusters is how hardware is shared, and more specifically, how storage is shared. There are three main cluster categories:

- Clusters using mirrored disks—Volume manager software is used to create mirrored disks across all the machines in the cluster. Each server writes to the disks that it owns and to the disks of the other servers that are part of the same cluster.
- Shared nothing clusters—At any given time, only one node owns a disk. When a node fails, another node in the cluster has access to the same disk. Typical examples include IBM High Availability Cluster Multiprocessing (HACMP) and Microsoft Cluster Server (MSCS).
- Shared disk—All nodes have access to the same storage. A locking mechanism protects against race conditions and data corruption. Typical examples include IBM Mainframe Sysplex technology and Oracle Real Application Cluster.

Technologies that may be required to implement shared disk clusters include a *distributed volume manager,* which is used to virtualize the underlying storage for all servers to access the same storage; and the *cluster file system,* which controls read/write access to a single file system on the shared SAN.

More sophisticated clustering technologies offer shared-everything capabilities, where not only the file system is shared, but memory and processors, thus offering to the user a *single system image (SSI)*. In this model, applications do not need to be cluster-aware. Processes are launched on any of the available processors, and if a server/processor becomes unavailable, the process is restarted on a different processor.

The following list provides a partial list of clustering software from various vendors, including the architecture to which it belongs, the operating system on which it runs, and which application it can support:

- HP MC/Serviceguard—Clustering software for HP-UX (the OS running on HP Integrity servers and PA-RISC platforms) and Linux. HP Serviceguard on HP-UX provides clustering for Oracle, Informix, Sybase, DB2, Progress, NFS, Apache, and Tomcat. HP Serviceguard on Linux provides clustering for Apache, NFS, MySQL, Oracle, Samba, PostgreSQL, Tomcat, and SendMail. For more information, see the following URL: http://h71028.www7.hp.com/enterprise/cache/4189-0-0-0-121.html.
- HP NonStop computing—Provides clusters that run with the HP NonStop OS. NonStop OS runs on the HP Integrity line of servers (which uses Intel Itanium processors) and the NonStop S-series servers (which use MIPS processors). NonStop uses a shared nothing architecture and was developed by Tandem Computers. For more information, see the following URL: http://h20223.www2.hp.com/nonstopcomputing/cache/76385-0-0-0-121.aspx
- HP OpenVMS High Availability Cluster Service—This clustering solution was originally developed for VAX systems, and now runs on HP Alpha and HP Integrity servers. This is an OS-level clustering that offers an SSI. For more information, see the following URL: http://h71000.www7.hp.com/.

- HP TruCluster—Clusters for Tru64 UNIX (aka Digital UNIX). Tru64 Unix runs on HP Alpha servers. This is an OS-level clustering that offers an SSI. For more information, see the following URL: http://h30097.www3.hp.com/cluster/
- IBM HACMP—Clustering software for servers running AIX and Linux. HACMP is based on a shared nothing architecture. For more information, see the following URL: http://www-03.ibm.com/systems/p/software/hacmp.html
- MSCS—Belongs to the category of clusters that are referred to as shared nothing. MSCS can provide clustering for applications such as file shares, Microsoft SQL databases, and Exchange servers. For more information, see the following URL: http://www.microsoft.com/windowsserver2003/technologies/clustering/default.mspx
- Oracle Real Application Cluster (RAC) provides a shared disk solution that runs on Solaris, HP-UX, Windows, HP Tru64 UNIX, Linux, AIX, and OS/390. For more information about Oracle RAC 10g, see the following URL: http://www.oracle.com/technology/products/database/clustering/index.html
- Solaris SUN Cluster—Runs on Solaris and supports many applications including Oracle, Siebel, SAP, and Sybase. For more information, see the following URL: http://wwws.sun.com/software/cluster/index.html
- Veritas (now Symantec) Cluster Server—Veritas is a "mirrored disk" cluster. Veritas supports applications such as Microsoft Exchange, Microsoft SQL Databases, SAP, BEA, Siebel, Oracle, DB2, Peoplesoft, and Sybase. In addition to these applications you can create agents to support custom applications. It runs on HP-UX, Solaris, Windows, AIX, and Linux. For more information, see the following URL: http://www.veritas.com/us/products/clusterserver/prodinfo.html and http://www.veritas.com/Products/www?c=product&refId=20.

Note A single server can run several server clustering software packages to provide high availability for different server resources.

Note For more information about the performance of database clusters, see the following URL: http://www.tpc.org

Clusters can be "stretched" to distances beyond the local data center facility to provide metro or regional clusters. Virtually any cluster software can be configured to run as a *stretch cluster*, which means a cluster at metro distances. Vendors of cluster software often offer a geoclusters version of their software that has been specifically designed to have no intrinsic distance limitations. Examples of geoclustering software include the following:

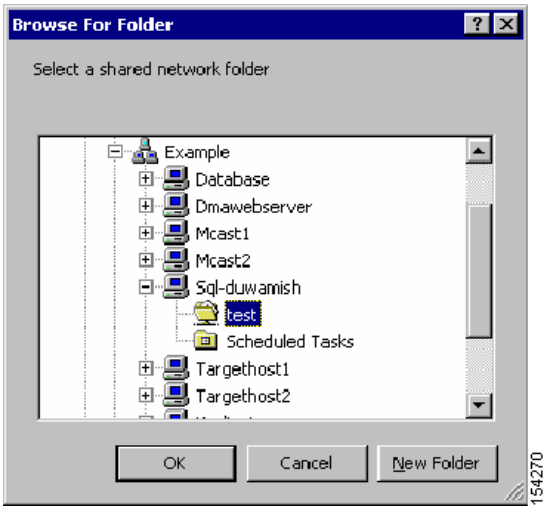
- EMC Automated Availability Manager Data Source (also called AAM)—This HA clustering solution can be used for both local and geographical clusters. It supports Solaris, HP-UX, AIX, Linux, and Windows. AAM supports several applications including Oracle, Exchange, SQL Server, and Windows services. It supports a wide variety of file systems and volume managers. AAM supports EMC SRDF/S and SRDF/A storage-based replication solutions. For more information, see the following URL: http://www.legato.com/products/autostart/
- Oracle Data Guard—Provides data protection for databases situated at data centers at metro, regional, or even continental distances. It is based on redo log shipping between active and standby databases. For more information, see the following URL: http://www.oracle.com/technology/deploy/availability/htdocs/DataGuardOverview.html
- Veritas (now Symantec) Global Cluster Manager—Allows failover from local clusters in one site to a local cluster in a remote site. It runs on Solaris, HP-UX, and Windows. For more information, see the following URL: http://www.veritas.com/us/products/gcmanager/

- HP Continental Cluster for HP-UX—For more information, see the following URL: http://docs.hp.com/en/B7660-90013/index.html
- IBM HACMP/XD (Extended Distance)—Available with various data replication technology combinations such as HACMP/XD Geographic Logical Volume Manager (GLVM) and HACMP/XD HAGEO replication for geographical distances. For more information, see the following URL: http://www-03.ibm.com/servers/systems/p/ha/disaster_tech.html

HA Clusters Basics

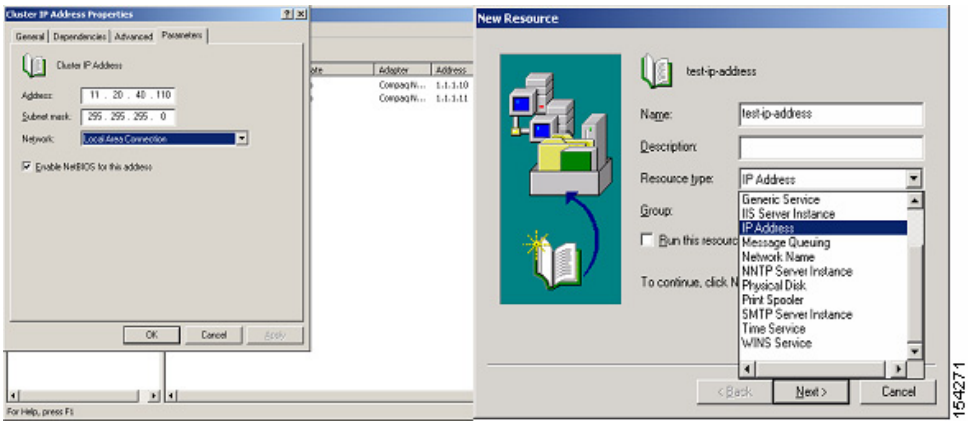
HA clusters are typically made of two servers such as the configuration shown in Figure 1-1. One server is actively processing client requests, while the other server is monitoring the main server to take over if the primary one fails. When the cluster consists of two servers, the monitoring can happen on a dedicated cable that interconnects the two machines, or on the network. From a client point of view, the application is accessible via a name (for example, a DNS name), which in turn maps to a virtual IP address that can float from a machine to another, depending on which machine is active. Figure 1-2 shows a clustered file-share.

Figure 1-2 Client Access to a Clustered Application—File Share Example



In this example, the client sends requests to the machine named “sql-duwamish”, whose IP address is a virtual address, which could be owned by either node1 or node2. The left of Figure 1-3 shows the configuration of a cluster IP address. From the clustering software point of view, this IP address appears as a monitored resource and is tied to the application, as described in Concept of Group, page 1-7. In this case, the IP address for the “sql-duwamish” is 11.20.40.110, and is associated with the clustered application “shared folder” called “test”.

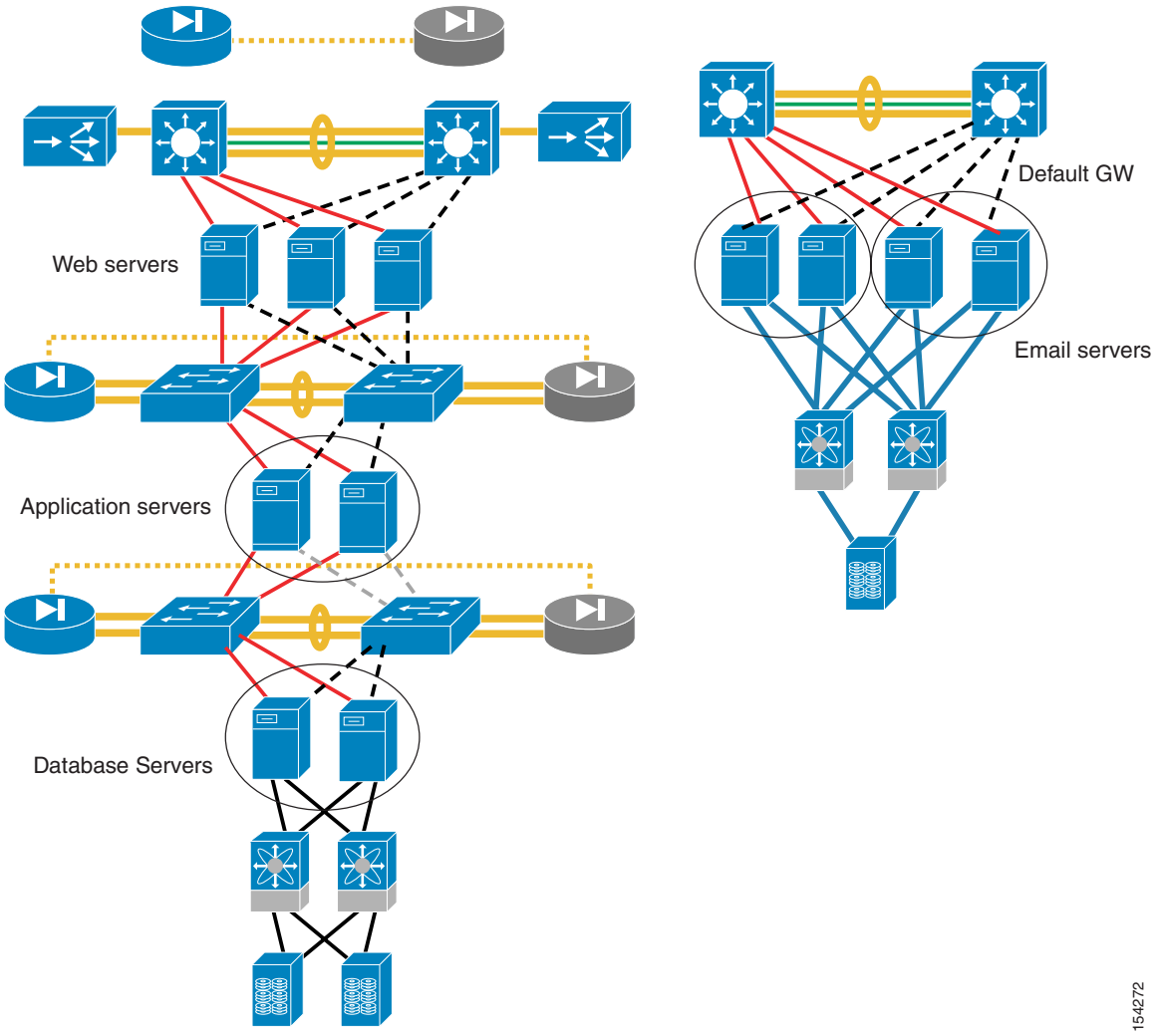
Figure 1-3 Virtual Address Configuration with MSCS



HA Clusters in Server Farms

Figure 1-4 shows where HA clusters are typically deployed in a server farm. Databases are typically clustered to appear as a single machine to the upstream web/application servers. In multi-tier applications such as a J2EE based-application and Microsoft .NET, this type of cluster is used at the very bottom of the processing tiers to protect application data.

Figure 1-4 HA Clusters Use in Typical Server Farms



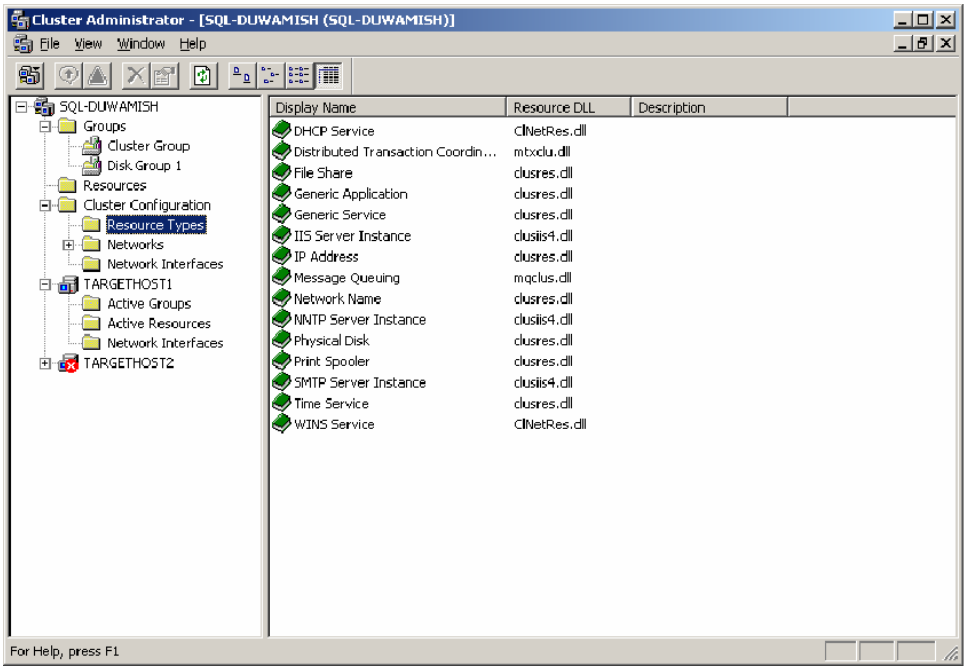
Applications

An application running on a server that has clustering software installed does not mean that the application is going to benefit from the clustering. Unless an application is cluster-aware, an application process crashing does not necessarily cause a failover to the process running on the redundant machine. Similarly, if the public network interface card (NIC) of the main machine fails, there is no guarantee that the application processing will fail over to the redundant server. For this to happen, you need an application that is cluster-aware.

Each vendor of cluster software provides immediate support for certain applications. For example, Veritas provides enterprise agents for the SQL Server and Exchange, among others. You can also develop your own agent for other applications. Similarly, EMC AAM provides application modules for Oracle, Exchange, SQL Server, and so forth.

In the case of MSCS, the cluster service monitors all the resources by means of the Resource Manager, which monitors the state of the application via the "Application DLL". By default, MSCS provides support for several application types, as shown in Figure 1-5. For example, MSCS monitors a clustered SQL database by means of the distributed transaction coordinator DLL.

Figure 1-5 Example of Resource DLL from MSCS

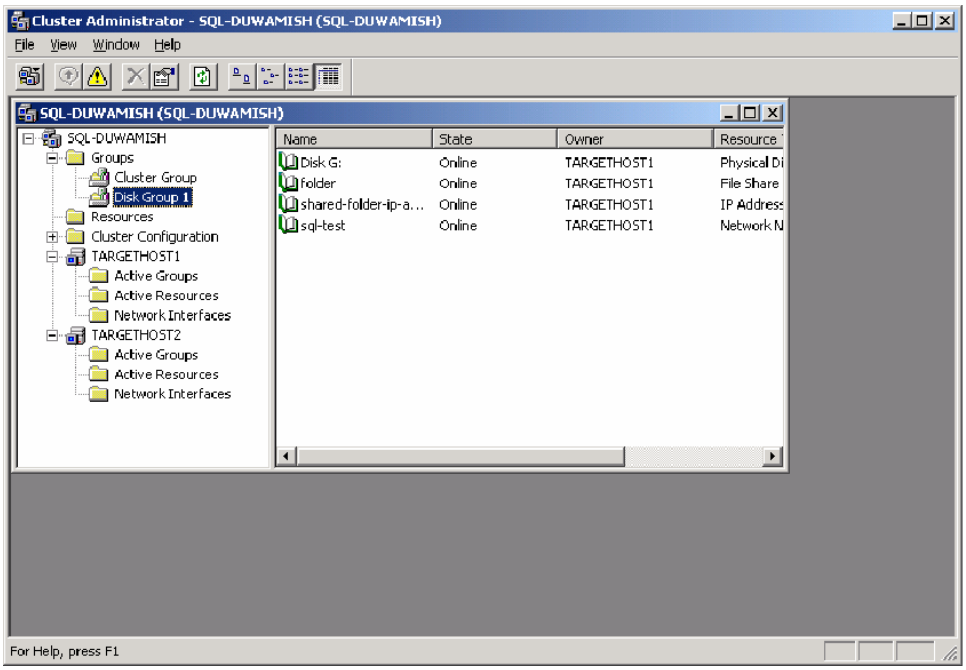


It is not uncommon for a server to run several clustering applications. For example, you can run one software program to cluster a particular database, another program to cluster the file system, and still another program to cluster a different application. It is out of the scope of this document to go into the details of this type of deployment, but it is important to realize that the network requirements of a clustered server might require considering not just one but multiple clustering software applications. For example, you can deploy MSCS to provide clustering for an SQL database, and you might also install EMC SRDF Cluster Enabler to failover the disks. The LAN communication profile of the MSCS software is different than the profile of the EMC SRDF CE software.

Concept of Group

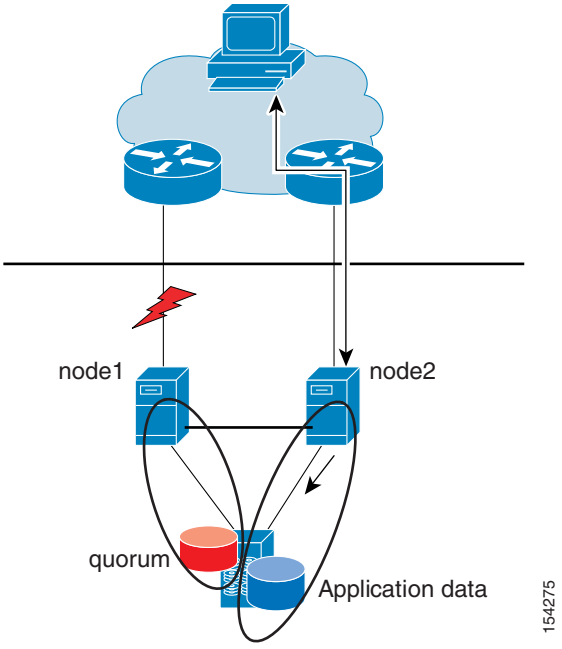
One key concept with clusters is the *group*. The group is a unit of failover; in other words, it is the bundling of all the resources that constitute an application, including its IP address, its name, the disks, and so on. Figure 1-6 shows an example of the grouping of resources: the "shared folder" application, its IP address, the disk that this application uses, and the network name. If any one of these resources is not available, for example if the disk is not reachable by this server, the group fails over to the redundant machine.

Figure 1-6 Example of Group



The failover of a group from one machine to another one can be automatic or manual. It happens automatically when a key resource in the group fails. Figure 1-7 shows an example: when the NIC on node1 goes down, the application group fails over to node2. This is shown by the fact that after the failover, node2 owns the disk that stores the application data. When a failover happens, node2 mounts the disk and starts the application by using the API provided by the Application DLL.

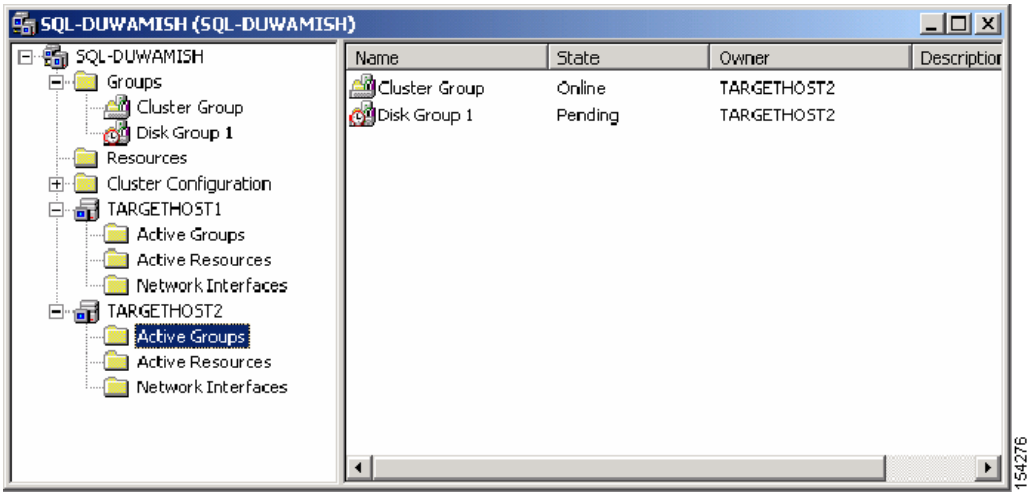
Figure 1-7 Failover of Group



The failover can also be manual, in which case it is called a *move*. Figure 1-8 shows a group (DiskGroup1) failing over to a node or "target2" (see the *owner* of the group), either as the result of a move or as the result of a failure.

After the failover or move, nothing changes from the client perspective. The only difference is that the machine that receives the traffic is node2 or target2, instead of node1 (or target1, as it is called in these examples).

Figure 1-8 Move of a Group



LAN Communication

The LAN communication between the nodes of a cluster obviously depends on the software vendor that provides the clustering function. As previously stated, to assess the network requirements, it is important to know all the various software components running on the server that are providing clustering functions.

Virtual IP Address

The virtual IP address (VIP) is the floating IP address associated with a given application or *group*. Figure 1-3 shows the VIP for the clustered shared folder (that is, DiskGroup1 in the group configuration). In this example, the VIP is 11.20.40.110. The physical address for node1 (or target1) could be 11.20.40.5, and the address for node2 could be 11.20.40.6. When the VIP and its associated group are active on node1, when traffic comes into the public network VLAN, either router uses ARP to determine the VIP and node1 answer. When the VIP *moves* or *fails over* to node2, then node2 answers the ARP requests from the routers.

Note From this description, it appears that the two nodes that form the cluster need to be part of the same subnet, because the VIP address stays the same after a failover. This is true for most clusters, except when they are geographically connected, in which case certain vendors allow solutions where the IP address can be different at each location, and the DNS resolution process takes care of mapping incoming requests to the new address.

The following trace helps explaining this concept:

```
11.20.40.6    11.20.40.1    ICMP      Echo (ping) request
11.20.40.1    11.20.40.6    ICMP      Echo (ping) reply
11.20.40.6    Broadcast     ARP       Who has 11.20.40.110?  Tell 11.20.40.6
11.20.40.6    Broadcast     ARP       Who has 11.20.40.110?  Gratuitous ARP
```

When 11.20.40.5 fails, 11.20.40.6 detects this by using the heartbeats, and then verifies its connectivity to 11.20.40.1. It then announces its MAC address, sending out a gratuitous ARP that indicates that 11.20.40.110 has moved to 11.20.40.6.

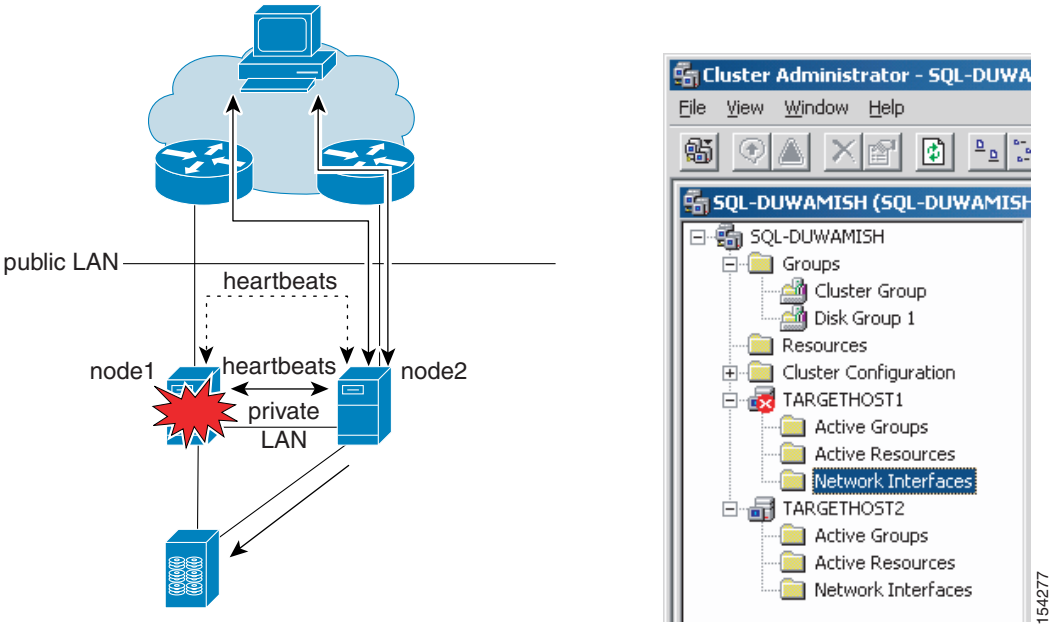
Public and Private Interface

As previously mentioned, the nodes in a cluster communicate over a public and a private network. The public network is used to receive client requests, while the private network is mainly used for monitoring. Node1 and node2 monitor the health of each other by exchanging heartbeats on the private network. If the private network becomes unavailable, they can use the public network. You can have more than one private network connection for redundancy. Figure 1-1 shows the public network, and a direct connection between the servers for the private network. Most deployments simply use a different VLAN for the private network connection.

Alternatively, it is also possible to use a single LAN interface for both public and private connectivity, but this is not recommended for redundancy reasons.

Figure 1-9 shows what happens when node1 (or target1) fails. Node2 is monitoring node1 and does not hear any heartbeats, so it declares target1 failed (see the right side of Figure 1-9). At this point, the client traffic goes to node2 (target2).

Figure 1-9 Public and Private Interface and a Failover



Heartbeats

From a network design point of view, the type of heartbeats used by the application often decide whether the connectivity between the servers can be routed. For *local* clusters, it is almost always assumed that the two or more servers communicate over a Layer 2 link, which can be either a direct cable or simply a VLAN.

The following traffic traces provide a better understanding of the traffic flows between the nodes:

```
1.1.1.11 1.1.1.10    UDP      Source port: 3343  Destination port: 3343
1.1.1.10 1.1.1.11    UDP      Source port: 3343  Destination port: 3343
1.1.1.11 1.1.1.10    UDP      Source port: 3343  Destination port: 3343
1.1.1.10 1.1.1.11    UDP      Source port: 3343  Destination port: 3343
```

1.1.1.10 and 1.1.1.11 are the IP addresses of the servers on the private network. This traffic is unicast. If the number of servers is greater or equal to three, the heartbeat mechanism typically changes to multicast. The following is an example of how the server-to-server traffic might appear on either the public or the private segment:

```
11.20.40.5  239.255.240.185  UDP  Source port: 3343  Destination port: 3343
11.20.40.6  239.255.240.185  UDP  Source port: 3343  Destination port: 3343
11.20.40.7  239.255.240.185  UDP  Source port: 3343  Destination port: 3343
```

The 239.255.x.x range is the site local scope. A closer look at the payload of these UDP frames reveals that the packet has a time-to-live (TTL)=1:

```
Internet Protocol, Src Addr: 11.20.40.5 (11.20.40.5), Dst Addr: 239.255.240.185
(239.255.240.185)
[…]
Fragment offset: 0
Time to live: 1
Protocol: UDP (0x11)
Source: 11.20.40.5 (11.20.40.5)
Destination: 239.255.240.185 (239.255.240.185)
```

The following is another possible heartbeat that you may find:

```
11.20.40.5  224.0.0.127  UDP Source port: 23  Destination port: 23
11.20.40.5  224.0.0.127  UDP Source port: 23  Destination port: 23
11.20.40.5  224.0.0.127  UDP Source port: 23  Destination port: 23
```

The 224.0.0.127 address belongs to the link local address range, which is generated with TTL=1.

These traces show that the private network connectivity between nodes in a cluster typically requires Layer 2 adjacency between the nodes; in other words, a non-routed VLAN. The Design chapter outlines options where routing can be introduced between the nodes when certain conditions are met.

Layer 2 or Layer 3 Connectivity

Based on what has been discussed in Virtual IP Address, page 1-9 and Heartbeats, page 1-11, you can see why Layer 2 adjacency is required between the nodes of a local cluster. The documentation from the cluster software vendors reinforces this concept.

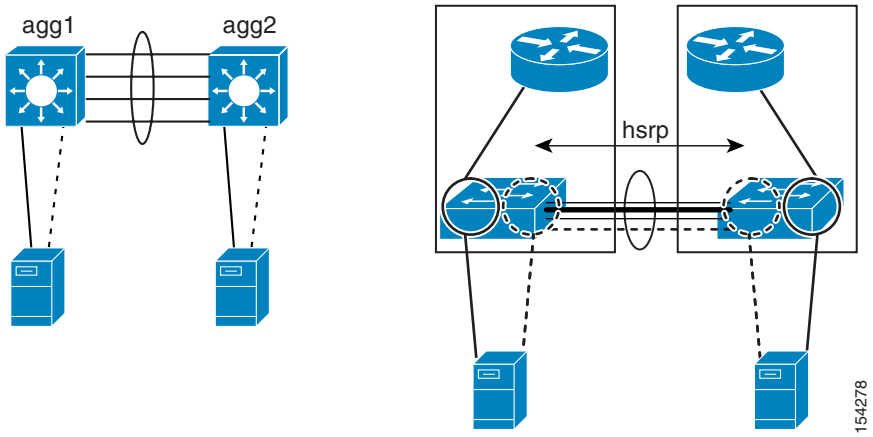
Quoting from the IBM HACMP documentation: “Between cluster nodes, do not place intelligent switches, routers, or other network equipment that do not transparently pass through UDP broadcasts and other packets to all cluster nodes. This prohibition includes equipment that optimizes protocol such as Proxy ARP and MAC address caching, transforming multicast and broadcast protocol requests into unicast requests, and ICMP optimizations.”

Quoting from the MSCS documentation: "The private and public network connections between cluster nodes must appear as a single, non-routed LAN that uses technologies such as virtual LANs (VLANs). In these cases, the connections network must be able to provide a guaranteed, maximum round-trip latency between nodes of no more than 500 milliseconds. The Cluster Interconnect must appear as a standard LAN". For more information, see the following URL: http://support.microsoft.com/kb/280743/EN-US/. According to Microsoft, future releases might address this restriction to allow building clusters across multiple L3 hops.

Note Some Cisco technologies can be used in certain cases to introduce Layer 3 hops in between the nodes. An example is a feature called Local Area Mobility (LAM). LAM works for unicast traffic only and it does not necessarily satisfy the requirements of the software vendor because it relies on Proxy ARP.

As a result of this requirement, most cluster networks are currently similar to those shown in Figure 1-10; to the left is the physical topology, to the right the logical topology and VLAN assignment. The continuous line represents the public VLAN, while the dotted line represents the private VLAN segment. This design can be enhanced when using more than one NIC for the private connection. For more details, see Complete Design, page 1-22.

Figure 1-10 Typical LAN Design for HA Clusters



Disk Considerations

Figure 1-7 displays a typical failover of a group. The disk ownership is moved from node1 to node2. This procedure requires that the disk be shared between the two nodes, such that when node2 becomes active, it has access to the same data as node1. Different clusters provide this functionality differently: some clusters follow a *shared disk* architecture where every node can write to every disk (and a sophisticated lock mechanism prevents inconsistencies which could arise from concurrent access to the same data), or *shared nothing*, where only one node owns a given disk at any given time.

Shared Disk

With either architecture (shared disk or shared nothing), from a storage perspective, the disk needs to be connected to the servers in a way that any server in the cluster can access it by means of a simple software operation.

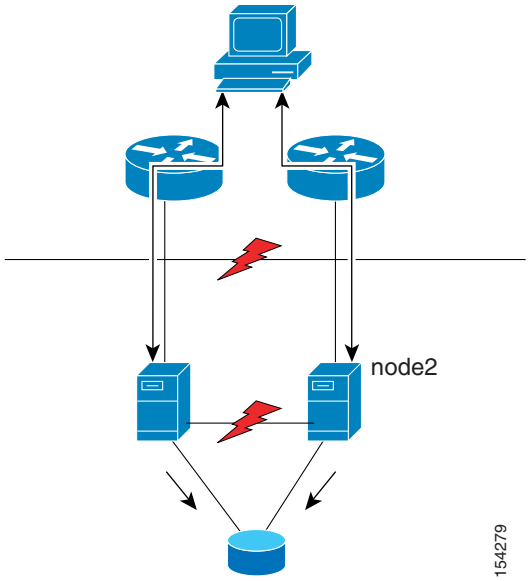
The disks to which the servers connect are typically protected with redundant array of independent disks (RAID): RAID1 at a minimum, or RAID01 or RAID10 for higher levels of I/O. This approach minimizes the chance of losing data when a disk fails as the disk array itself provides disk redundancy and data mirroring.

You can provide access to shared data also with a shared SCSI bus, network access server (NAS), or even with iSCSI.

Quorum Concept

Figure 1-11 shows what happens if all the communication between the nodes in the cluster is lost. Both nodes bring the same group online, which results in an active-active scenario. Incoming requests go to both nodes, which then try to write to the shared disk, thus causing data corruption. This is commonly referred to as the *split-brain* problem.

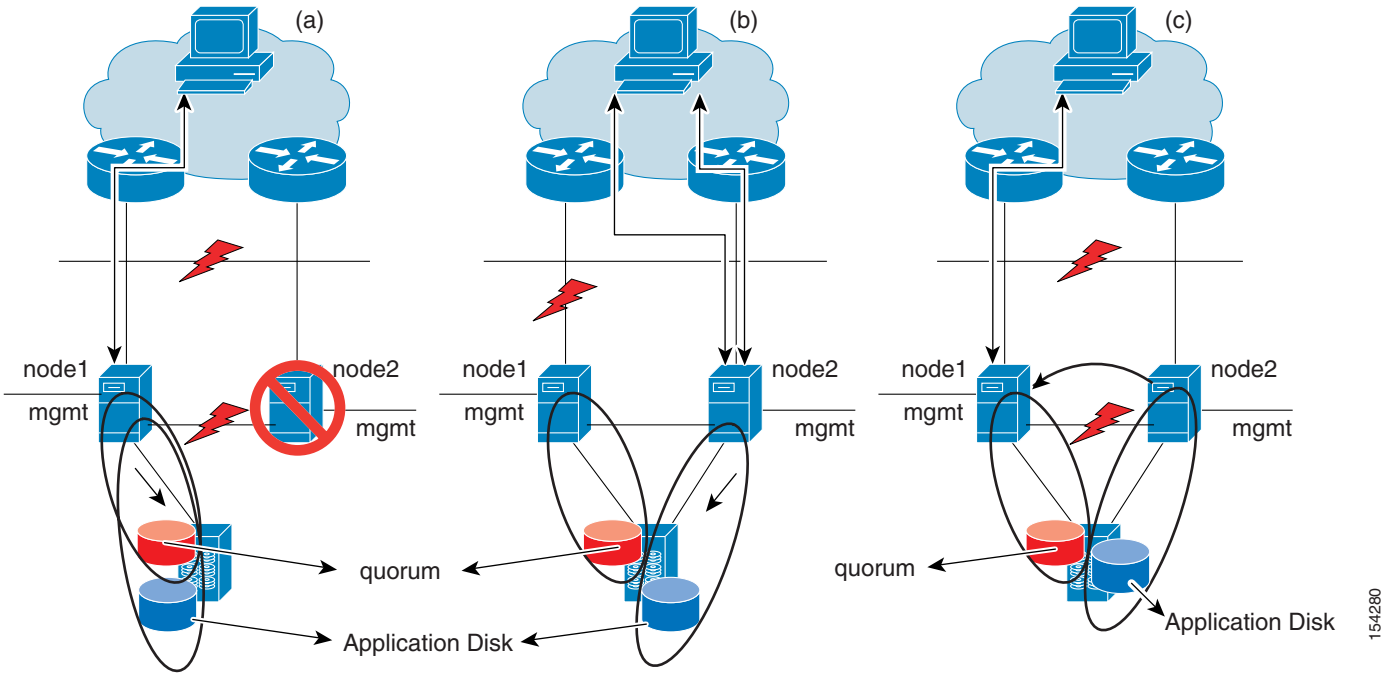
Figure 1-11 Theoretical Split-Brain Scenario



The mechanism that protects against this problem is the *quorum*. For example, MSCS has a *quorum disk* that contains the database with the cluster configuration information and information on all the objects managed by the clusters.

Only one node in the cluster owns the quorum at any given time. Figure 1-12 shows various failure scenarios where despite the fact that the nodes in the cluster are completely isolated, there is no data corruption because of the quorum concept.

Figure 1-12 LAN Failures in Presence of Quorum Disk



In scenario (a), node1 owns the quorum and that is also where the group for the application is active. When the communication between node1 and node2 is cut, nothing happens; node2 tries to reserve the quorum, but it cannot because the quorum is already owned by node1.

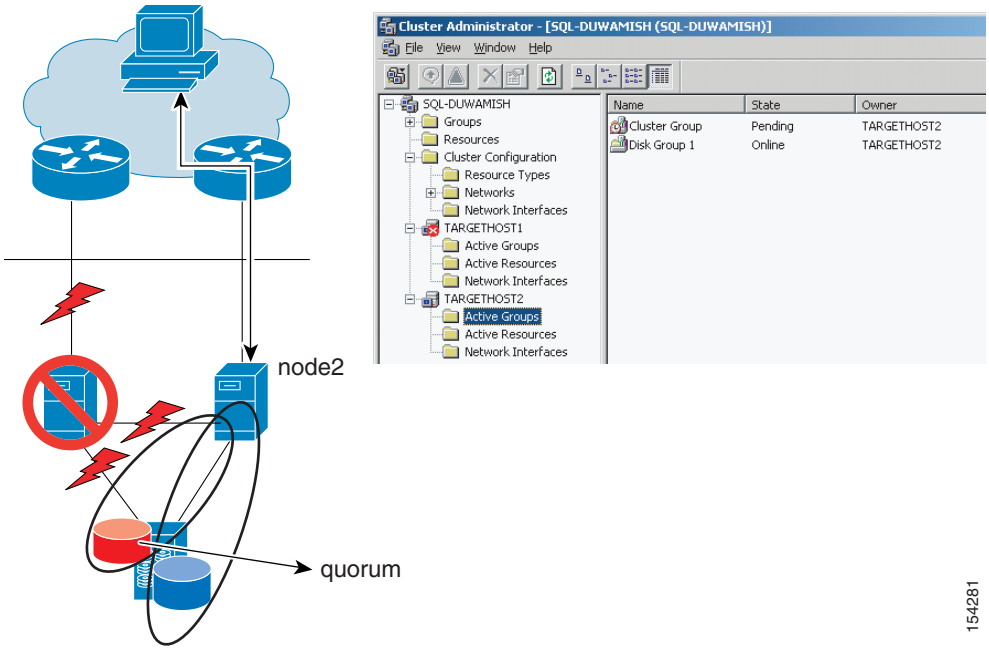
Scenario (b) shows that when node1 loses communication with the public VLAN, which is used by the *application group,* it can still communicate with node2 and instruct node2 to take over the disk for the application group. This is because node2 can still talk to the default gateway. For management purposes, if the quorum disk as part of the *cluster group* is associated with the public interface, the quorum disk can also be transferred to node2, but it is not necessary. At this point, client requests go to node2 and everything works.

Scenario (c) shows what happens when the communication is lost between node1 and node2 where node2 owns the application group. Node1 owns the quorum, thus it can bring resources online, so the application group is brought up on node1.

The key concept is that when all communication is lost, the node that owns the quorum is the one that can bring resources online, while if partial communication still exists, the node that owns the quorum is the one that can initiate the move of an application group.

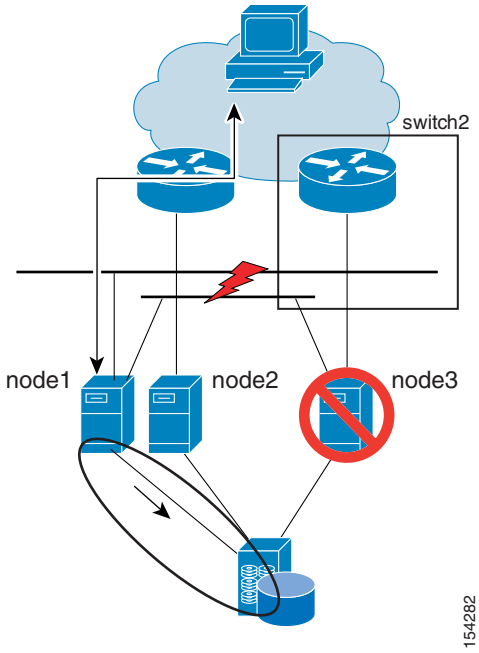
When all communication is lost, the node that does not own the quorum (referred to as the *challenger*) performs a SCSI reset to get ownership of the quorum disk. The owning node (referred to as the *defender*) performs SCSI reservation at the interval of 3s,and the challenger retries after 7s. As a result, if a node owns the quorum, it still holds it after the communication failure. Obviously, if the defender loses connectivity to the disk, the challenger can take over the quorum and bring all the resources online. This is shown in Figure 1-13.

Figure 1-13 Node1 Losing All Connectivity on LAN and SAN



There are several options related to which approach can be taken for the quorum implementation; the quorum disk is just one option. A different approach is the *majority node set*, where a copy of the quorum configuration is saved on the local disk instead of the shared disk. In this case, the arbitration for which node can bring resources online is based on being able to communicate with at least more than half of the nodes that form the cluster. Figure 1-14 shows how the majority node set quorum works.

Figure 1-14 Majority Node Set Quorum



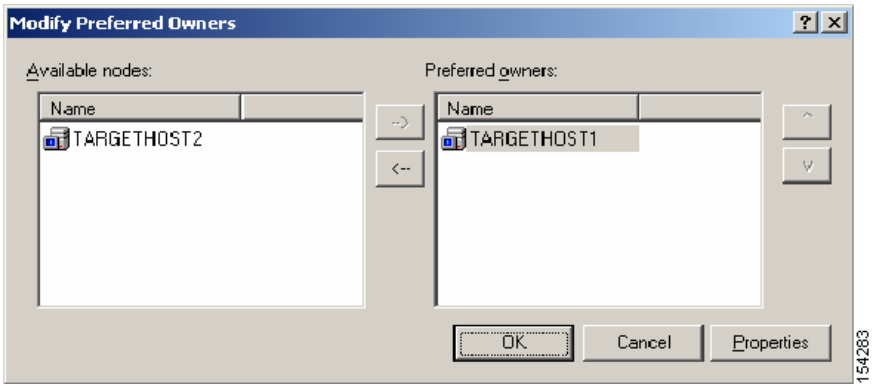
Each local copy of the quorum disk is accessed via the network by means of server message block (SMB). When nodes are disconnected, as in Figure 1-14, node1 needs to have the vote of the majority of the nodes to be the master. This implies that this design requires an odd number of nodes. Also notice that there is no quorum disk configured on the storage array.

Network Design Considerations

Routing and Switching Design

Figure 1-12 through Figure 1-14 show various failure scenarios and how the quorum concept helps prevent data corruption. As the diagrams show, it is very important to consider the implications of the routing configuration, especially when dealing with a geocluster (see subsequent section in this document). It is very important to match the routing configuration to ensure that the traffic enters the network from the router that matches the node that is preferred to own the quorum. By matching quorum and routing configuration, when there is no LAN connectivity, there is no chance that traffic is routed to the node whose resources are offline. Figure 1-15 shows how to configure the preferred owner for a given resource; for example, the quorum disk. This configuration needs to match the routing configuration.

Figure 1-15 Configuring the Preferred Owner for a Resource—Quorum Example

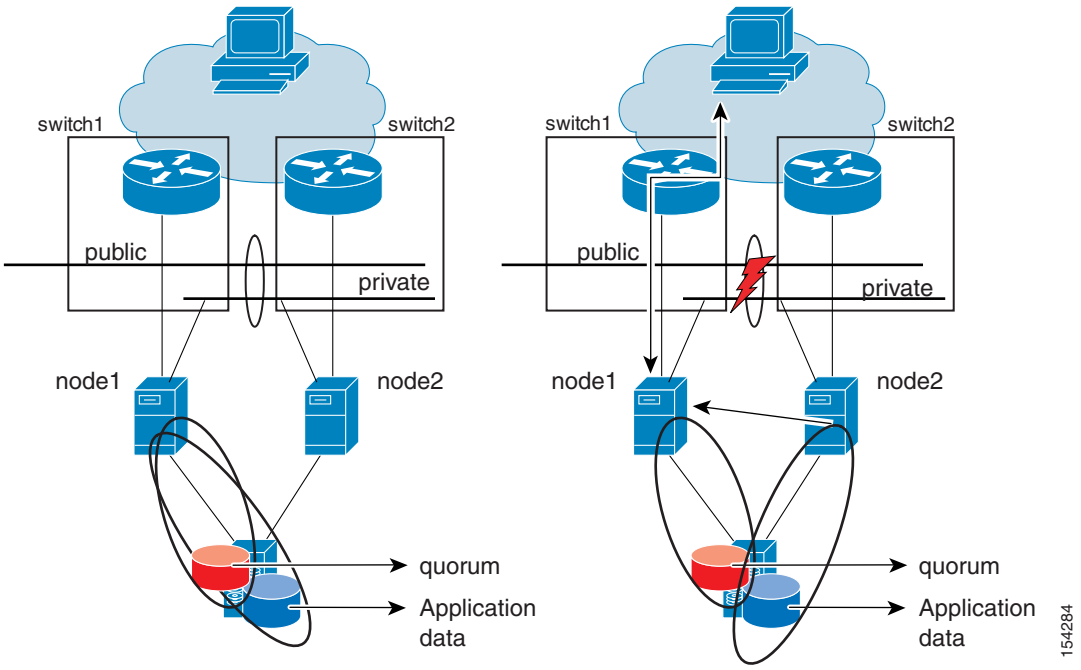


Controlling the inbound traffic from a routing point of view and matching the routing configuration to the quorum requires the following:

- Redistributing the connected subnets
- Filtering out the subnets where there are no clusters configured (this is done with route maps)
- Giving a more interesting cost to the subnets advertised by switch1

Figure 1-16 (a) shows a diagram with the details of how the public and private segment map to a typical topology with Layer 3 switches. The public and private VLANs are trunked on an EtherChannel between switch1 and switch2. With this topology, when the connectivity between switch1 and switch2 is lost, the nodes cannot talk with each other on either segment. This is actually preferable to having a LAN disconnect on only, for example, the public segment. The reason is that by losing both segments at the same time, the topology converges as shown in Figure 1-16 (b) no matter which node owned the disk group for the application.

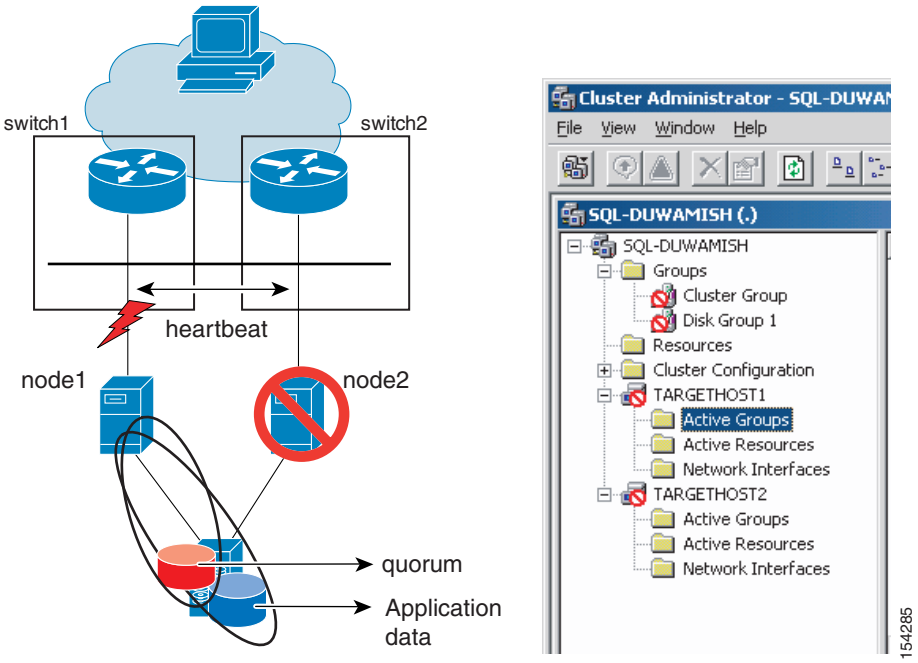
Figure 1-16 Typical Local Cluster Configuration



Importance of the Private Link

Figure 1-17 shows the configuration of a cluster where the public interface is used both for client-to-server connectivity and for the heartbeat/interconnect. This configuration does not protect against the failure of a NIC or of the link that connects node1 to the switch. This is because the node that owns the quorum cannot instruct the other node to take over the application group. The result is that both nodes in the cluster go offline.

Figure 1-17 Cluster Configuration with a Promiscuous Port—No Private Link



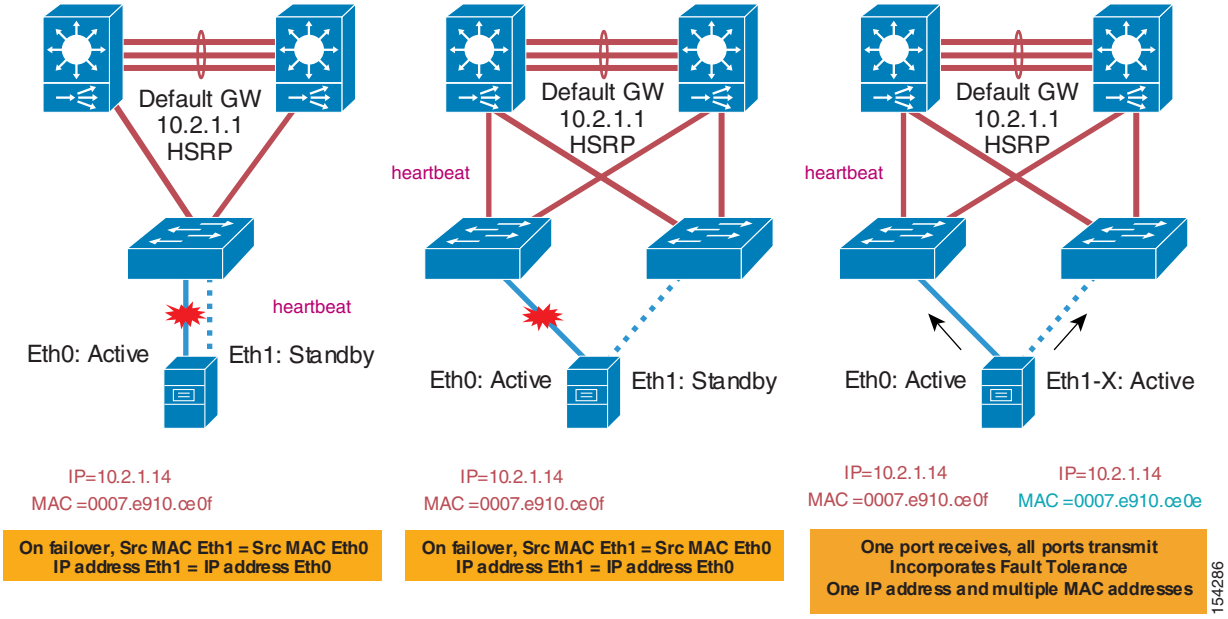
For this reason, Cisco highly recommends using at least two NICs; one for the public network and one for the private network, even if they both connect to the same switch. Otherwise, a single NIC failure can make the cluster completely unavailable, which is exactly the opposite of the purpose of the HA cluster design.

NIC Teaming

Servers with a single NIC interface can have many single points of failure, such as the NIC card, the cable, and the switch to which it connects. NIC teaming is a solution developed by NIC card vendors to eliminate this single point of failure by providing special drivers that allow two NIC cards to be connected to two different access switches or different line cards on the same access switch. If one NIC card fails, the secondary NIC card assumes the IP address of the server and takes over operation without disruption. The various types of NIC teaming solutions include active/standby and active/active. All solutions require the NIC cards to have Layer 2 adjacency with each other.

Figure 1-18 shows examples of NIC teaming configurations.

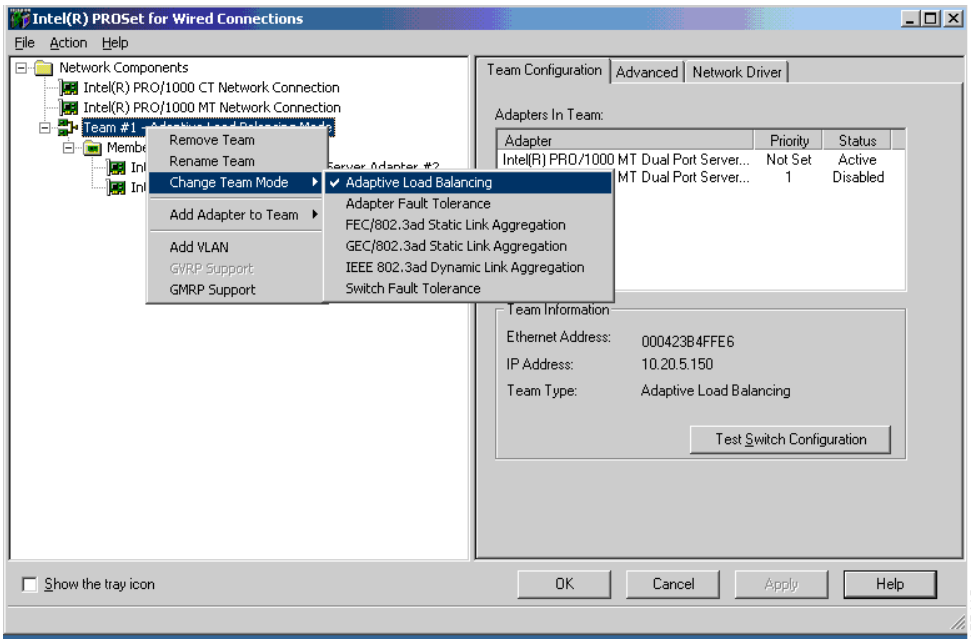
Figure 1-18 NIC Teaming Configurations



With Switch Fault Tolerance (SFT) designs, one port is active and the other is standby, using one common IP address and MAC address. With Adaptive Load Balancing (ALB) designs, one port receives and all ports transmit using one IP address and multiple MAC addresses.

Figure 1-19 shows an Intel NIC teaming software configuration where the user has grouped two interfaces (in this case from the same NIC) and has selected the ALB mode.

Figure 1-19 Typical NIC Teaming Software Configuration



Depending on the cluster server vendor, NIC teaming may or may not be supported. For example, in the case of MSCS, teaming is supported for the public-facing interface but not for the private interconnects. For this reason, it is advised to use multiple links for the private interconnect, as described at the following URL: http://support.microsoft.com/?id=254101.

Quoting from Microsoft: "Microsoft does not recommend that you use any type of fault-tolerant adapter or "Teaming" for the heartbeat. If you require redundancy for your heartbeat connection, use multiple network adapters set to Internal Communication Only and define their network priority in the cluster configuration. Issues have been seen with early multi-ported network adapters, so verify that your firmware and driver are at the most current revision if you use this technology. Contact your network adapter manufacturer for information about compatibility on a server cluster. For more information, see the following article in the Microsoft Knowledge Base: 254101 Network Adapter Teaming and Server Clustering."

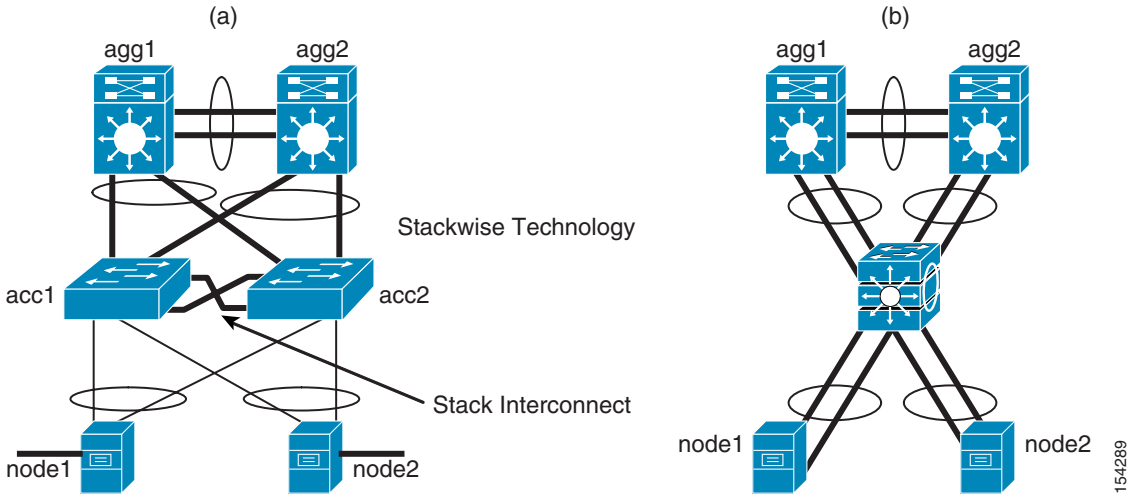
Another variation to the NIC teaming configuration consists in using *cross-stack EtherChannels*. For more information, see the following URL: http://www.cisco.com/en/US/docs/switches/lan/catalyst3750/software/release/12.2_25_sed/configuration/guide/swethchl.html.

Figure 1-20 (a) shows the network design with cross-stack EtherChannels. You need to use two or more Cisco Catalyst 3750 switches interconnected with the appropriate stack interconnect cable, as described at the following URL: http://www.cisco.com/en/US/docs/switches/lan/catalyst3750/software/release/12.2_25_sed/configuration/guide/swstack.html.

The aggregation switches are dual-connected to each stack member (access1 and access2); the servers are similarly dual-connected to each stack member. EtherChanneling is configured on the aggregation switches as well as the switch stack. Link Aggregation Protocol is not supported across switches, so the channel group must be configured in mode "on". This means that the aggregation switches also need to be configured with the channel group in mode on.

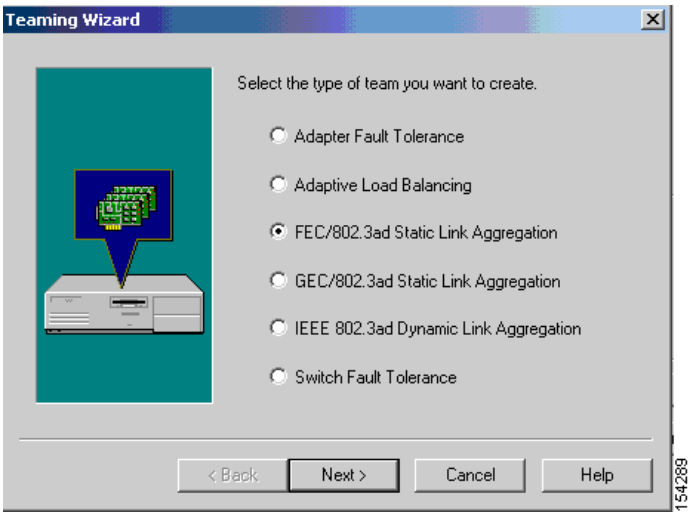
Figure 1-20 (b) shows the resulting equivalent topology to Figure 1-20 (a) where the stack of access switches appears as a single device to the eyes of the aggregation switches and the servers.

Figure 1-20 Configuration with Cross-stack EtherChannels



Configuration of the channeling on the server requires the selection of *Static Link Aggregation*; either FEC or GEC, depending on the type of NIC card installed, as shown in Figure 1-21.

Figure 1-21 Configuration of EtherChanneling on the Server Side



Compared with the ALB mode (or TLB, whichever name the vendor uses for this mechanism), this deployment has the advantage that all the server links are used both in the outbound and inbound direction, thus providing a more effective load balancing of the traffic. In terms of high availability, there is little difference with the ALB mode:

- With the stackwise technology, if one of the switches in the stack fails (for example the master), the remaining one takes over Layer 2 forwarding in 1s (see the following URL: http://www.cisco.com/en/US/docs/switches/lan/catalyst3750/software/release/12.2_25_sed/configuration/guide/swintro.html)

  The FEC or GEC configuration of the NIC teaming driver stops using the link connecting to the failed switch and continues on the remaining link.

- With an ALB configuration, when access1 fails, the teaming software simply forwards the traffic on the remaining link.

In both cases, the traffic drop amounts to few seconds.

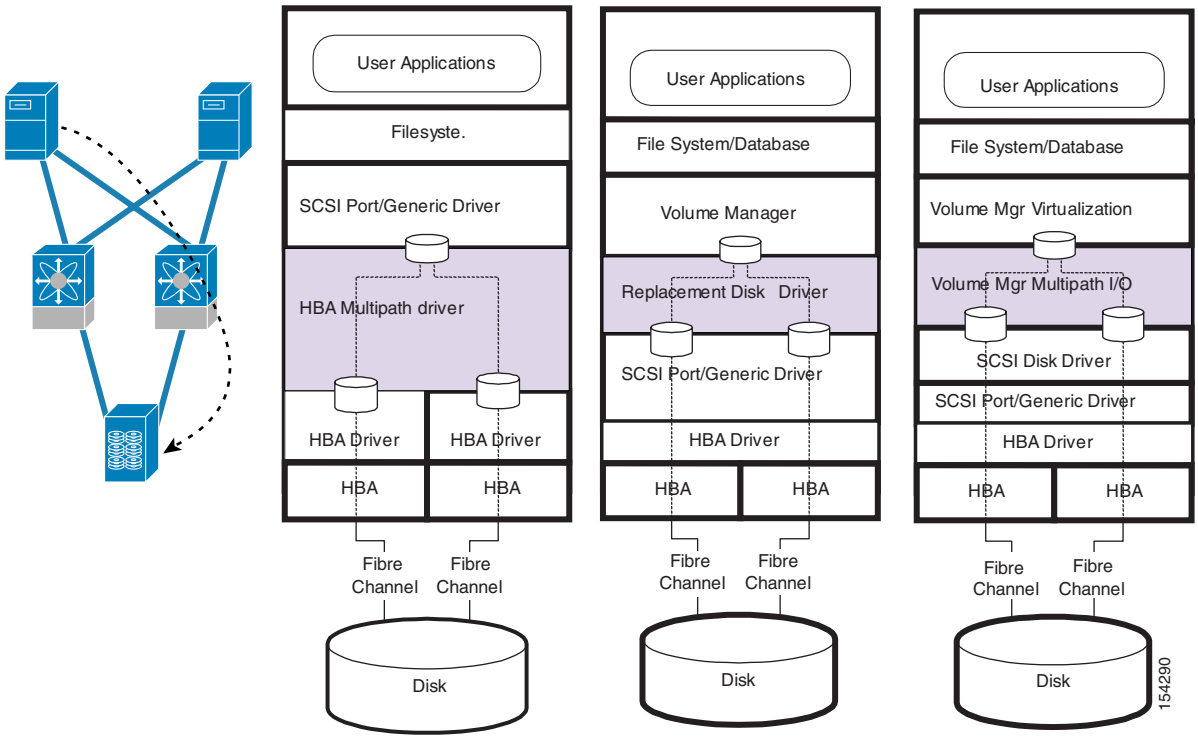
Storage Area Network Design

From a SAN point of view, the key requirement for HA clusters is that both nodes need to be able to see the same storage. Arbitration of which node is allowed to write to the disk happens at the cluster software level, as previously described in Quorum Concept, page 1-13.

HA clusters are often configured for multi-path I/O (MPIO) for additional redundancy. This means that each server is configured with two host-based adapters (HBAs) and connects to two fabrics. The disk array is in turn connected to each fabric. This means that each server has two paths to the same LUN. Unless special MPIO software is installed on the server, the server thinks that each HBA gives access to a different disk.

The MPIO software provides a single view of the disk via these two paths and load balancing between them. Two examples of this type of software include EMC Powerpath and HP Autopath. The MPIO software can be provided by HBA vendors, storage vendors, or by Volume Manager vendors. Each product operates in a different layer of the stack, as shown in Figure 1-22. Several mechanisms can be used by this software to identify the same disk that appears on two different HBAs.

Figure 1-22 SAN Configuration



MPIO can use several load distribution/HA algorithms: Active/Standby, Round Robin, Least I/O (referred to the path with fewer I/O requests), or Least Blocks (referred to the path with fewer blocks).

Not all MPIO software is compatible with clusters, because sometimes the locking mechanisms required by the cluster software cannot be supported with MPIO. To discover whether a certain cluster software is compatible with a specific MPIO solution, see the hardware and software compatibility matrix provided by the cluster vendor. As an example, in the case of Microsoft, see the following URLs:

- http://www.microsoft.com/whdc/hcl/search.mspx
- http://www.microsoft.com/windows2000/datacenter/HCL/default.asp
- http://www.microsoft.com/WindowsServer2003/technologies/storage/mpio/faq.mspx

Besides verifying the MPIO compatibility with the cluster software, it is also important to verify which mode of operation is compatible with the cluster. For example, it may be more likely that the active/standby configuration be compatible than the load balancing configurations.

Besides MPIO, the SAN configuration for cluster operations is fairly simple; you just need to configure zoning correctly so that all nodes in the cluster can see the same LUNs, and similarly on the storage array, LUN masking needs to present the LUNs to all the nodes in the cluster (if MPIO is present, the LUN needs to mapped to each port connecting to the SAN).

Complete Design

Figure 1-23 shows the end-to-end design with a typical data center network. Each clustered server is dual-homed to the LAN and to the SAN. NIC teaming is configured for the public interface; with this design, it might be using the ALB mode (also called TLB depending on the NIC vendor) to take

advantage of the forwarding uplinks of each access switch; MPIO is configured for storage access. The private connection is carried on a different port. If you require redundancy for the private connection, you would configure an additional one, without the two being teamed together.

Figure 1-23 Design Options with Looped Access with (b) being the Preferred Design

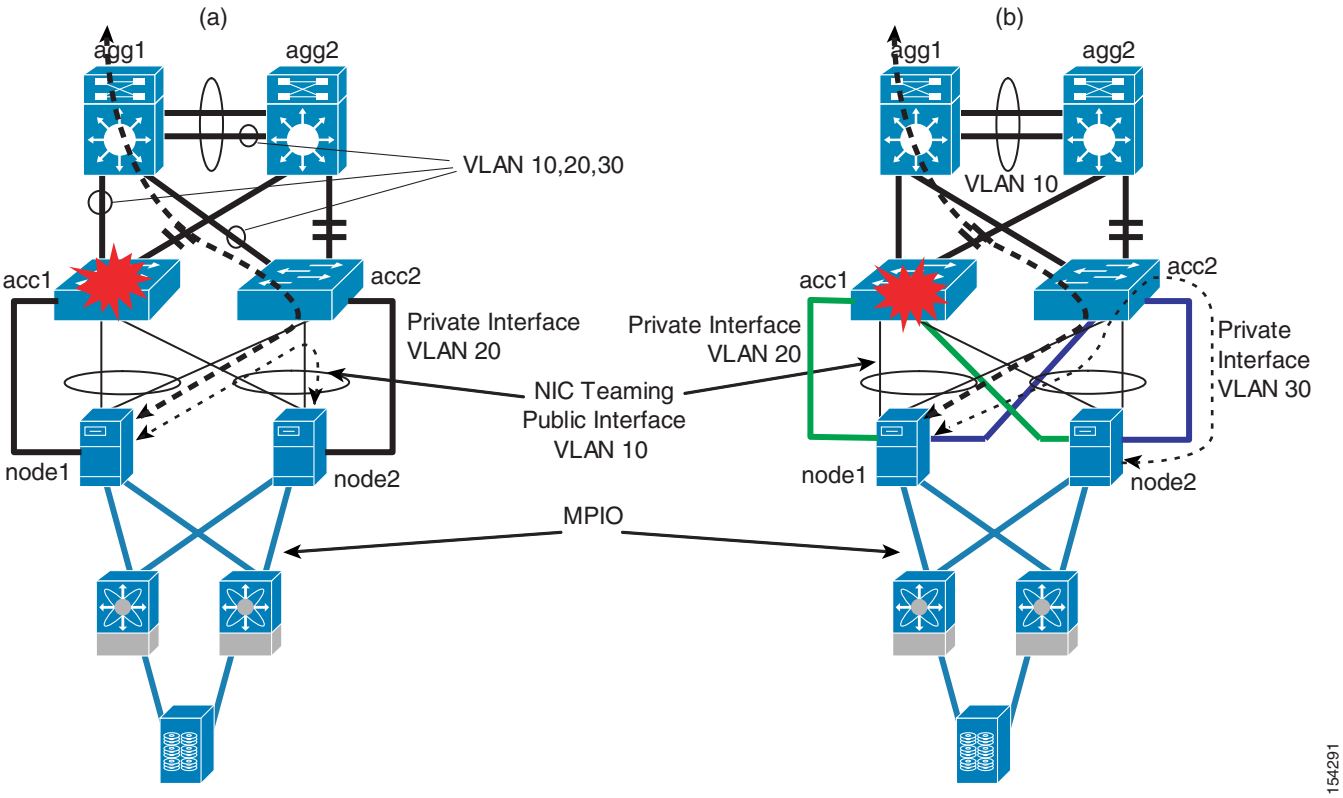
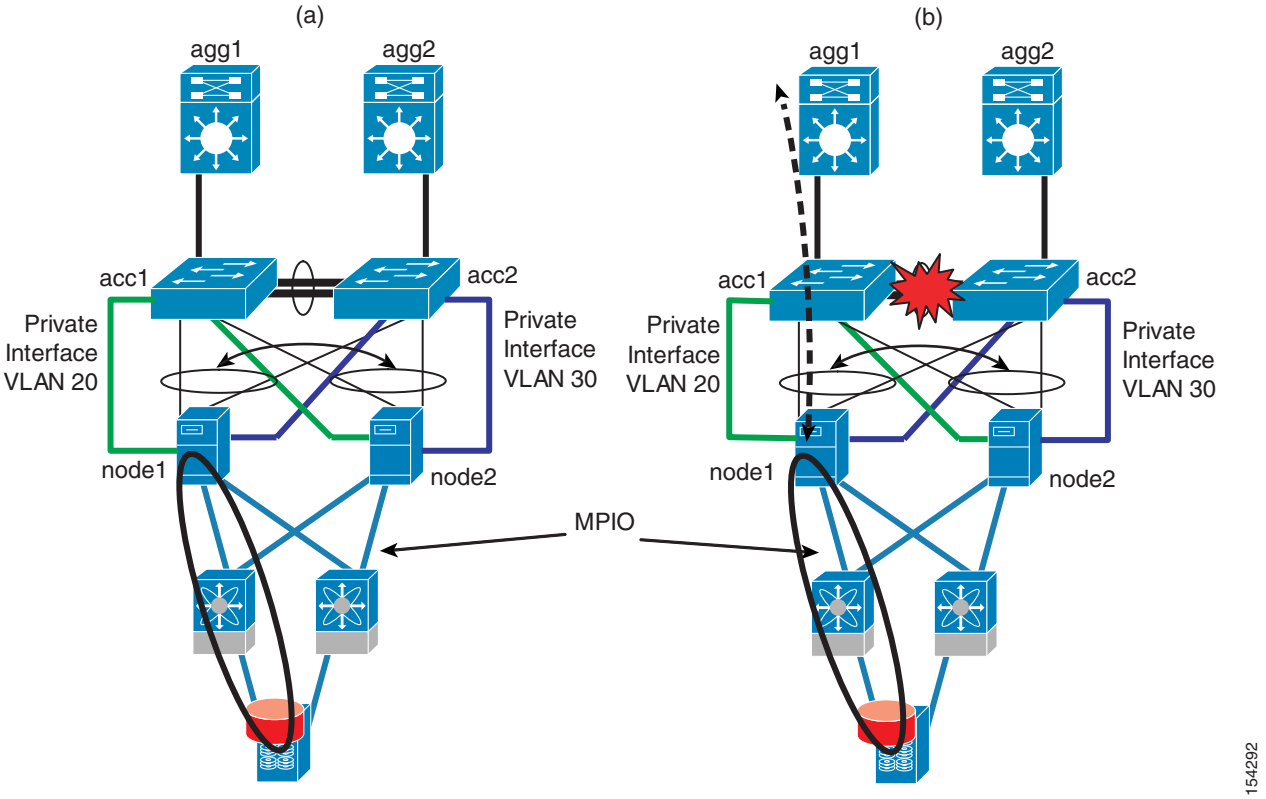


Figure 1-23 (a) shows a possible design where each server has a private connection to a single switch. This design works fine except when one of the two switches fails, as shown. In this case, the heartbeat (represented as the dash line in the picture) needs to traverse the remaining link in the teamed public interface. Depending on the clustering software vendor, this configuration might or might not work. As previously stated, Microsoft, for example, does not recommend carrying the heartbeat on a teamed interface. Figure 1-23 (b) shows a possible alternative design with redundancy on the private links. In this case, there are three VLANs: Vlan 10 for the public interface, and VLAN 20 and 30 for the private links. VLAN 20 is local to the access switch to the left and VLAN 30 is local to the access switch to the right. Each node has a private link to each access switch. In case one access switch fails, the heartbeat communication (represented as the dash line in the picture) continues on the private links connected to the remaining access switch.

Figure 1-24 (a) shows the design with a loop-free access.

Figure 1-24 Design with a Loop-free Access (a) and an Important Failure Scenario (b)



This design follows the same strategy as Figure 1-23 (b) for the private links. The teaming configuration most likely leverages Switch Fault Tolerance, because there is no direct link between the access switch to the right towards the left aggregation switch where HSRP is likely to be the primary. One important failure scenario is the one shown in Figure 1-24 (b) where the two access switches are disconnected, thus creating a split subnet. To address this problem and make sure that the cluster can continue to work, it may be a good design best practice to match the preferred owner for the quorum disk to the aggregation switch that advertises the path with the best metric. This configuration is not the normal default configuration for the aggregation switches/routers. You have to explicitly configure the routing in a way that aggregation1 is the preferred path to the cluster. This is achieved, for example, by using the command **redistribute connected** to filter out all the subnets except the cluster subnet, and by using route maps to assign a better cost to the route advertised by agg1 compared to the one advertised by agg2.

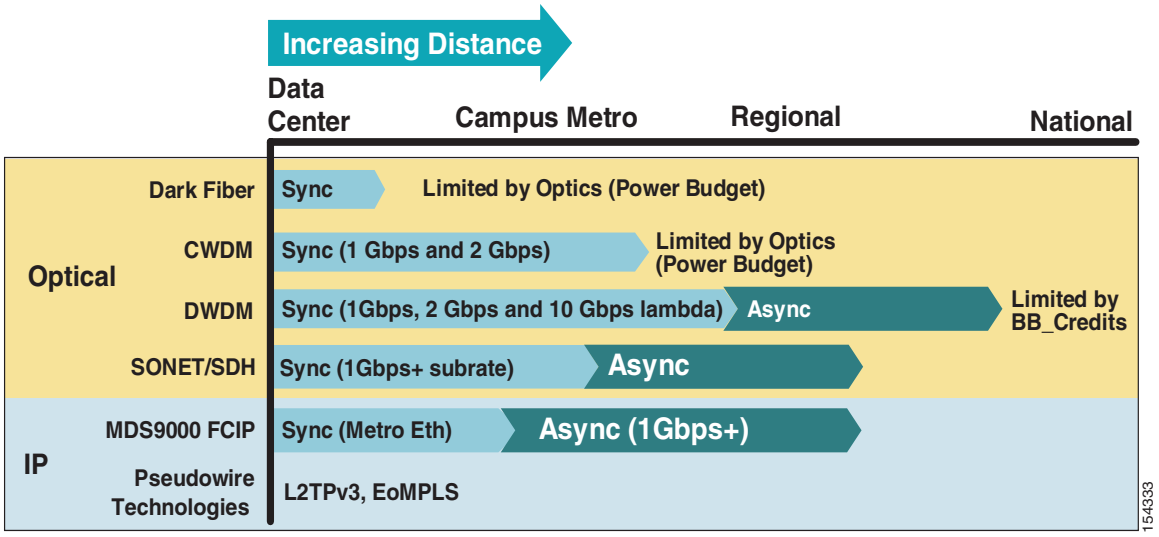
Data Center Transport Technologies

A wide variety of transport options for interconnecting the data centers provide various features and allow many different distances. Achievable distances depend on many factors such as the power budget of the optics, the lambda used for the transmission, the type of fiber, buffer-to-buffer credits, and so forth.

Before discussing some of the available technologies, it is important to consider the features of the LAN and SAN switches that provide higher availability for the data center interconnect. The required convergence time from the application that is going to use these features is also important.

Figure 2-1 shows the various transport technologies and distances.

Figure 2-1 Transport Technologies and Distances



Redundancy and Client Protection Technologies

EtherChanneling on the LAN switches and port channeling on the Cisco MDS Fibre Channel switches are two typical technologies that are used to provide availability and increased bandwidth from redundant fibers, pseudowires, or lambda.

EtherChannels allow you to bundle multiple ports for redundancy and/or increased bandwidth. Each switch connects to the other switch, with up to eight links bundled together as a single port with eight times the throughput capacity (if these are gigabit ports, an 8-Gigabit port results).

The following are benefits of channeling:

- Sub-second convergence for link failures—If you lose any of the links in the channel, the switch detects the failure and distributes the traffic on the remaining links.
- Increased bandwidth—Each port channel link has as much bandwidth as the sum of the bundled links.
- All links are active.

You can configure EtherChannels manually, or you can use Port Aggregation Protocol (PAgP) or Link Aggregation Control Protocol (LACP) to form EtherChannels. The EtherChannel protocols allow ports with similar characteristics to form an EtherChannel through dynamic negotiation with connected network devices. PAgP is a Cisco-proprietary protocol and LACP is defined in IEEE 802.3ad.

EtherChannel load balancing can use the following:

- MAC addresses or IP addresses
- Layer 4 port numbers
- Either source or destination, or both source and destination addresses or ports

The selected mode applies to all EtherChannels configured on the switch. EtherChannel load balancing can also use the Layer 4 port information. An EtherChannel can be configured to be an IEEE 802.1q trunk, thus carrying multiple VLANs.

For more information, see the following URL:
http://www.cisco.com/en/US/docs/switches/lan/catalyst6500/ios/12.2SXF/native/configuration/guide/channel.html.

When an EtherChannel link goes down, and there are at least *min-links* up (which by default is 1), the EtherChannel stays up, and spanning tree or the routing protocols running on top of the EtherChannel do not have to reconverge. The detection speed of the link failure is immediate if the devices are connected directly via a fiber or via an optical transport technology. The detection might take longer on a *pseudo-wire*.

Fibre Channel port channeling provides the ability to aggregate multiple physical inter-switch links (ISLs) into a logical ISL (up to 16 ports). The load sharing on the link members is based on source and destination ID (SID/DID) and exchange ID(SID/DID/OXID). If one link fails, the traffic is redistributed among the remaining member links in the channel and is transparent to the end applications. The *Port Channel* feature supports both E_port and TE_port modes, creating a virtual ISL or EISL that allows transporting multiple virtual storage area networks (VSANs).

When a port channel link goes down and at least one link within the channel group is still functional, there is no topology change in the fabric.

Dark Fiber

Dark fiber is a viable method for SAN extension over data center or campus distances. The maximum attainable distance is a function of the optical characteristics (transmit power and receive sensitivity) of the LED or laser that resides in a Small Form-Factor Pluggable (SFP) or Gigabit Interface Converter (GBIC) transponder, combined with the number of fiber joins, and the attenuation of the fiber. Lower cost MultiMode Fiber (MMF) with 850 nm SX SFPs/GBICs are used in and around data center rooms. SingleMode Fiber (SMF) with 1310 nm or 1550 nm SFPs/GBICs are used over longer distances.

Pluggable Optics Characteristics

The following list provides additional information about the wavelength and the distance achieved by various GigabitEthernet, 10 GigabitEthernet, Fibre Channel 1 Gbps, and Fibre Channel 2 Gbps GBICs and SFPs. For data center connectivity, the preferred version is obviously the long wavelength or extra long wavelength version.

- 1000BASE-SX GBIC and SFP—GigabitEthernet transceiver that transmits at 850 nm on MMF. The maximum distance is 550 m on MMF with core size of 50 um and *multimodal bandwidth.distance* of 500 MHz.km.
- 1000BASE-LX/LH GBIC and SFP—GigabitEthernet transceiver that transmits at 1300 nm on either MMF or SMF. The maximum distance is 550 m on MMF fiber with core size of 62.5 um or 50 um and *multimodal bandwidth.distance* respectively of 500 MHz.km and 400 MHz.km and 10 km on SMF with 9/10 um mode field diameter, ~8.3 um core (ITU-T G.652 SMF).
- 1000BASE-ZX GBIC and SFP—GigabitEthernet transceiver that transmits at 1550 nm on SMF. The maximum distance is 70 km on regular ITU-T G.652 SMF (9/10 um mode field diameter, ~8.3 um core) and 100 km on with dispersion shifted SMF.
- 10GBASE-SR XENPAK—10 GigabitEthernet transceiver that transmits at 850 nm on MMF. The maximum distance is 300 m on 50 um core MMF with *multimodal bandwidth.distance* of 2000 MHz.km.
- 10GBASE-LX4 XENPAK—10 GigabitEthernet transceiver that transmits at 1310 nm on MMF. The maximum distance is 300 m with 50 um or 62.5 um core and *multimodal bandwidth.distance* of 500 MHz.km.
- 10BASE-LR XENPAK—10 GigabitEthernet transceiver that transmits at 1310 nm on ITU-T G.652 SMF. The maximum distance is ~10 km.
- 10BASE-ER XENPAK—10 GigabitEthernet transceiver that transmits at 1550 nm on ITU-T G.652 SMF. The maximum distance is 40 km.
- 10BASE-ER XENPAK—10 GigabitEthernet transceiver that transmits at 1550 nm on any SMF type. The maximum distance is ~80 km.
- SFP-FC-2G-SW—1 Gbps or 2 Gbps Fibre Channel transceiver that transmits at 850 nm on MMF. The maximum distance on 50 um core MMF is 500 m at 1.06 Gbps and 300 m at 2.125 Gbps.
- SFP-FC-2G-LW—1 Gbps or 2 Gbps Fibre Channel transceiver that transmits at 1310 nm on SMF. The maximum distance is 10 km on 9 um mode field diameter SMF for either speed.
- SFP-FCGE-SW—Triple-Rate Multiprotocol SFP that can be used as Gigabit Ethernet or Fibre Channel transceiver. It transmits at 810 nm on MMF. The maximum distance is 500 m on MMF with core of 50 um.
- SFP-FCGE-LW—Triple-Rate Multiprotocol SFP that can be used as Gigabit Ethernet or Fibre Channel transceiver. It transmits at 1310 nm on SMF. The maximum distance is 10 km on SMF with mode field diameter of 9 um.

For a complete list of Cisco Gigabit, 10 Gigabit, Course Wave Division Multiplexing (CWDM), and Dense Wave Division Multiplexing (DWDM) transceiver modules, see the following URL: http://www.cisco.com/en/US/products/hw/modules/ps5455/products_data_sheets_list.html.

For a list of Cisco Fibre Channel transceivers, see the following URL: http://www.cisco.com/warp/public/cc/pd/ps4159/ps4358/prodlit/mds9k_ds.pdf

Note On MMF, the modal bandwidth that characterizes different fibers is a limiting factor in the maximum distance that can be achieved. The *bandwidth.distance* divided by the bandwidth used for the transmission gives the maximum distance.

CWDM

When using dark fiber with long wavelength transceivers, the maximum achievable distance is ~10 km. CWDM and DWDM allow greater distances. Before discussing CWDM and DWDM, it is important to be familiar with the ITU G.694.2 CWDM grid, and more specifically the transmission bands (most systems operate in the 1470–1610 nm range):

- O-band—Original band, which ranges from 1260 nm to 1360 nm
- E-band—Extended band, which ranges from 1360 nm to 1460 nm
- S-band—Short band, which ranges 1460 nm to 1530 nm
- C-band—Conventional band, which ranges from 1530 nm to 1565 nm
- L-band—Long band, which ranges from 1565 nm to 1625 nm
- U-band—Ultra long band, which ranges from 1625 nm to 1675 nm

CWDM allows multiple 1 Gbps or 2 Gbps channels (or colors) to share a single fiber pair. Channels are spaced at 20 nm, which means that there are 18 possible channels between 1260 nm and 1610 nm. Most systems support channels in the 1470–1610 nm range. Each channel uses a differently colored SFP or GBIC. These channels are networked with a variety of wavelength-specific add-drop multiplexers to enable an assortment of ring or point-to-point topologies. Cisco offers CWDM GBICs, SFPs, and add-drop multiplexers that work with the following wavelengths spaced at 20 nm: 1470, 1490, 1510, 1530, 1550, 1570, 1590, and 1610 nm:

- CWDM 1470-nm SFP; Gigabit Ethernet and 1 Gbps and 2 Gbps Fibre Channel, gray
- CWDM 1490-nm SFP; Gigabit Ethernet and 1 Gbps and 2 Gbps Fibre Channel, violet
- CWDM 1510-nm SFP; Gigabit Ethernet and 1 Gbps and 2 Gbps Fibre Channel, blue
- CWDM 1530-nm SFP; Gigabit Ethernet and 1 Gbps and 2-Gbps Fibre Channel, green
- CWDM 1550-nm SFP; Gigabit Ethernet and 1Gbps and 2 Gbps Fibre Channel, yellow
- CWDM 1570-nm SFP; Gigabit Ethernet and 1 Gbps and 2 Gbps Fibre Channel, orange
- CWDM 1590-nm SFP; Gigabit Ethernet and 1 Gbps and 2 Gbps Fibre Channel, red
- CWDM 1610-nm SFP; Gigabit Ethernet and 1 Gbps and 2 Gbps Fibre Channel, brown

For a complete list of Cisco Gigabit, 10 Gigabit, CWDM, and DWDM transceiver modules, see the following URL:
http://www.cisco.com/en/US/products/hw/modules/ps5455/products_data_sheets_list.html.

For a list of Cisco Fibre Channel transceivers, see the following URL:
http://www.cisco.com/warp/public/cc/pd/ps4159/ps4358/prodlit/mds9k_ds.pdf

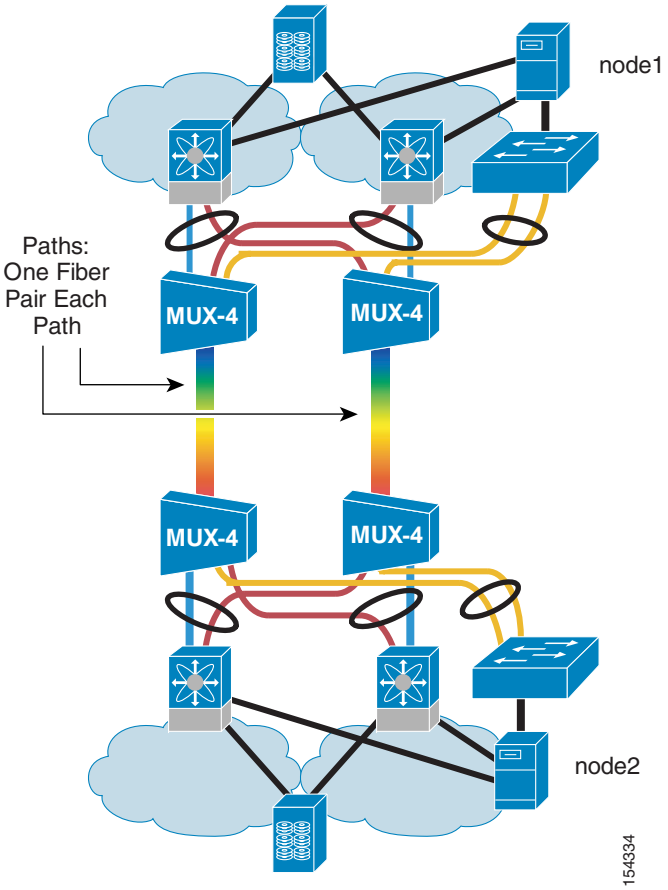
CWDM works on the following SMF fibers:

- ITU-T G.652 (standard SMF)
- ITU-T G.652.C (zero water peak fiber)
- ITU-T G.655 (non-zero dispersion shifted fiber)
- ITU-T G.653 (dispersion shifted fiber)

The CWDM wavelengths are not amplifiable and thus are limited in distance according to the number of joins and drops. A typical CWDM SFP has a 30dB power budget, so it can reach up to ~90 km in a point-to-point topology, or around 40 km in a ring topology with 0.25 db/km fiber loss, and 2x 0.5 db connector loss .

CWDM technology does not intrinsically offer redundancy mechanisms to protect against fiber failures. Redundancy is built with *client protection*. In other words, the device connecting to the CWDM "cloud" must work around fiber failures by leveraging technologies such as EtherChanneling. Figure 2-2 shows an example of a cluster with two nodes, where the SAN and the LAN are extended over ~90 km with CWDM. This topology protects against fiber cuts because port channeling on the Cisco MDS or the Catalyst switch detects the link failure and sends the traffic to the remaining link. When both fibers are available, the traffic can take both paths.

Figure 2-2 Client Protection with CWDM



DWDM

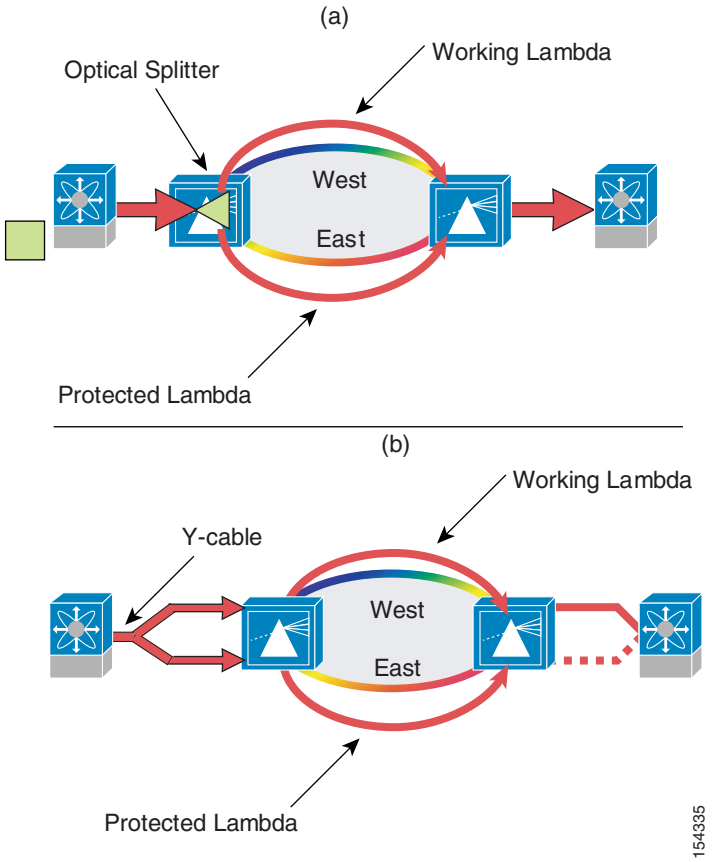
DWDM enables up to 32 channels (lambdas) to share a single fiber pair. Each of the 32 channels can operate at up to 10 Gbps. DWDM networks can be designed either as multiplexing networks similar to CWDM or with a variety of protection schemes to guard against failures in the fiber plant. DWDM is

Erbium-Doped Fiber Amplifier (EDFA)-amplifiable, which allows greater distances. DWDM can transport Gigabit Ethernet, 10 Gigabit Ethernet, Fibre Channel 1 Gbps and 2 Gbps, FICON, ESCON, and IBM GDPS. DWDM runs on SMF ITU-T G.652 and G.655 fibers.

DWDM offers the following protection mechanisms:

- Client protection—Leveraging EtherChanneling and Fibre Channel Port Channeling, this mechanism protects against fiber or line card failures by using the remaining path, without causing spanning tree or routing protocol recalculations, a new principle selection, or FSPF recalculation. With client protection, you can use both west and east links simultaneously, thus optimizing the bandwidth utilization (be careful if the west and east path have different lengths because this can cause out of order exchanges). For example, you can build a two-port port channel where one port uses a lambda on the west path and the other port uses a lambda on the east path.
- Optical splitter protection—Assume that the DWDM optical devices are connected in a ring topology such as is shown in Figure 2-3 (a). The traffic is split and sent out both a west and east path, where one is the working path and one is the “protected” path. The lambda used on both paths is the same because this operation is performed by a single transponder; also, the power of the signal is 50 percent on each path. The receiving transponder chooses only one of the two signals and sends it out to the client. Traffic is switched from a working path to a protected path in the event of a fiber failure. Switchover times for DWDM are ~50 ms or less and may cause a link up/down. This mechanism does not protect against line card failures.
- Y-cable and redundant transponders—Assume the DWDM optical devices are connected in a ring topology as shown in Figure 2-3 (b). The transceiver connects to two DWDM transponders, which in their turn respectively connect to the west mux and the east mux. The signal is sent on both the west and east path with the same power (because there is one transponder per cable termination). Each side can use a different lambda. Only one of the two receiving transponders transmits to the client.

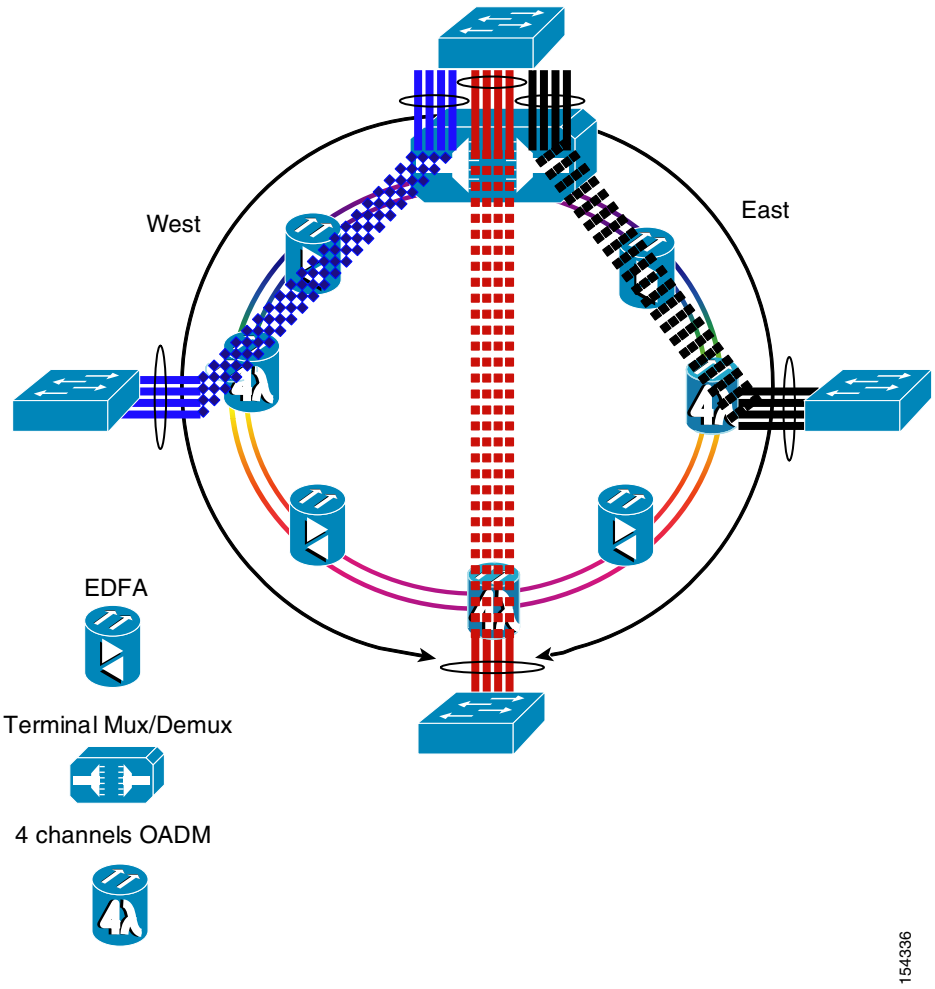
Figure 2-3 DWDM Protection Mechanisms



The basic building blocks for DWDM designs include transponders, optical multiplexers, and demultiplexers, optical add/drop multiplexers (OAMs), optical amplifiers and attenuators, variable optical attenuators (VOA), dispersion compensators (DMC/DCU), and muxponders (which are the devices that multiplex multiple client inputs onto a single channel). With these tools, it is possible to implement several designs.

Figure 2-4 shows a sample DWDM topology connecting four sites.

Figure 2-4 Sample DWDM Hub-and-Spoke Topology



From the primary site, four lambdas are carried to each site and pulled out of the ring by the OAM. In this theoretical design, a Layer 2 switch is present at each site and EtherChanneled to the switch in the primary site. The dotted lines show the logical point-to-point connections to the primary site.

Note A great tool to plan DWDM networks is Metro Planner (MP), which is a GUI-based tool to design Cisco ONS 15454 MSTP DWDM networks. MP calculates power budgets and dispersion, determines the optimized locations for amplifiers and dispersion compensation units, and produces a configuration file that can be exported to the Cisco ONS 15454. Additional details on how to build a DWDM metropolitan area network (MAN) are out of the scope of this document.

A topology such as this one is applicable to an enterprise that manages a MAN. You can configure a physical DWDM ring to provide point-to-point logical topologies, hub-and spoke logical topologies, and fully-meshed logical topologies.

Notice that in Figure 2-4, the Catalyst switches are using protected channels, so they need separate unique channels. You can use a logical ring topology where the same channel is re-used. In this case, you lose the DWDM protection and spanning tree re-routes around link failures.

If the goal of using DWDM is simply to interconnect two sites, it is possible to use a much simpler point-to-point topology with DWDM GBICs and Xenpaks to take full advantage of the available dark fiber via DWDM multiplexers.

For a list of the 32 Cisco DWDM GBICs and DWDM 10 Gigabit Ethernet Xenpaks, see the following URL: http://www.cisco.com/en/US/products/hw/modules/ps5455/products_data_sheets_list.html.

Maximum Distances and BB Credits Considerations

DWDM can provide longer distances than CWDM. Factors that affect the maximum achievable distance include power budget, chromatic dispersion, optical signal-to-noise ratio, and non-linearities.

For example, given the chromatic dispersion of the fiber and the tolerance of the transponder, the maximum achievable distance equals the tolerance of transponder divided by the coefficient of dispersion of the fiber. There are several techniques to compensate the chromatic dispersion with appropriate optical design, but they are out of the scope of this document.

Another limiting factor is the optical signal-to-noise ratio (OSNR), which degrades each time that the signal is amplified. This means that even if from a power budget point of view you can amplify the signal several times, the final OSNR might not be good enough for the receiving transponder. Because of OSNR, it makes sense to place only a maximum number of optical amplifiers in the path, after which you need full O-E-O regeneration.

Also consider that the amplification applies to all the channels, which implies that the maximum achievable length per fiber span depends also on the wavelength speed (for example, 2.5 Gbps is different from 10 Gbps), and on the total number of channels.

For example, with a single span of fiber and one single 10 Gbps channel, you can potentially achieve a maximum of 41dB loss with co-located EDFA amplification, which, on a fiber with 0.25dB/km loss, equals ~164 km. By adding amplifiers in the path, you can design for example a four-span connection with total power loss of 71dB, which equals ~284 km without regeneration.

Note As previously stated, increasing the number of channels and changing the speed of the channel changes the calculations. Many more factors need to be considered, starting with the fiber type, the chromatic dispersion characteristics of the fiber, and so on. A Cisco optical designer can assist the definition of all the required components and the best design.

Distances of thousands of kilometers can be achieved by using O-E-O regeneration with single spans of ~165 km of fiber. When EDFA or regeneration points are required, enterprises may co-locate them in the POP of the service provider from which they have purchased the fiber.

When carrying Fibre Channel on a DWDM network, the limiting factor becomes buffer-to-buffer credits (BB_credits). All data networks employ flow control to prevent data overruns in intermediate and end devices. Fibre Channel networks use BB_credits on a hop-by-hop basis with Class 3 storage traffic. Senders are permitted to send up to the negotiated number of frames (equal to the BB_credit value) to the receiver before waiting for Receiver Readys (R_RDY) to return from the receiver to replenish the BB_credits for the sender. As distance increases, so does latency, so the number of BB_credits required to maintain the flow of data increases. Fibre Channel line cards in many storage arrays have limited BB_credits, so Fibre Channel directors such as the Cisco MDS 9000, which have sufficient BB_credits, are required if extension is required beyond a few kilometers. The MDS 9000 16-port Fibre Channel line card supports up to 255 BB_credits per port, allowing DWDM metro optical fabric extension over 200 km without BB_Credit starvation and resulting performance degradation. The MDS 9000 14/2-port Multiprotocol Services Module offers Extended BB_credits, which allows distances up to 7000 km @ 1 G

FC or 3500 km @ 2 G FC. Up to 2400 BB_credits can be configured on any one port in a four-port quad with remaining ports maintaining 255 BB_credits. Up to 3500 BB_credits can be configured on any one port in a four-port quad when remaining ports shut down.

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 2 km at 1 Gbps
- 1.25 BB_credits for every 1 km at 2 Gbps

In addition, the 2.5 Gb and 10 Gb datamux cards on the Cisco ONS 15454 provide buffer-to-buffer credit spoofing, allowing for distance extension up to 1600 km for 1Gb/s Fibre Channel and 800 km for 2 Gbps Fibre Channel.

CWDM versus DWDM

CWDM offers a simple solution to carry up to eight channels (1 Gbps or 2 Gbps) on the same fiber. These channels can carry Ethernet or Fibre Channel. CWDM does not offer protected lambdas, but client protection allows re-routing of the traffic on the functioning links when a failure occurs. CWDM lambdas can be added and dropped, thus allowing the creation of hub-and-spoke, ring, and meshed topologies. The maximum achievable distance is ~100 km with a point-to-point physical topology and 40 km with a ring physical topology.

DWDM offers more channels than CWDM (32), more protection mechanisms (splitter protection and Y-cable protection), and the possibility to amplify the channels to reach greater distances. Each channel can operate at up to 10 Gbps.

In addition to these considerations, a transponder-based DWDM solution such as a solution based on ONS-15454 offers better management for troubleshooting purposes than a CWDM or DWDM solution simply based on muxes.

The two main DWDM deployment types are as follows:

- DWDM GBICs and Xenpaks (IPoDWDM) connected to a MUX (for example, Cisco ONS 15216 products)—Enterprises can take advantage of the increased bandwidth of DWDM by connecting DWDM 10 GigE Xenpaks directly to a passive MUX. This approach offers increased bandwidth over CWDM, potentially greater distances (~160 km in a single fiber span with EDFA amplifiers co-located at the data center premises and more), and several hundred kilometers with amplification and multiple fiber spans. The main disadvantage of a DWDM GBCI/MUX-based solution as compared with a DWDM transponder solution or a CWDM solution, is the fact that there is currently no DWDM transceiver for Fibre Channel, which limits the deployment to IP over DWDM.
- Regular Ethernet GBICs and Fibre Channel connected to a transponder (for example, ONS 15454 products)–Enterprises can take advantage of this type of solution to build a MAN, and can use transponders to build a transparent multiservice (including Ethernet and Fibre Channel) transport infrastructure with virtually no distance limits. A transponder-based solution has virtually no distance limitations, and allows building a Layer 2 transport (with O-E-O regeneration) across sites at thousands of kilometers of distance. As previously stated, when buying dark fiber from an SP, an enterprise may be able to co-locate their EDFA or O-E-O gear at the SP POP.

Table 2-1 compares the various solutions.

Table 2-1 Solution Comparison

	CWDM	DWDM GBIC/Xenpak	DWDM Transponder
Number of channels	8	32	32
Available speeds	2 Gbps	10 Gbps	10 Gbps
Protection	Client	Client, splitter, Y-cable	Client, splitter, Y-cable
Ethernet support	Yes	Yes	Yes
Fibre Channel support	Yes	No	Yes
Amplifiable	No	Yes	Yes
Buffer-to-buffer options	255 BB_credits from the MDS	Up to 3500 BB_credits on the MDS 14+2 cards for distances up to 7000 km	Extended BB_credits on MDS 14+2 for distances up to 7000 km or BB spoofing on the ONS 15454 for distances up to ~1600 km
Distance	~100 km	~160 km in a single span, virtually unlimited with regeneration stations	Virtually unlimited with regeneration stations
Management	None	Some management	Rich management and troubleshooting tools

Fiber Choice

When deploying an optical solution, it is important to identify the existing fiber type, or to choose the fiber type that needs to be deployed. Before using the fiber, verify the fiber conditions by using the optical time domain reflectometer (OTDR). It is also important to test the fiber for polarization mode dispersion (PMD). For example, standard fiber at 1550 nm has a dispersion of 17ps/nm/km, which may be inadequate, depending on the desired distance. For example, if the dispersion tolerance is 1800 ps/nm for the path between a transmitter and receiver and the fiber dispersion is 18 ps/nm-km, the maximum distance is 100 km between end nodes (1800/18). The dispersion tolerance or limitation is inversely proportional to the data rate, and is typically not an issue at speeds below OC-192 (10 Gbps).

For SMF, you typically have to choose between the following types:

- ITU-T G.652 (standard SMF, also known as SMF28)—Optimized for 1310 nm (SONET). Works with CWDM and DWDM. There is an attenuation peak at 1383 nm (0.50dB/km).
- ITU-T G.652.C (zero water peak fiber)—Optimized for CWDM. Most systems support CWDM in the 1470–1610 range. From a chromatic dispersion point of view, this is just like a G.652. This is also referred to as extended band, because it eliminates the water peak. Also works with DWDM.
- ITU-T G.655 (non-zero dispersion shifted fiber)—Best for DWDM. There is a little dispersion at 1550 nm, and 4ps/nm/km in the 1530–1570 nm range. This type of fiber addresses non-linearity in DWDM, and more specifically four-wave mixing (FWM). Works with CWDM, and for TDM at 1310 nm and TDM at 1550 nm.
- ITU-T G.653 (dispersion-shifted fiber [DSF])—Changes the chromatic and waveguide dispersion to cancel at 1550 nm. Works with CWDM, good for TDM at 1550 nm. Not good for DWDM.

For more information on various fibers, see the following URLs:

- Lucent/OFS fibers—http://www.ofsoptics.com/product_info/ofs-fitel.shtml
- Corning fibers— http://www.corning.com/opticalfiber/
- Alcatel fibers—http://www.alcatel.com/opticalfiber/index.htm
- Pirelli fibers— http://www.pirelli.com/en_42/cables_systems/telecom/product_solutions/optical_fibres.jhtml

SONET/SDH

Although DWDM allows Layer 2 connectivity at continental distances, it is more typical for enterprises to connect sites at continental distances via SONET offerings from service providers (SPs). SONET can transport several interface types, such as T1, DS3, N x STS-1, and so on. The Cisco ONS15454 platform offers SONET/SDH client connection options, in addition to Gigabit Ethernet and Fibre Channel (FC).

SONET/SDH-based architectures can support both sub-rate (less than 1 Gbps) and line rate (1 Gbps or 2 Gbps) services. SPs have already installed large SONET/SDH rings and can leverage this existing architecture to provide storage, Ethernet, and data center connectivity. Ethernet support includes 10/100/1000 Mbps interfaces. SAN interfaces include Fibre Channel, FICON, and ESCON.

The following line cards are deployed to support Ethernet and storage support on SONET/SDH networks:

- E-Series Ethernet card
- G-Series gigabit Ethernet card (G1000-4/G1K-4)
- ML-Series Ethernet cards (Ethernet /FCIP)
- SL-Series FC card

From the standpoint of the Fibre Channel switch, the connection looks like any other optical link. However, it differs from other optical solutions in that it *spoofs* R_RDY frames to extend the distance capabilities of the Fibre Channel transport. All Fibre Channel over optical links use BB_credits to control the flow of data between switches. R_RDY frames control the BB_credits. As the distance increases, so must the number of BB_credits. The spoofing capability of the SL line card extends this distance capability to 2800 km at 1 Gbps.

The Cisco ONS 15454 offers a number of SONET/SDH protection options, in addition to client-level protection through Fibre Channel switches. Port channels in conjunction with VSAN trunking are recommended where multiple links are used.

Just as with other optical solutions, FC over SONET/SDH is suitable for synchronous replication deployments subject to application performance constraints. Latency through the FC over SONET/SDH network is only negligibly higher than other optical networks of the same distance because each frame is serialized in and out of the FC over SONET/SDH network. The latency is 10µs per maximum-sized FC frame at 2 Gbps.

SONET/SDH Basics

The details of SONET and its history are out of the scope of this document, but it is important to consider some of the key principles behind this technology. SONET offers the following hierarchy of transport rates:

- STS-1 (51.84 Mbps)
- STS-3 (155.52 Mbps)

- STS-12 (622.08 Mbps)
- STS-24 (1244.16 Mbps)
- STS-48 (2488.32 Mbps)
- STS-192 (9.953.28 Mbps)

Note This guide refers to STS-1 and OC-1 interchangeably, and similarly for STS-3 and OC-3, and so on.

STS-3 is made of 3 STS-1s. One STS-1 can be transporting Fibre Channel, another can be transporting voice, and so on, or these three channels can be "bundled"; that is, *concatenated* in an OC-3c.

SONET/SDH has been designed to facilitate multiplexing and demultiplexing of multiple low-rate traffic. It has also been designed to carry legacy traffic such as DS1 (often referred to as T1), DS3, ATM, and so forth. This is done by subdividing an STS-1 into multiple virtual tributary groups (VTG). VTGs can transport VT1.5 (to carry DS1), VT2 (to carry E1 frames), and so on.

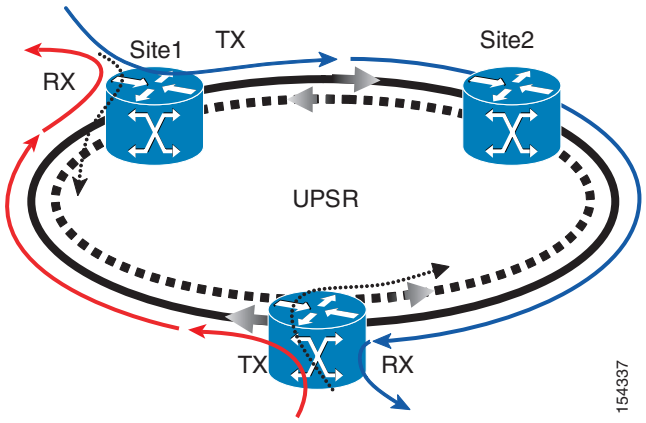
Several components comprise a SONET network, and for the purpose of this document, it is enough to focus simply on the use of add-drop-multiplexers; these are the device that insert them or remove DSs and STSs from an OC-N network.

SONET UPSR and BLSR

The two most fundamentals topologies used in SONET are the unidirectional path-switched ring (UPSR) and the bidirectional line-switched ring (BLSR).

With UPSR, the traffic between two nodes travels on two different paths. For example, in Figure 2-5, the traffic on the outer fiber travels clockwise, and the protection path (inner ring) travels counter-clockwise. As a result, Site1-to-Site3 traffic takes a different path than Site3-to-Site1 traffic.

Figure 2-5 SONET UPSR Topology



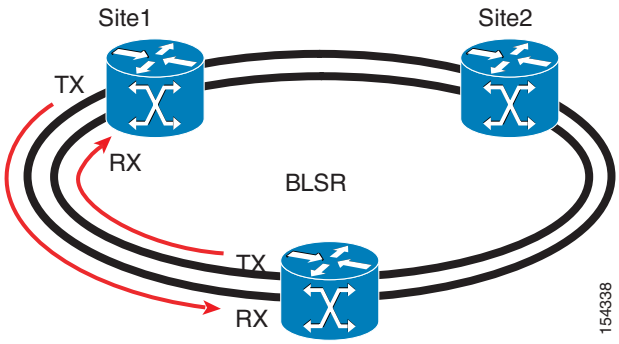
Note that Site1 TX sends traffic on both the outer and the inner ring. Site3 selects only the traffic coming from the outer ring. In case a failure occurs, the receiving end selects the traffic from the protected path. This topology works well for short distances, because the TX-RX paths can yield very different latency values, which can affect flow control.

From a bandwidth utilization perspective, the communication between two nodes always involves the full ring. For example, if Site1-to-Site2 is using one STS-1 and the ring is an OC-3 ring, there are only 2 STS-1s left for the other nodes to use.

From a logical topology point of view, USPR rings are more suited for hub-and-spoke topologies.

Figure 2-6 shows the BLSR topology.

Figure 2-6 SONET BLSR Topology



In this case, the communication between Site1 and Site3 does not involve the full ring. This means that if Site1-to-Site3 is using 6 STS-1s on an OC-12 ring, Site1-to-Site2 can still use 6 STS-1s and Site2-to-Site3 can also use 6 STS-1s.

Note that only half of the available channels can be used for protection reasons. In other words, if one link between Site1-and-Site3 fails, Site1 and Site3 need to be able to communicate over the path that goes through Site2. Six STS-1s need to be available along the alternate path.

BLSR offer several advantages: TX and RX between two sites travel on the same path, and bandwidth utilization is more optimized.

Ethernet Over SONET

Several options are available to transport Ethernet Over SONET. With the ONS products, the following three families of line cards are often used for this purpose:

- *E-Series* cards include the E100T-12/E100T-G and the E1000-2/E1000-2. An E-Series card operates in one of three modes: multi-card EtherSwitch group, single-card EtherSwitch, or port-mapped. E-Series cards in multicard EtherSwitch group or single-card EtherSwitch mode support Layer 2 features, including virtual local area networks (VLANs), IEEE 802.1Q, STP, and IEEE 802.1D. Port-mapped mode configures the E-Series to operate as a straight mapper card and does not support these Layer 2 features. Within a node containing multiple E-Series cards, each E-Series card can operate in any of the three separate modes.
- *G-Series* cards on the Cisco ONS 15454 and ONS 15454 SDH map up to four Gigabit Ethernet ports onto a SONET/SDH transport network and provide scalable and provisionable transport bandwidth at signal levels up to STS-48c/VC4-16 per card. The G-Series cards provide line rate forwarding for all Ethernet frames (unicast, multicast, and broadcast) and can be configured to support Jumbo frames (defined as a maximum of 10,000 bytes). The card maps a single Ethernet port to a single STS circuit. You can independently map the four ports on a G-Series card to any combination of STS-1, STS-3c, STS-6c, STS-9c, STS-12c, STS-24c, and STS-48c circuit sizes, provided that the sum of the circuit sizes that terminate on a card do not exceed STS-48c.

- *ML-Series* cards are independent Gigabit Ethernet (ML1000-2) or Fast Ethernet (ML100T-12 and ML100X-8) Layer 3 switches that process up to 5.7 Mpps. The cards are integrated into the ONS 15454 SONET or the ONS 15454 SDH. The ML-Series card uses Cisco IOS Release 12.2(28) SV, and the Cisco IOS command-line interface (CLI) is the primary user interface for the ML-Series card. The ML100T-12 features twelve RJ-45 interfaces, and the ML100X-8 and ML1000-2 feature two SFP slots to support short wavelength (SX) and long wavelength (LX) optical modules. All three cards use the same hardware and software base and offer similar feature sets. The ML-Series card features two virtual packet-over-SONET/SDH (POS) ports, which function in a manner similar to OC-N card ports.

  The ML-Series cards support the following:

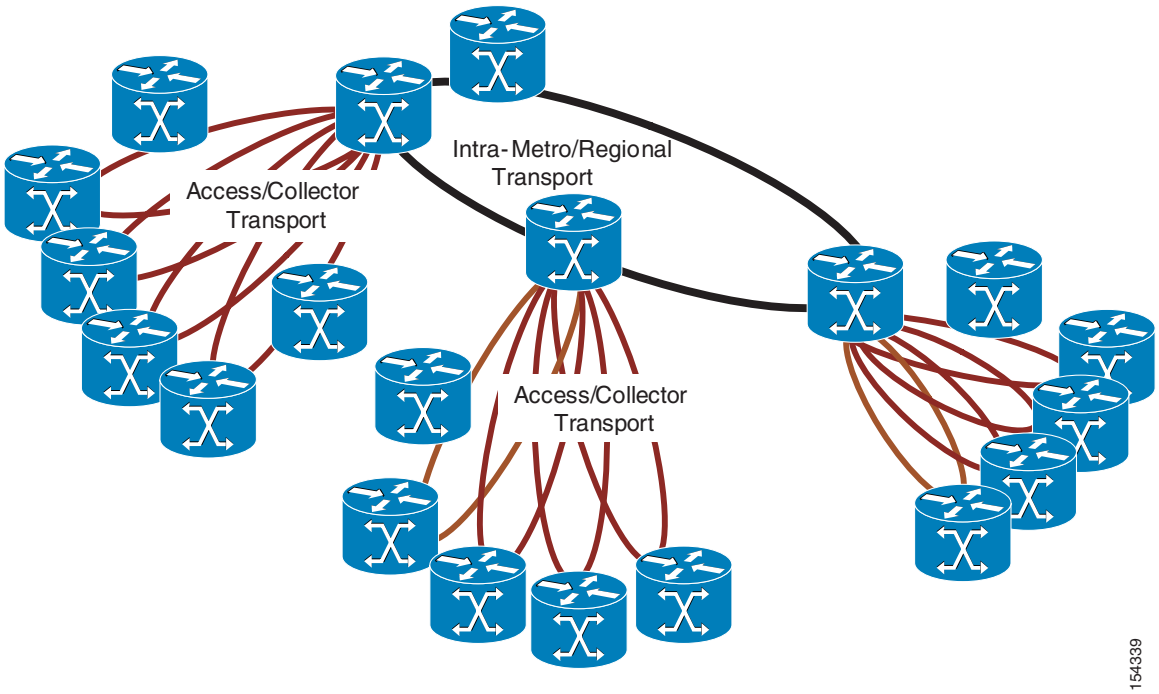
  - Transparent bridging
  - MAC address learning
  - Aging and switching by hardware
  - Multiple Spanning-Tree (MST) protocol tunneling
  - 255 bridge groups maximum
  - IEEE 802.1q VLAN tunneling
  - IEEE 802.1d and IEEE 802.1w Rapid Spanning-Tree Protocol (RSTP)
  - Resilient packet ring (RPR)
  - Ethernet over Multiprotocol Label Switching (EoMPLS)
  - EtherChanneling
  - Layer 3 unicast and multicast forwarding
  - Access control lists
  - Equal-cost multipath (ECMP)
  - VPN Routing and Forwarding (VRF)-lite
  - EIGRP, OSPF, IS-IS, PIM, BGP, QoS and more

Service Provider Topologies and Enterprise Connectivity

Most service providers deploy SONET rings where UPSRs are used at the edge of the network, while the backbone may be provided by BLSR rings. From an enterprise point of view, the end-to-end connectivity between sites over an SP SONET offering can consist of several rings of different types.

Figure 2-7 shows a typical SP topology with multiple SONET rings.

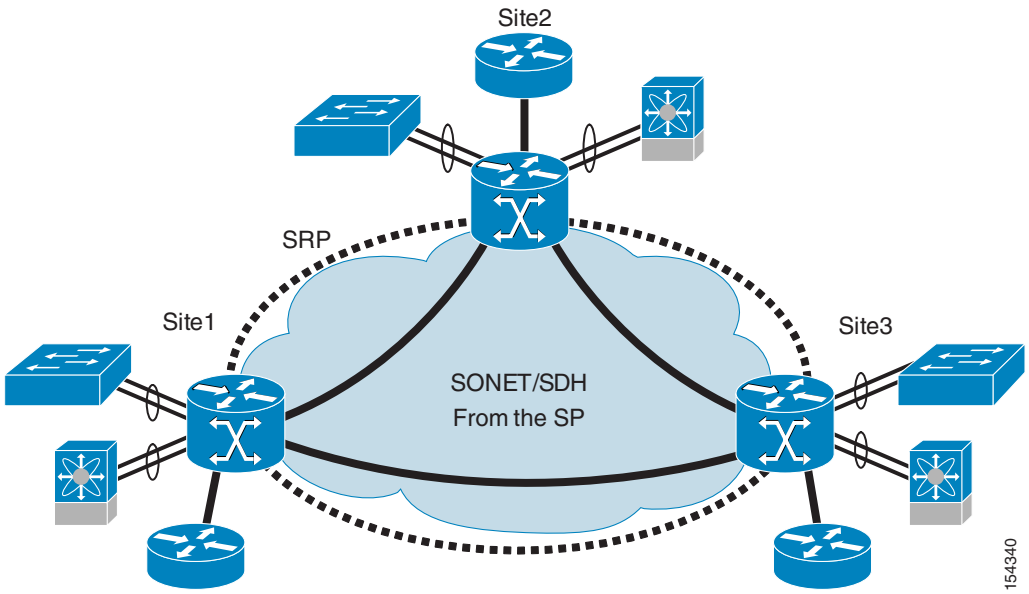
Figure 2-7 Typical SP Topology with Multiple SONET Rings



An enterprise with SONET connectivity between sites and with the need to extend an Ethernet segment across sites as well as the Fibre Channel network can consider using Ethernet over SONET and Fibre Channel over SONET by means of the ML-series, G-series, and SL-series cards connected to an ONS-15454 device.

Figure 2-8 shows the use of SONET to bridge Ethernet and Fibre Channel.

Figure 2-8 Use of SONET to Bridge Ethernet and Fibre Channel



Resilient Packet Ring/Dynamic Packet Transport

The Resilient Packet Ring (RPR)/Dynamic Packet Transport (DPT) protocol overcomes the limitations of SONET/SDH and spanning tree in packet-based networks. RPR/DPT convergence times are comparable to SONET/SDH and faster than 802.1w. RPR/DPT uses two counter-rotating rings where fibers are concurrently used to transport both data and control traffic. DPT rings run on a variety of transport technology including SONET/SDH, wavelength division multiplexing (WDM), and dark fiber. IEEE 802.17 is the working group that defines the standard for RPRs. The first major technical proposal for an RPR protocol was submitted by Cisco based on the DPT protocol.

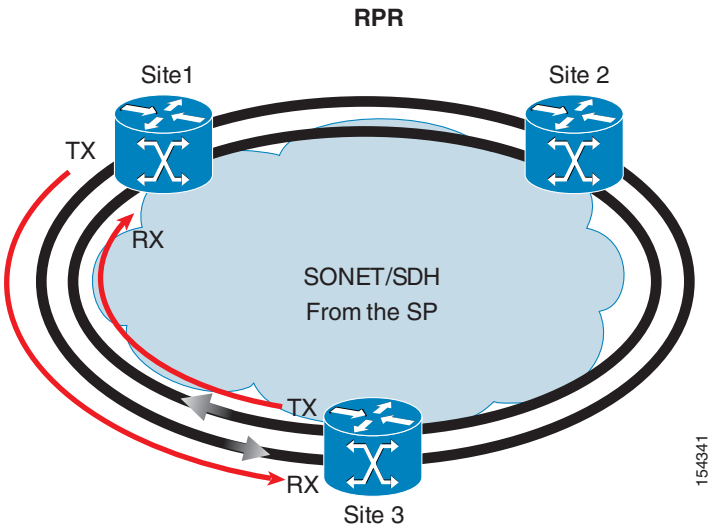
RPR/DPT provides sophisticated protection switching for self-healing via the intelligent protection switching (IPS) algorithm. IPS enables sub-50 ms protection for rapid IP service restoration. RPR/DPT rings use automatic procedures for address assignment and resolution, ring topology and status discovery, and control message propagation, which optimizes ring traffic routing and management procedures.

Spatial Reuse Protocol

Spatial Reuse Protocol (SRP) was developed by Cisco for ring-based media, and is the underlying media-independent protocol used in the DPT products. SRP uses two counter-rotating rings (inner and outer ring), and performs topology discovery, protection switching (IPS), and bandwidth control. Spatial reuse is among the key characteristics of SRP.

Figure 2-9 shows the communication between Site1and Site3.

Figure 2-9 Use of SONET to Build an RPR Ring



This communication in traditional ring technologies involves the full ring. With SRP, the 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 wherein 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 looking at 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.

RPR and Ethernet Bridging with ML-series Cards on a SONET Network

RPR/DPT operates at the Layer 2 level and operates on top of protected or unprotected SONET/SDH. It is well-suited for transporting Ethernet over a SONET/SDH ring topology and enables multiple ML-Series cards to become one functional network segment or shared packet ring (SPR). Although the IEEE 802.17 draft was used as reference for the Cisco ML-Series RPR implementation, the current ML-Series card RPR protocol does not comply with all clauses of IEEE 802.17 because it supports enhancements for Ethernet bridging on an RPR ring.

The ML-Series cards in an SPR must connect directly or indirectly through point-to-point STS/STM circuits. The point-to-point STS/STM circuits are configured on the ONS node and are transported over the SONET/SDH topology of the ONS node with either protected or unprotected circuits. On circuits unprotected by the SONET/SDH mechanism, RPR provides resiliency without using the capacity of the redundant protection path that a SONET/SDH-protected circuit requires. This frees this capacity for additional traffic. RPR also utilizes the bandwidth of the entire ring and does not block segments as does spanning tree.

Differently from IEEE 802.17, the ML-Series cards perform destination stripping both for routed and bridged traffic. IEEE 802.17 performs destination stripping only for routed traffic; bridged frames are flooded on the ring. The Cisco DPT implementation has a local MAC address table on each node; if the traffic matches a MAC address that is local to the line card, it is not sent out on the ring, thus preserving the ring bandwidth. The size of the MAC address table is optimized because RPR transit traffic is not learned by the Layer 2 forwarding table. The Layer 2 forwarding table is a CAM table for wire rate Layer 2 forwarding.

Metro Offerings

Customers who need to connect data centers can choose from several service provider offerings, some of which have been already described in this guide: dark fiber (which can in turn be used in conjunction with CWDM or DWDM) and SONET/SDH point-to-point connectivity.

In addition to these offerings, enterprises can also buy Metro Ethernet connectivity. Ethernet is attractive because it allows for rate limiting in increments not provided by time division multiplexing (TDM) service providers. For example, enterprise customers can purchase a 10 Mbps service (or less) instead of committing to a DS3 (44.736 Mbps) connection. With this solution, enterprise customers looking for a transparent LAN service can obtain connections starting at 0.5–1000 Mbps. Metro Ethernet supports both SAN and LAN traffic. Typical applications include Ethernet connectivity (server-to-server and client-server communications) and asynchronous/synchronous storage applications. Metro Ethernet can be used to transport FC over IP (FCIP) along with the traditional server/client communication.

These are typically available in either of the following formats:

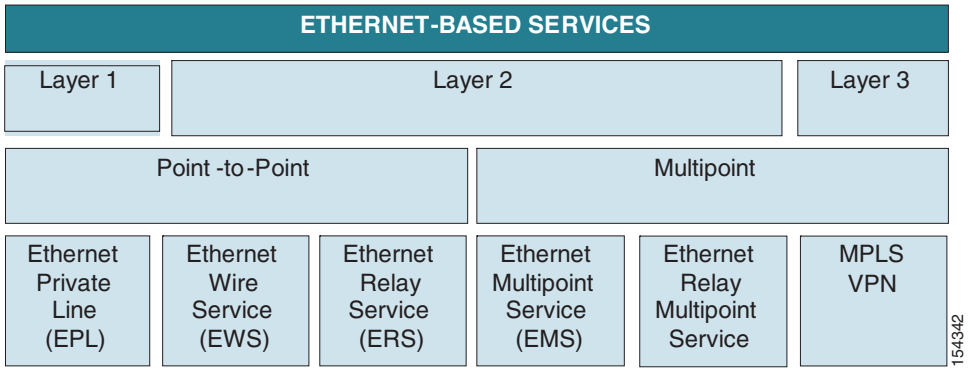
- Ethernet Relay Service (ERS) or Ethernet Virtual Circuit Service (EVCS)—Provides a point-to-point Ethernet circuit between customer premises equipment (CPEs) over the metro network. Multiple ERSes can be mapped from a single CPE over the user-to-network interface (UNI) of the SP. Each circuit is associated with and mapped to a single SP VLAN. In this way, ERS

emulates the traditional Frame Relay service in which the VLAN is analogous to the data-link connection identifier (DLCI). This type of transport does not carry Bridge Protocol Data Units (BPDUs), Cisco Discovery Protocol (CDP), VLAN Trunk Protocol (VTP), and so on.

- Ethernet Wire Service (EWS)—Emulates a point-to-point virtual wire connection, and appears to the customer as a "clear channel" pipe. This is sometimes referred to as an Ethernet private line. A customer using EWS does not see the SP network, and the connection appears as if it were a local Ethernet segment. All data passes transparently over the connection and the following Layer 2 (L2) control protocols STP, CDP, and VTP. Data transparency means that the data is transported intact with the VLAN ID untouched.
- Ethernet Multipoint Service (EMS), also known as Transparent LAN Service (TLS)—Multipoint-to-multipoint virtual wire connection. EMS is the multipoint extension of the EWS, and has the same service characteristics, such as data transparency and L2 control protocol tunneling. EMS is analogous to a multipoint Ethernet private line service.

Figure 2-10 shows the relation between Metro Ethernet services, their network layer, and point-to-point versus point-to-multipoint classification.

Figure 2-10 Metro Ethernet Services



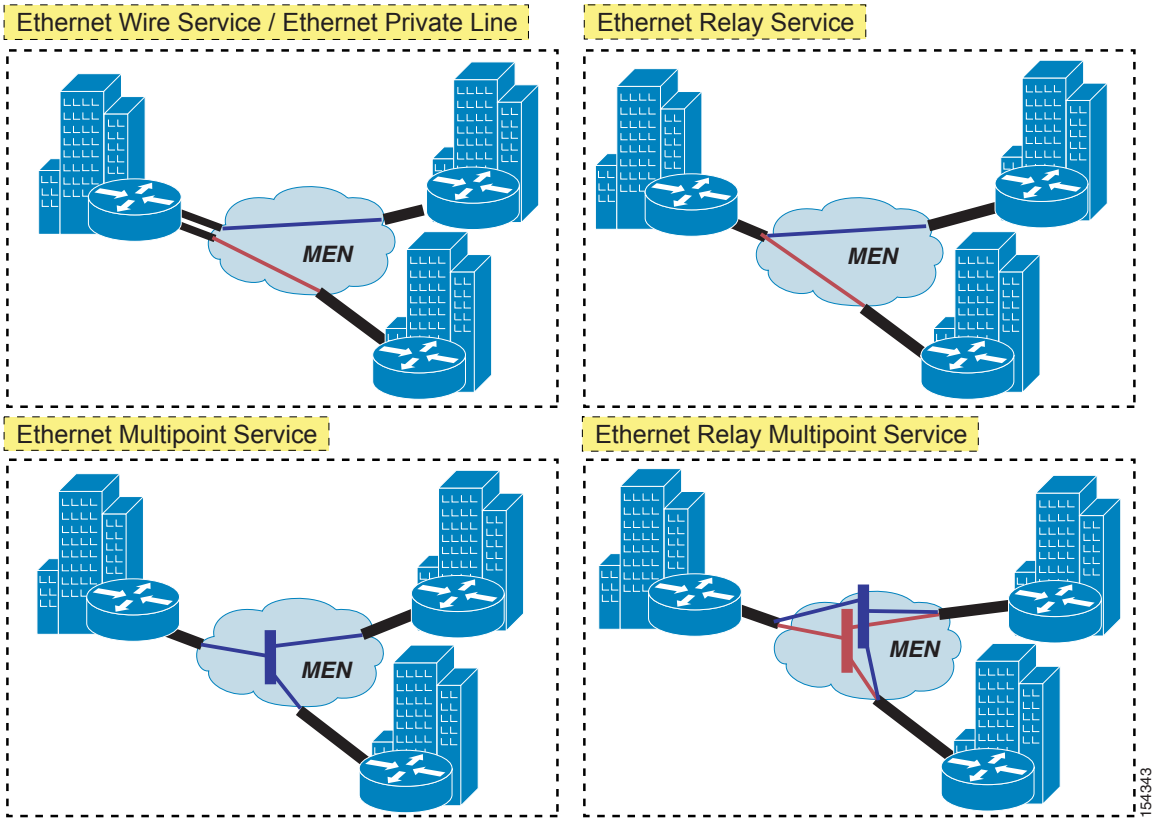
MEF identifies the following Ethernet services:

- Ethernet Line Service Type (E-Line)—Point-to-point Ethernet service; that is, a single point-to-point Ethernet circuit provisioned between two UNIs.
- Ethernet LAN Service Type (E-LAN)—Multipoint-to-multipoint Ethernet service; that is, a single multipoint-to-multipoint Ethernet circuit provisioned between two or more UNIs.
- Ethernet Private Line (EPL)—Port-based point-to-point E-Line service that maps Layer 2 traffic directly onto a TDM circuit.
- Ethernet Wire Service (EWS)—Point-to-point port-based E-Line service that is used primarily to connect geographically remote LANs over an SP network.
- Ethernet Relay Service (ERS)—Point-to-point VLAN-based E-Line service that is used primarily for establishing a point-to-point connection between customer routers.
- Ethernet Multipoint Service (EMS)—Multipoint-to-multipoint port-based E-LAN service that is used for transparent LAN applications.
- Ethernet Relay Multipoint Service (ERMS)—Multipoint-to-multipoint VLAN-based E-LAN service that is used primarily for establishing a multipoint-to-multipoint connection between customer routers.

- ERS Access to MPLS VPN—Mapping of an Ethernet connection directly onto an MPLS VPN. It provides Layer 2 access using an ERS UNI, but is a Layer 3 service because it traverses the MPLS VPN.
- ERS Access to ATM Service Interworking (SIW)—Point-to-point VLAN-based E-Line service that is used for Ethernet-to-ATM interworking applications.

Figure 2-11 shows a variety of metro Ethernet services.

Figure 2-11 Metro Ethernet Services



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.

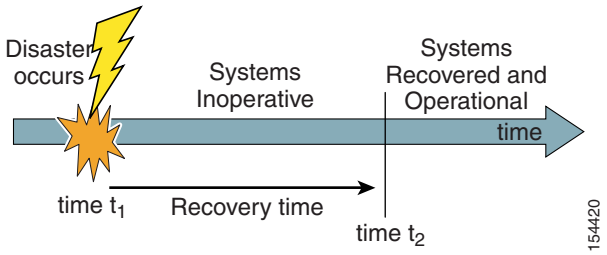
This document refers generically to *metro, regional,* and *continental clusters* as geoclusters.

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:

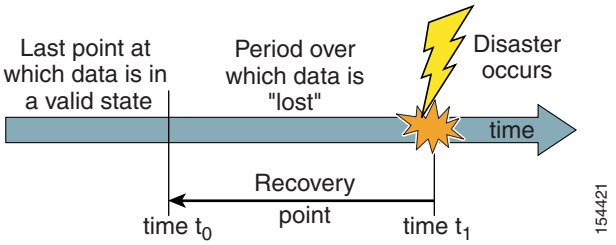
- Recovery time objective (RTO)—How long it takes for the enterprise systems to resume operation after a disaster, as shown in Figure 3-1. RTO is the longest time that your organization can tolerate.

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



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:

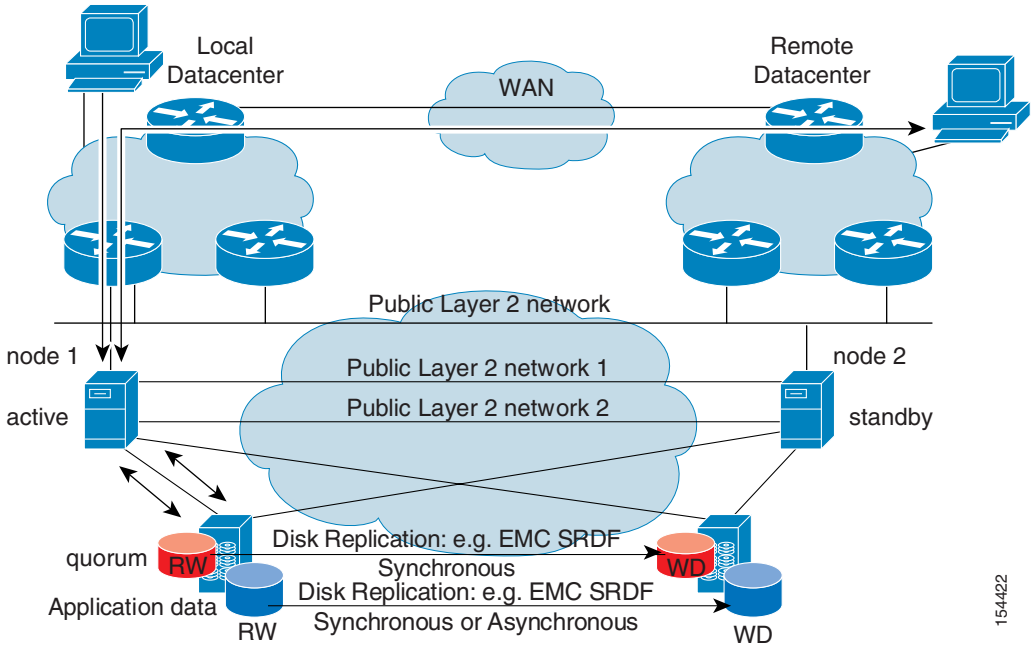
- Tape backup and restore
- Periodic replication and backups
- Asynchronous replication
- Synchronous replication

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.

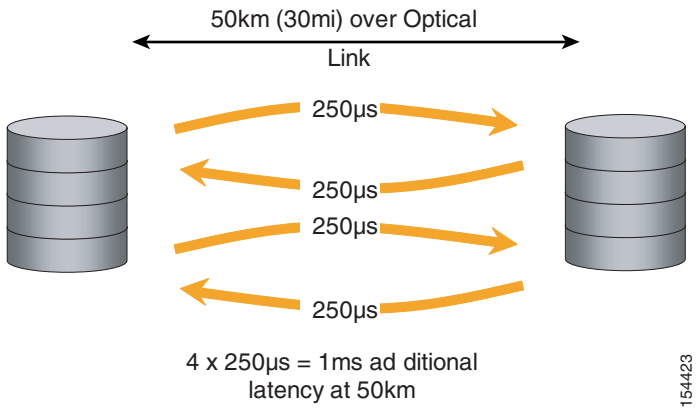
Host-based mirroring is a synchronous mirroring method where the host itself is responsible for duplicating writes to the storage arrays. However, because the two storage arrays are identical copies, reads are performed only against one array.

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

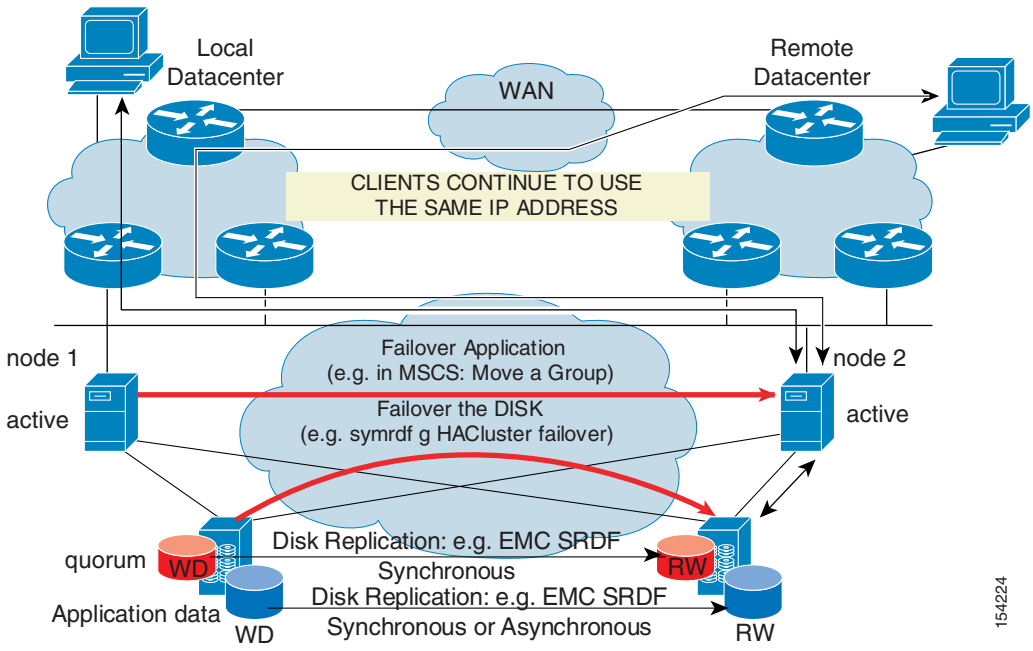
Disk replication and clustering software alone might not be enough to build a geocluster solution. When a failover happens, you may want the disks to failover also to the data center where the server nodes are active. This may require additional software or scripting. Typical products that are used for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

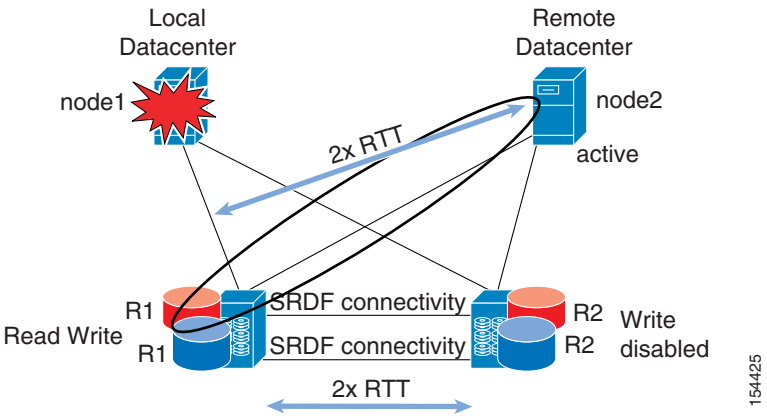
Figure 3-5 shows how a failure scenario can appear with geoclusters present. The administrator of the local data center might *move* a group to the remote data center. The disks should move together with the application; that is, the disks in the remote data center should become R/W, while the disks 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.

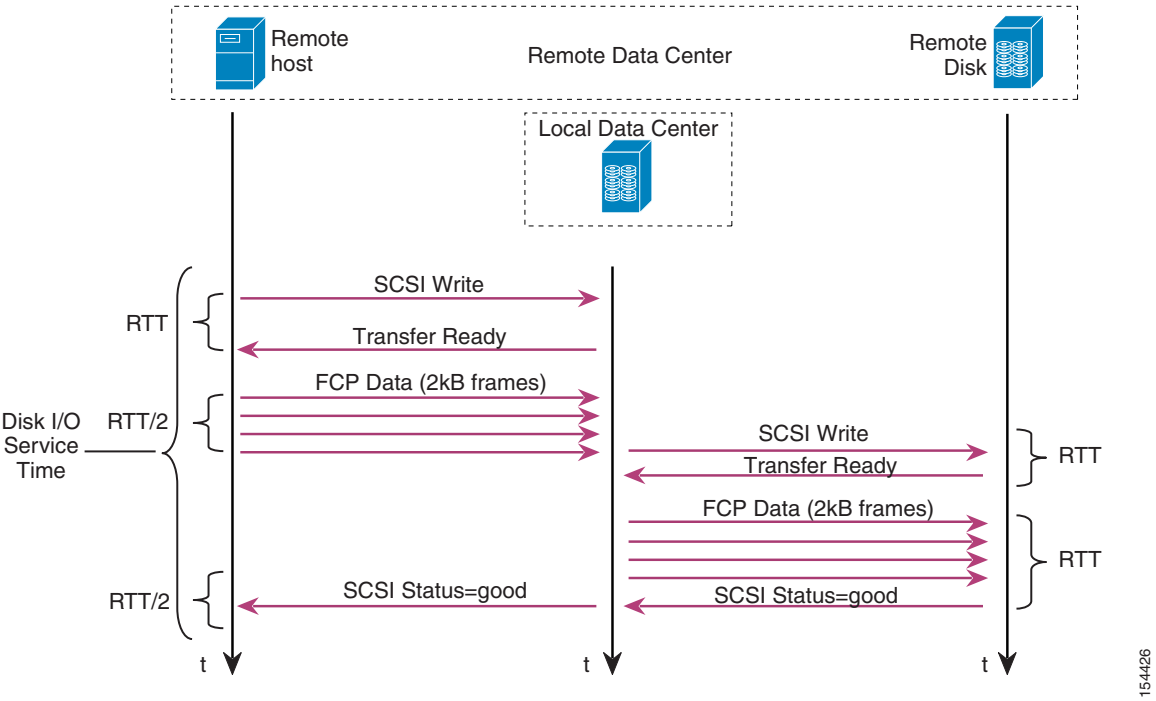
Figure 3-6 Node Failover without "Assisted Disk Failover" (Cluster Enabler and Similar Products)



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

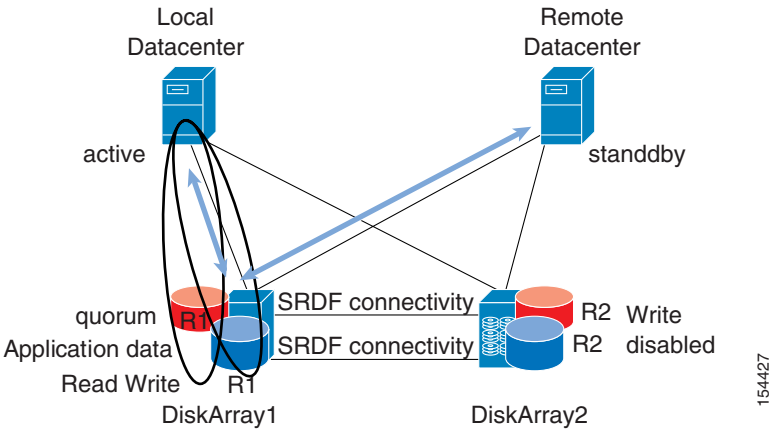
Various factors impact the performance of a geographically-clustered application, including the following:

- 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:

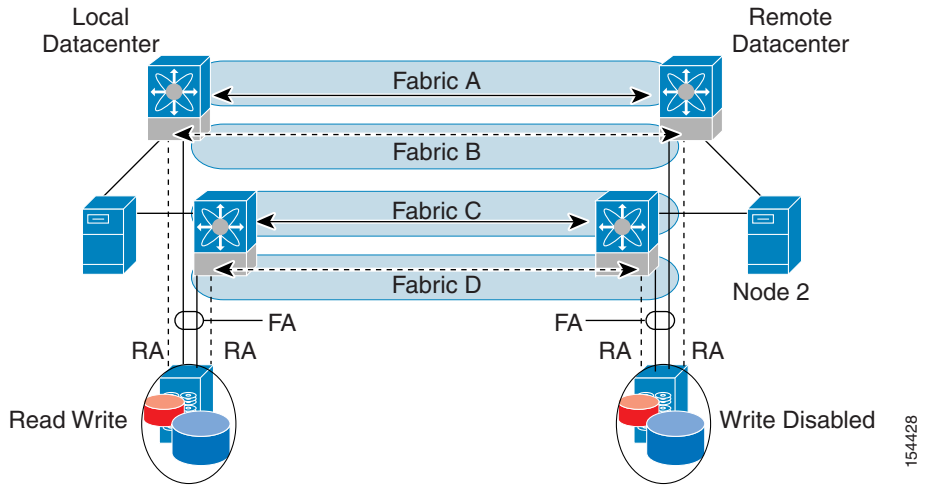
- Disk—Depends on several factors, such as revolutions per minute (RPM), redundant array of independent disks (RAID) type, type of access (*random*, as in the case of a database; *sequential*, as in the case of streaming), and disk connectivity to the SAN. The storage used in these tests consisted of 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

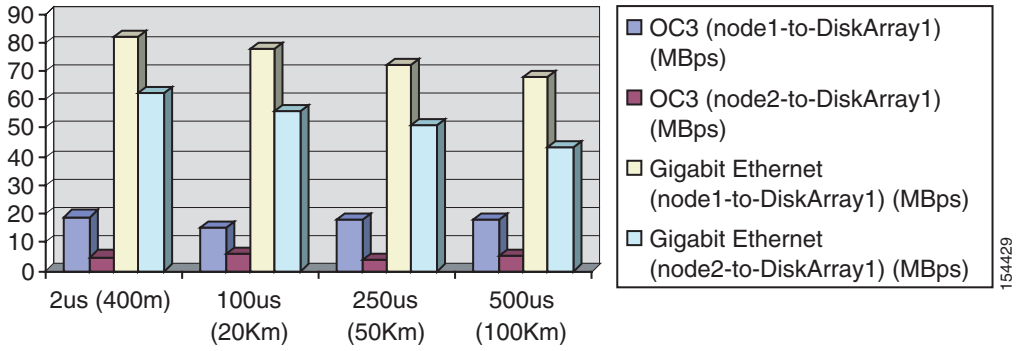


Figure 3-10 shows the variation of the maximum throughput at increasing distances, and contrasts the achievable performance with the following scenarios:

- 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.

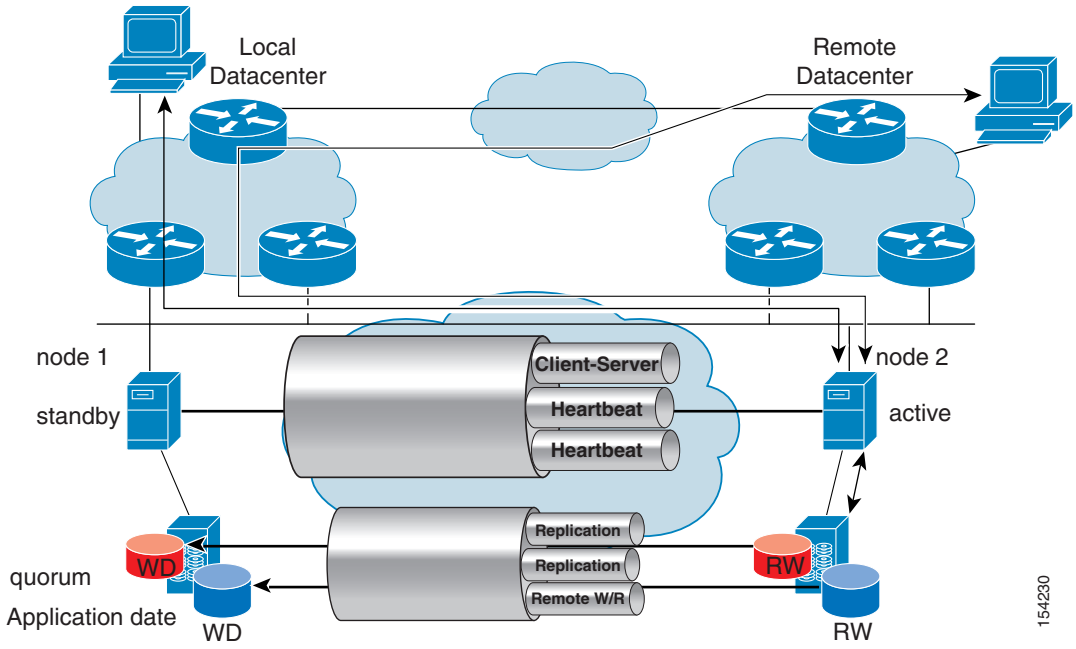
In sum, the test results show the following:

- 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

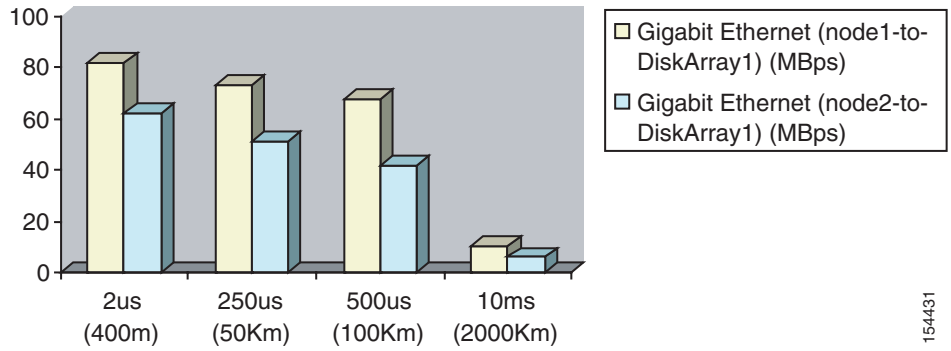


Distance Impact on the Application Throughput

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.

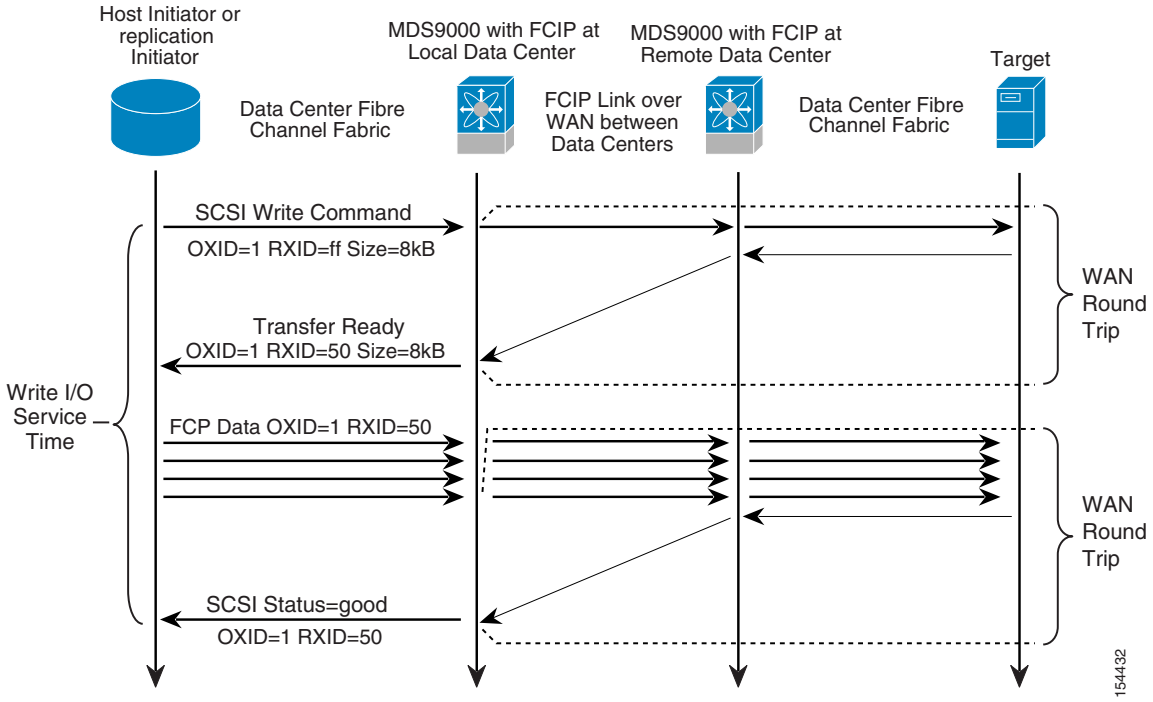
Benefits of Cisco FC-WA

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)

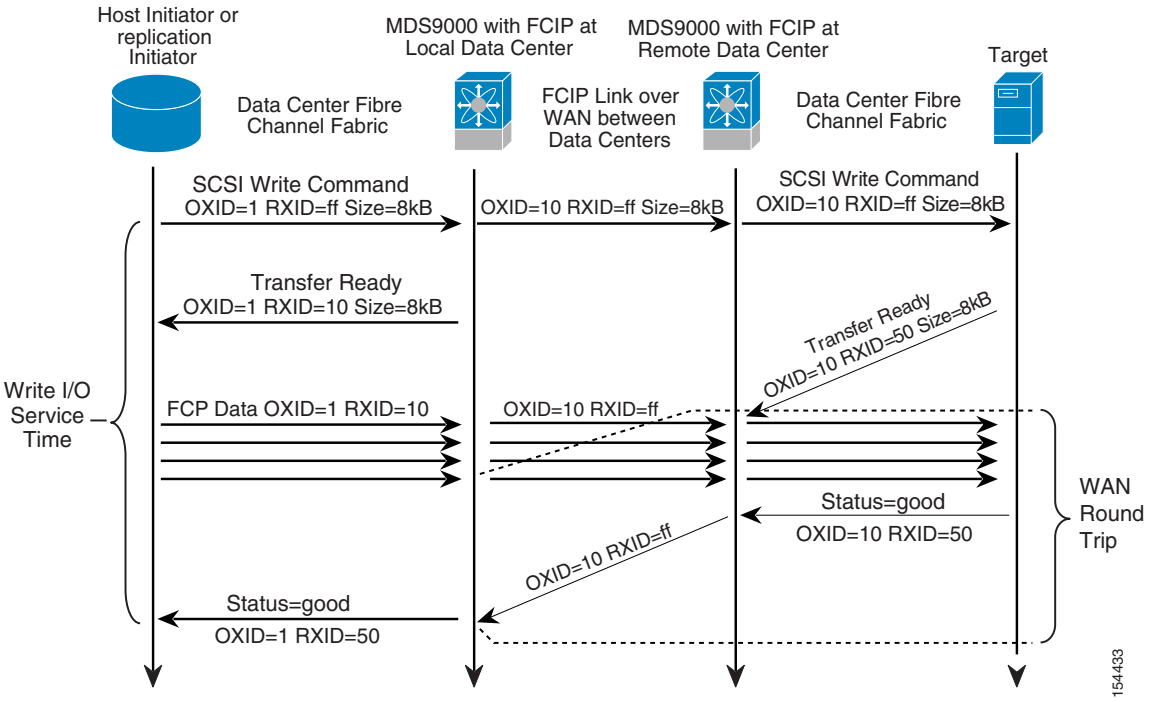


The protocol for a normal SCSI FCP Write without FC-WA is as follows:

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

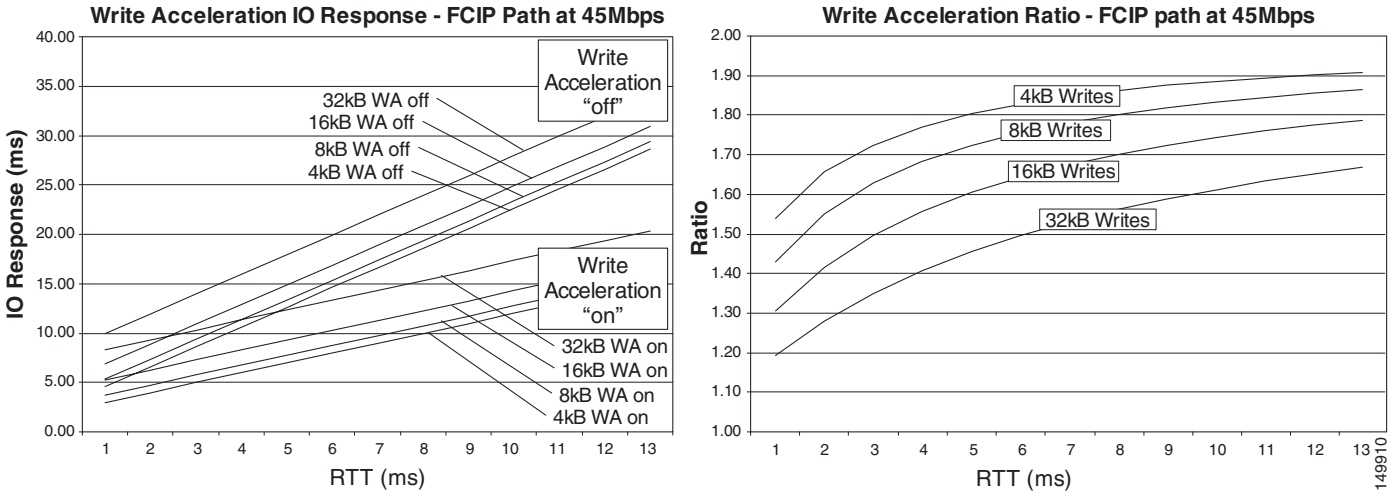


Figure 3-17 FC-WA I/O Response and Ratio at 155 Mbps (Single Outstanding I/O)—Approximate

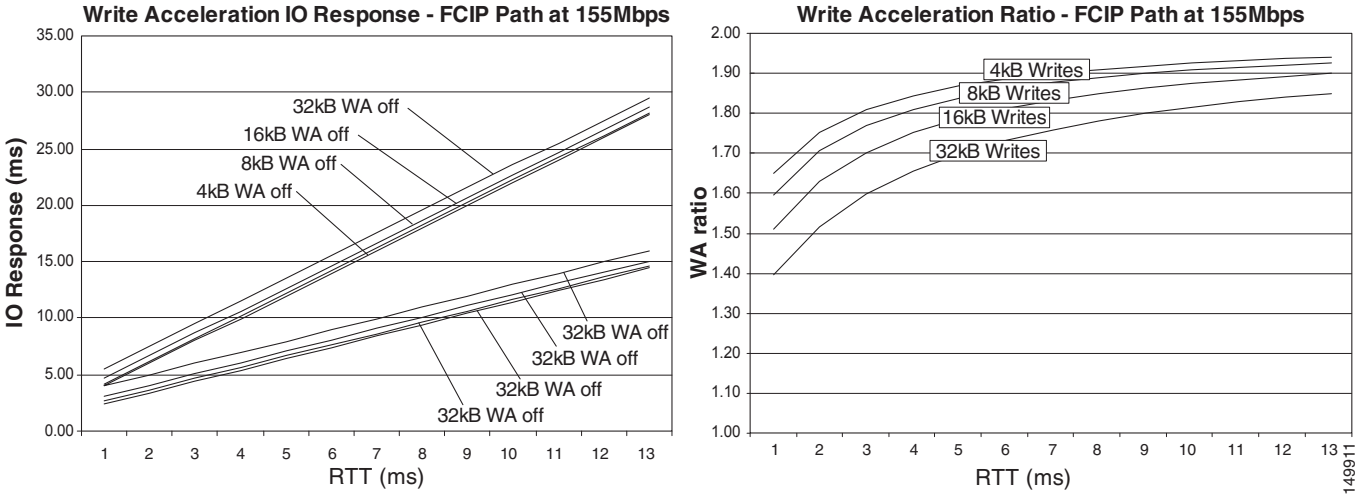


Figure 3-18 FC-WA I/O Response and Ratio at 622 Mbps (Single Outstanding I/O)—Approximate

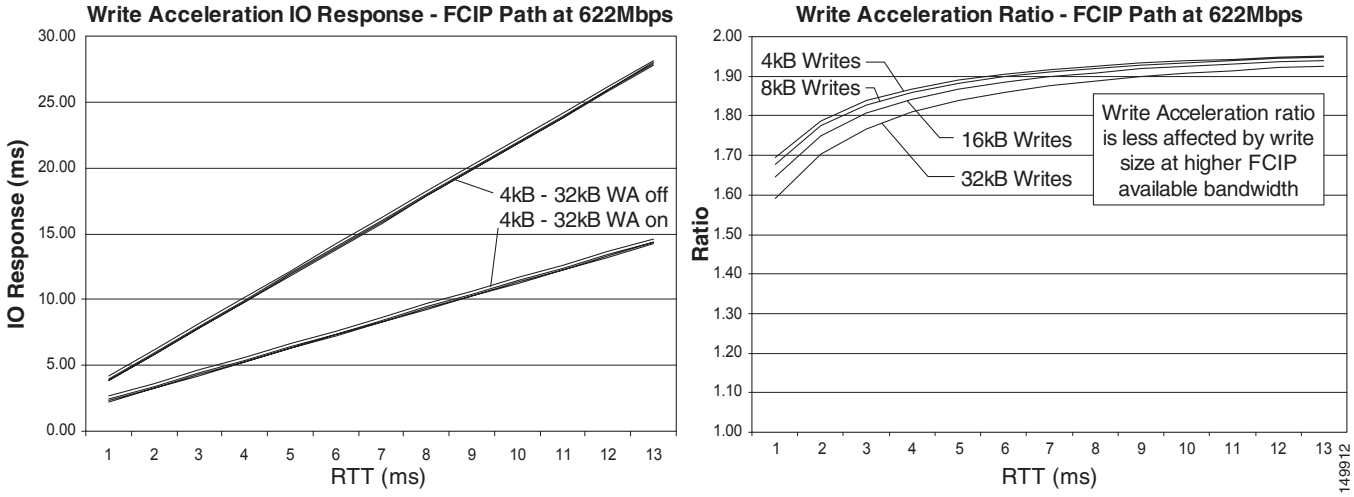
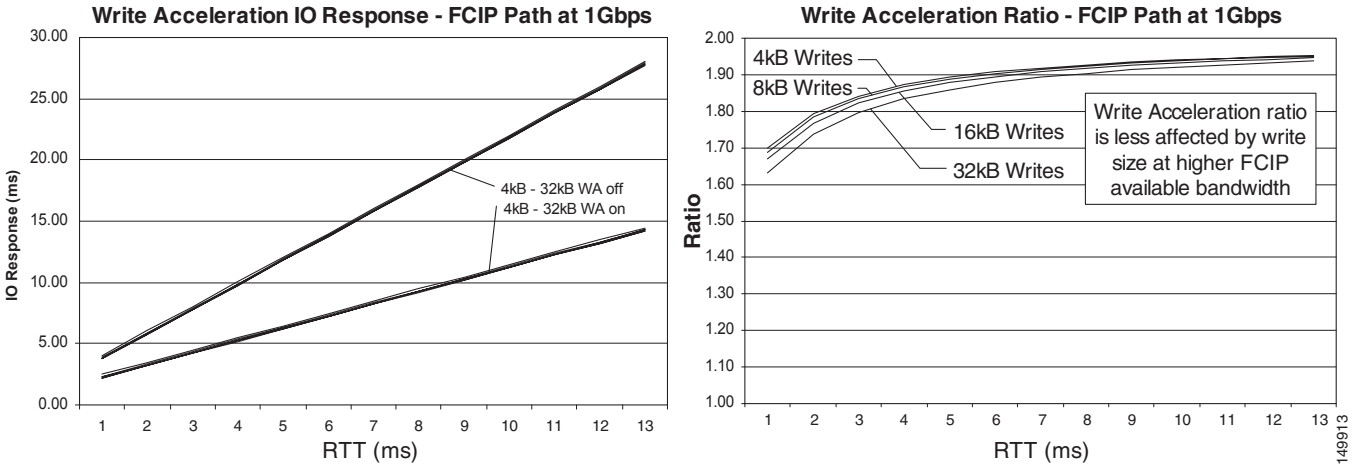


Figure 3-19 FC-WA I/O Response and Ratio at 1 Gbps (Single Outstanding I/O)—Approximate



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

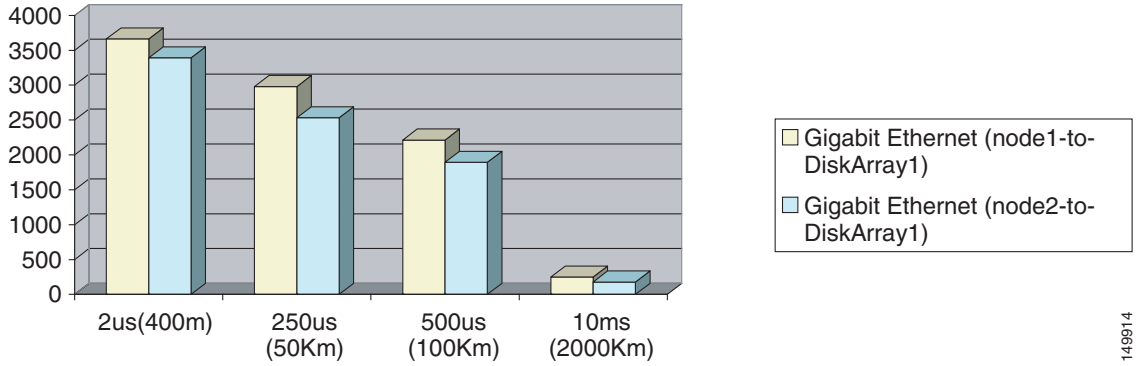
The distance between data centers decides what type of replication can be implemented on the disk arrays, but also constrains the performance of the local server to a certain level of response time (and IOPS as a result). Based on the distance, you may not be able to operate the application from the 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

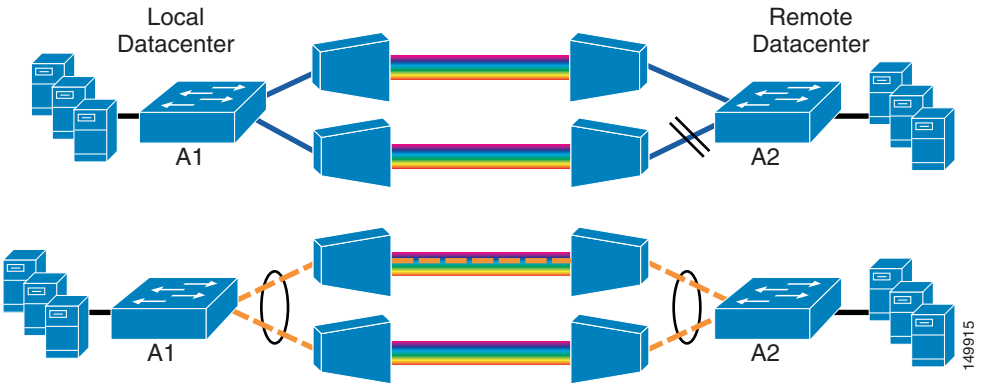
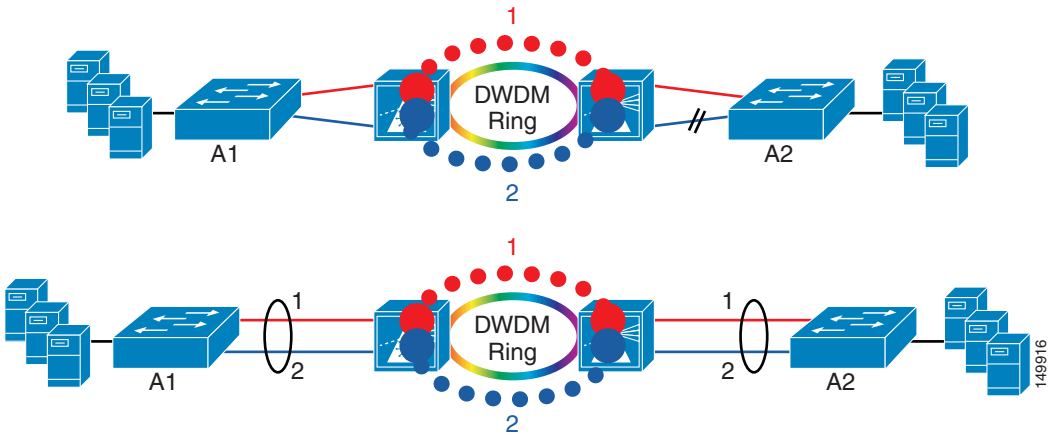


Figure 3-22 shows the Layer 2 extension topologies with DWDM. The lambdas for each port in the channel can be configured to use a different physical route on the ring. You can choose between using client protection (that is, using the port channel protection) or adding a further level of protection, 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

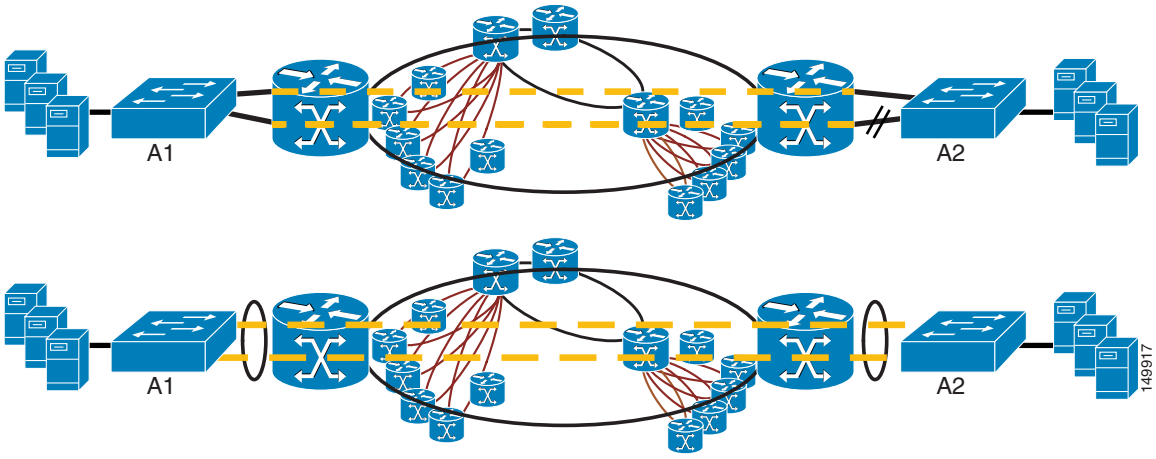


SONET Topologies

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.

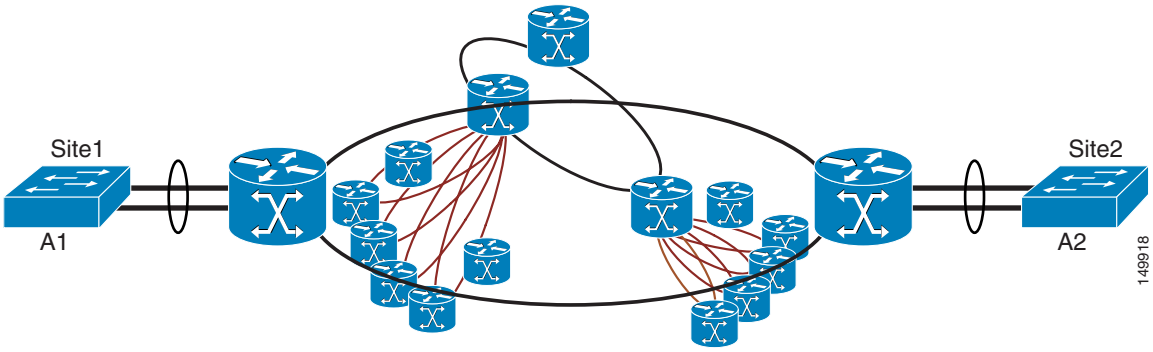
Figure 3-23 Layer 2 Extension Topologies with SONET (Layer 1 Circuits)



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)



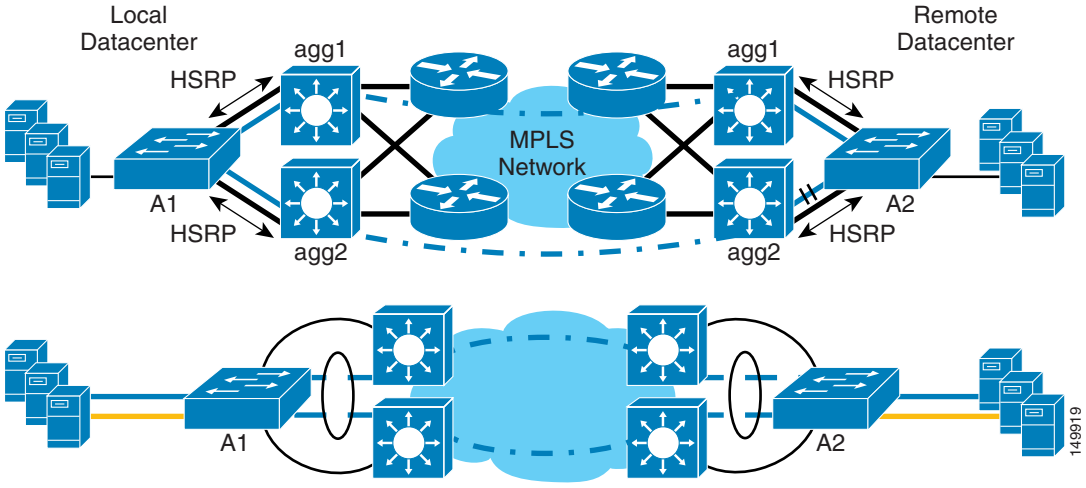
Multiprotocol Label Switching Topologies

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

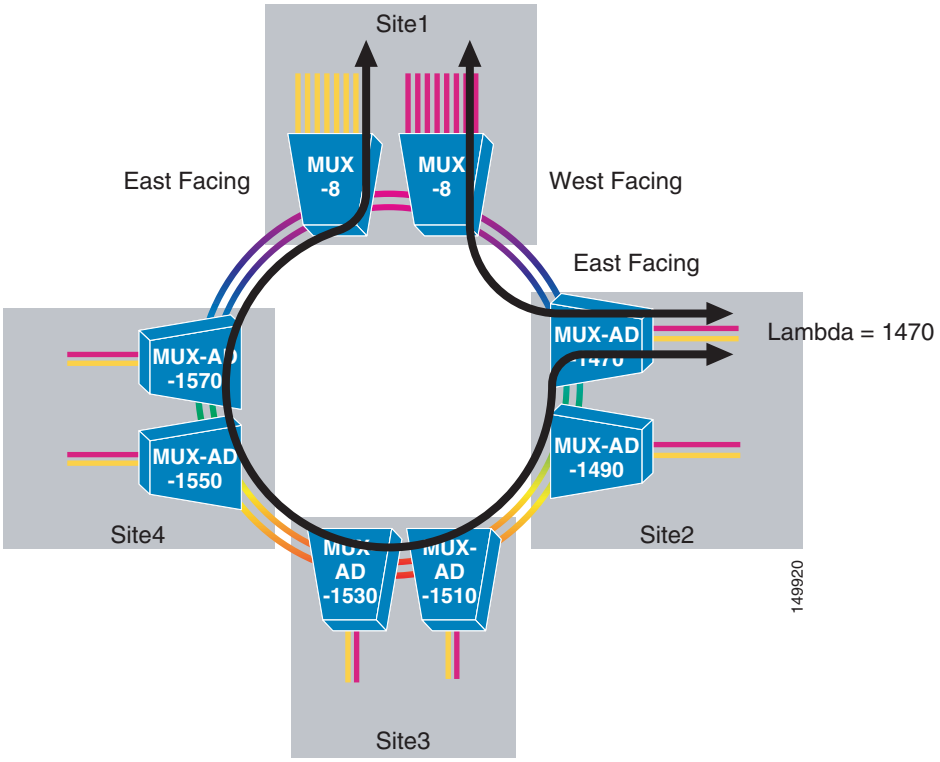
With three or more sites, the main topologies are hub-and-spoke and ring. The ring topology may match a physical topology, or may be just the topology of a specific lambda/circuit/pseudowire. The hub-and-spoke is typically a "logic" topology carried on top of one or several rings, or any multi-hop routed 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.

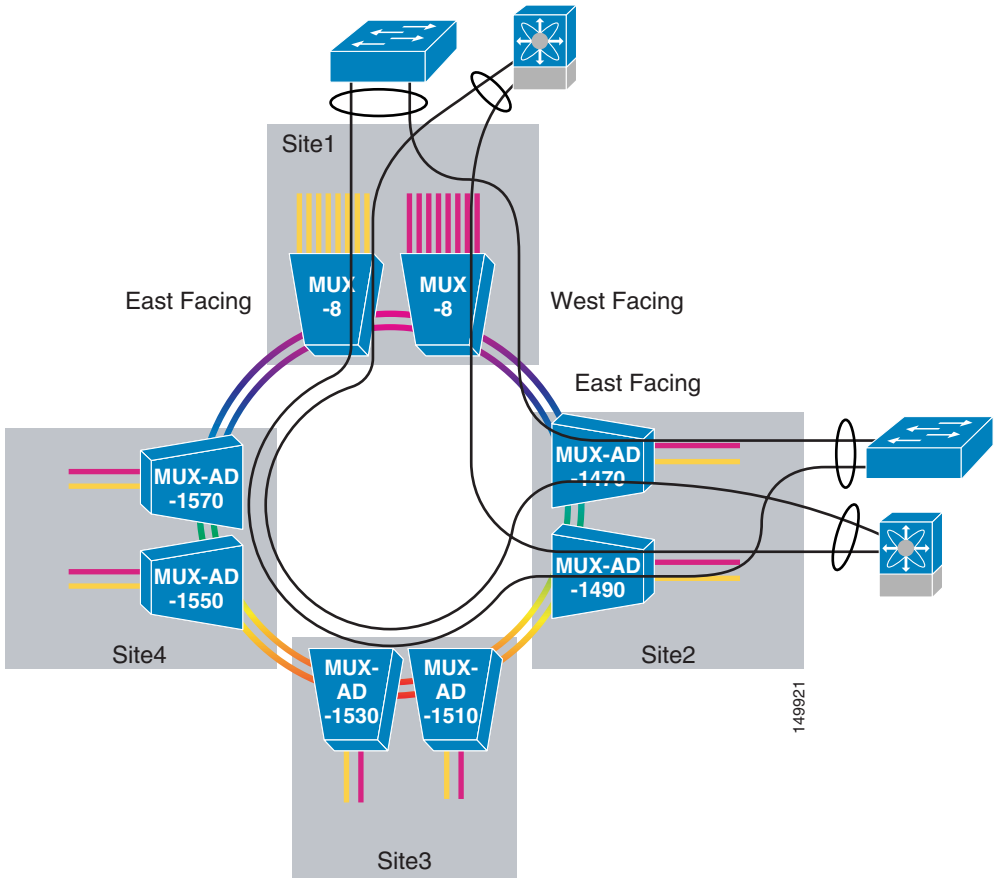
Figure 3-26 CWDM Ring Topology for Hub-and-Spoke Deployments



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

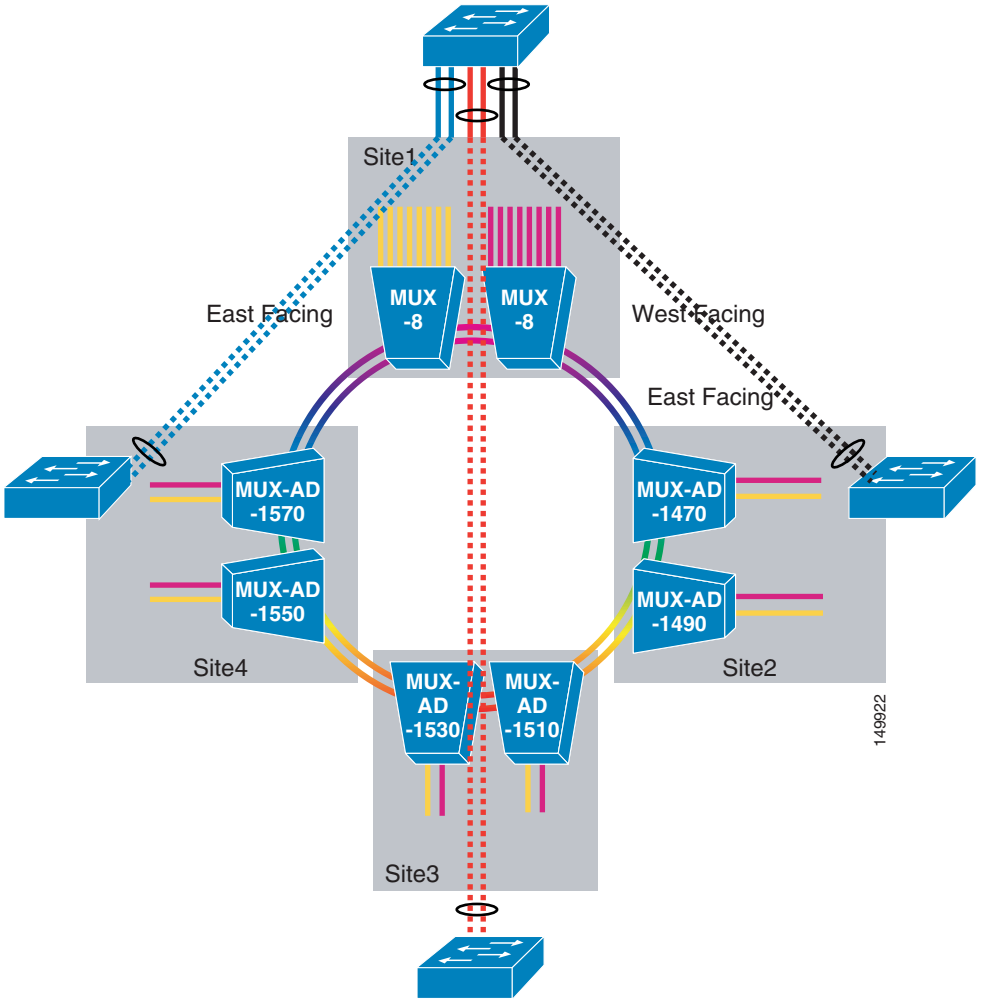
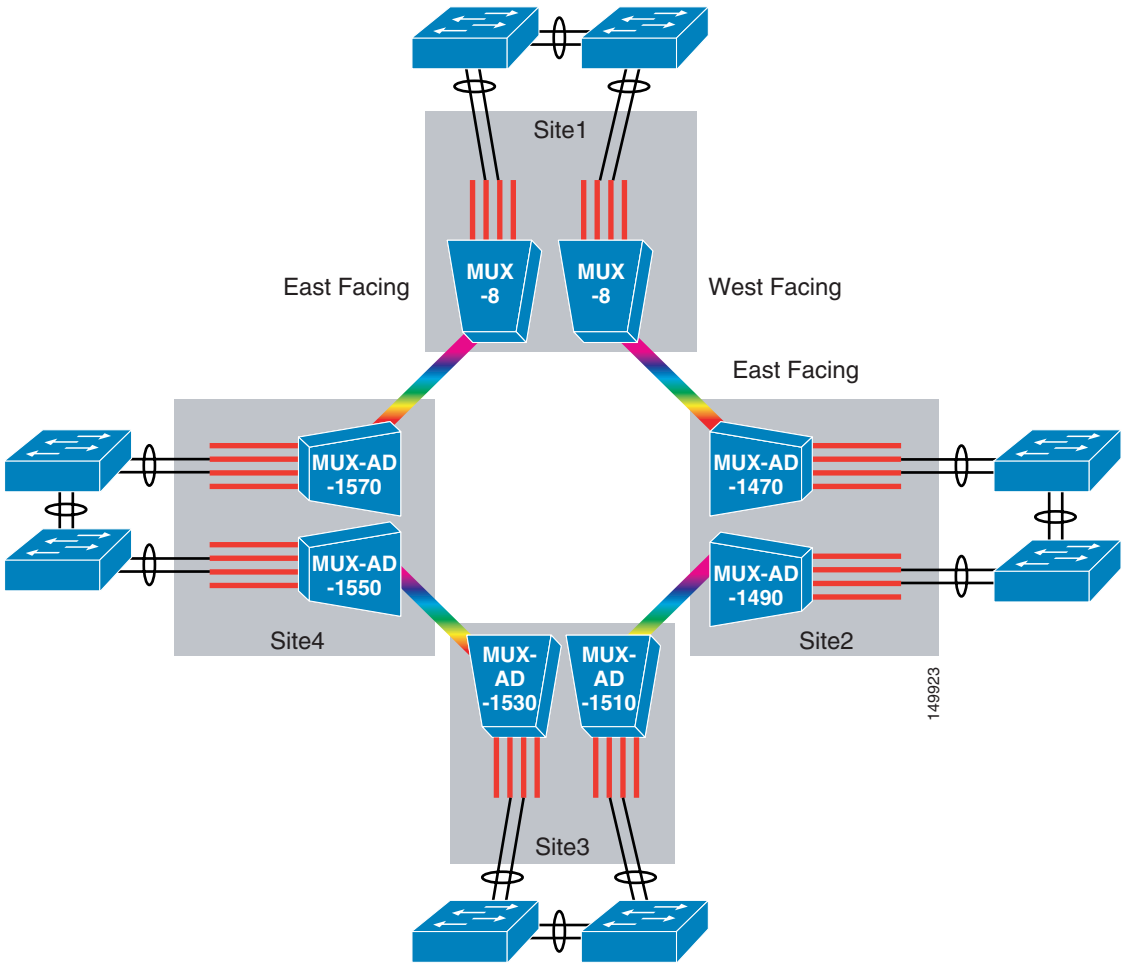


Figure 3-29 shows a topology with no single point of failure, which relies on spanning tree to provide a redundant Layer 2 path when one of the links fails. This topology is simply made of point-to-point links between the sites; these links are multiplexed with CWDM muxes.

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

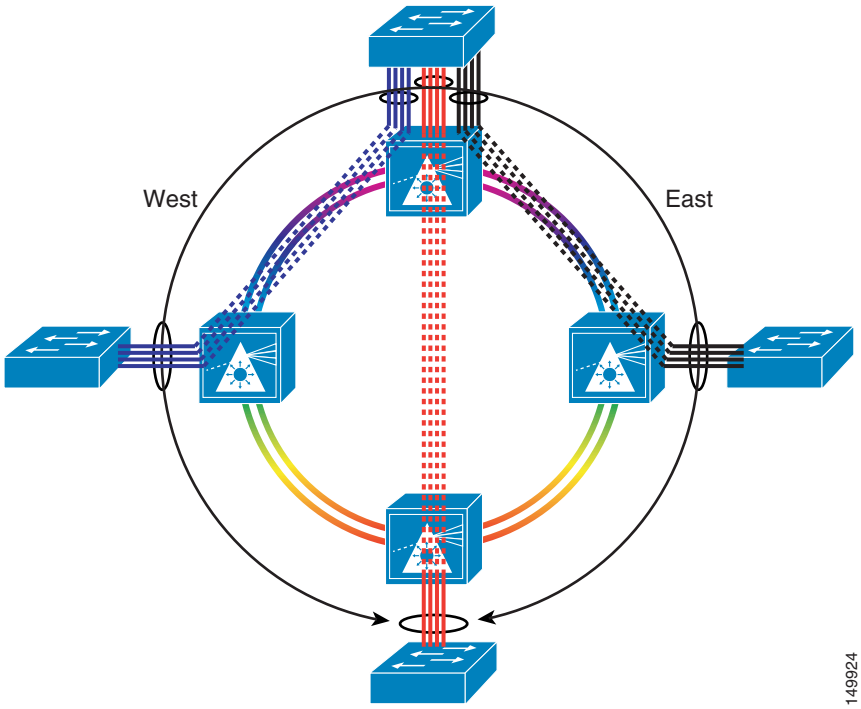
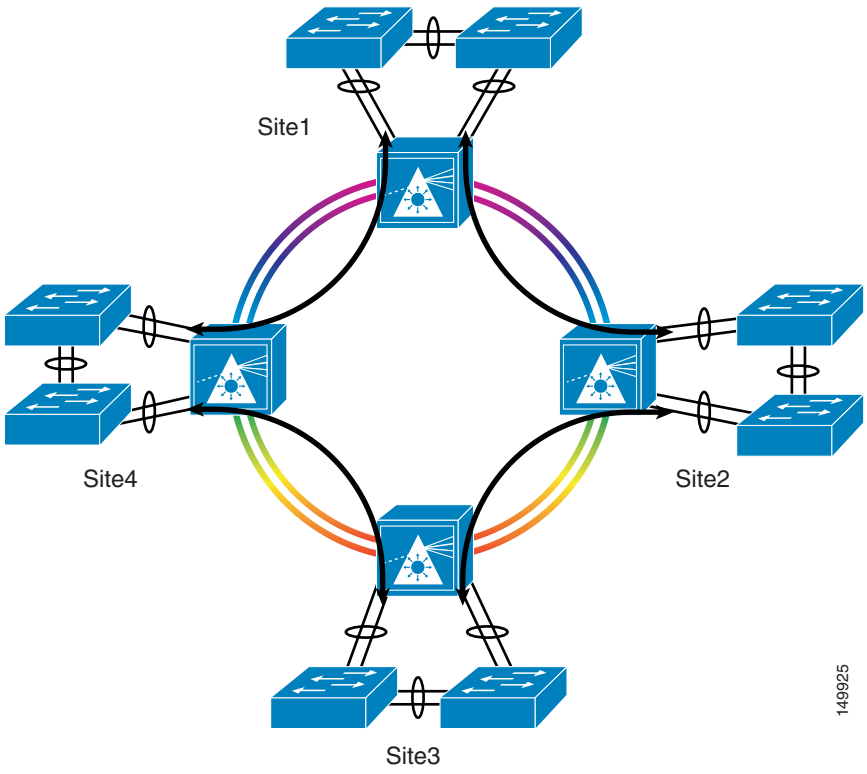


Figure 3-31 shows a ring topology with DWDM.

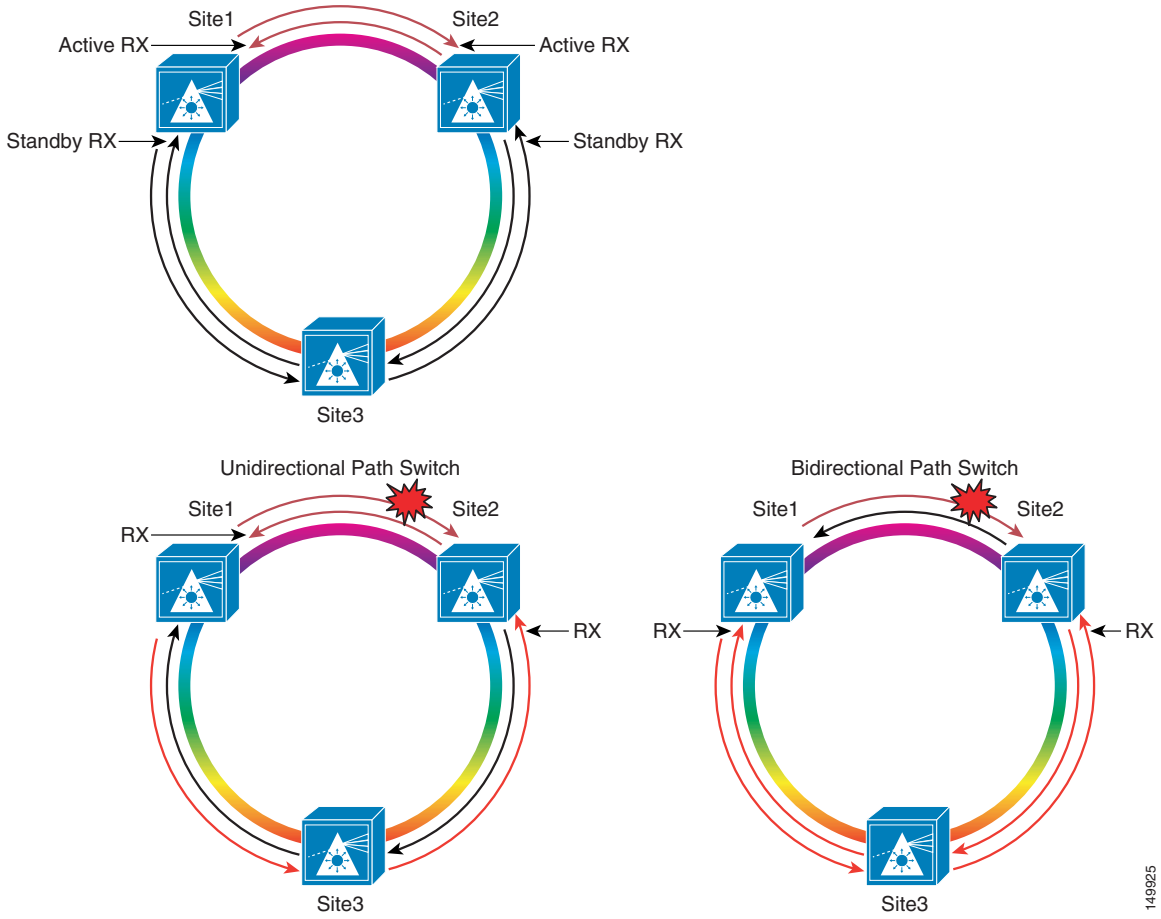
Differently from the CWDM topology, a DWDM ring can provide additional protection mechanisms than client protection.

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.

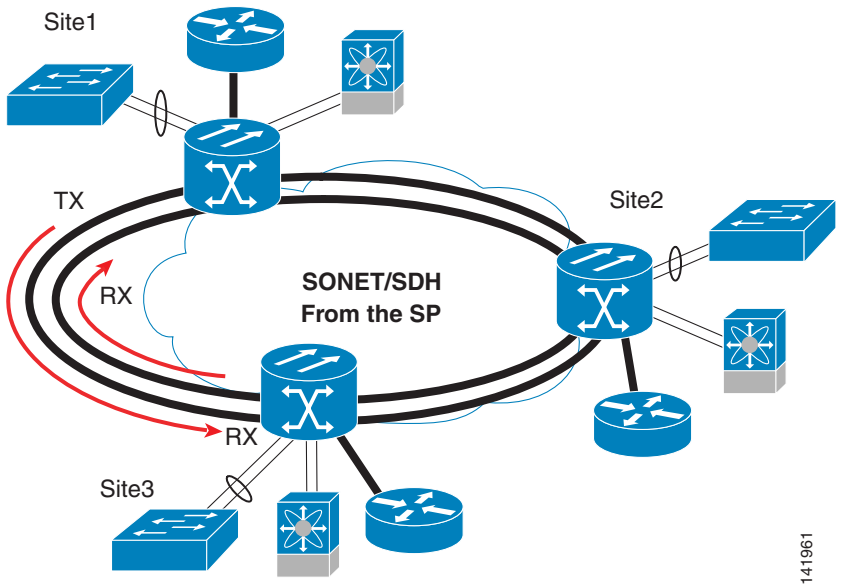
Figure 3-32 DWDM Protection Mechanisms



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.

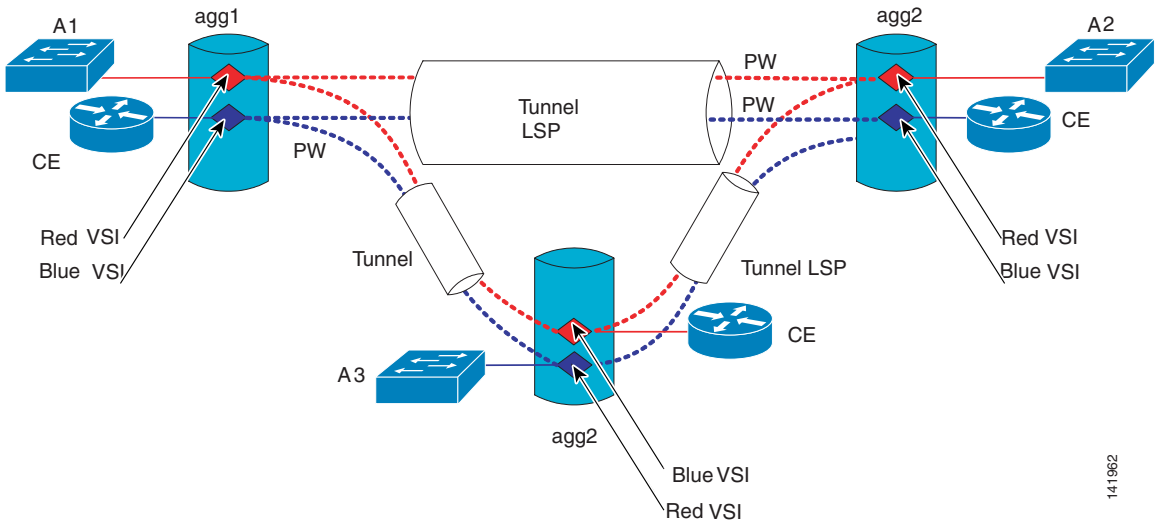
With SRP/RPR, the EtherChannels can be terminated locally on the Cisco ONS 15454 device. The ring appears to the Ethernet switches as a Layer 2 shared link. Neither spanning tree nor EtherChannel need to be carried across sites.

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)

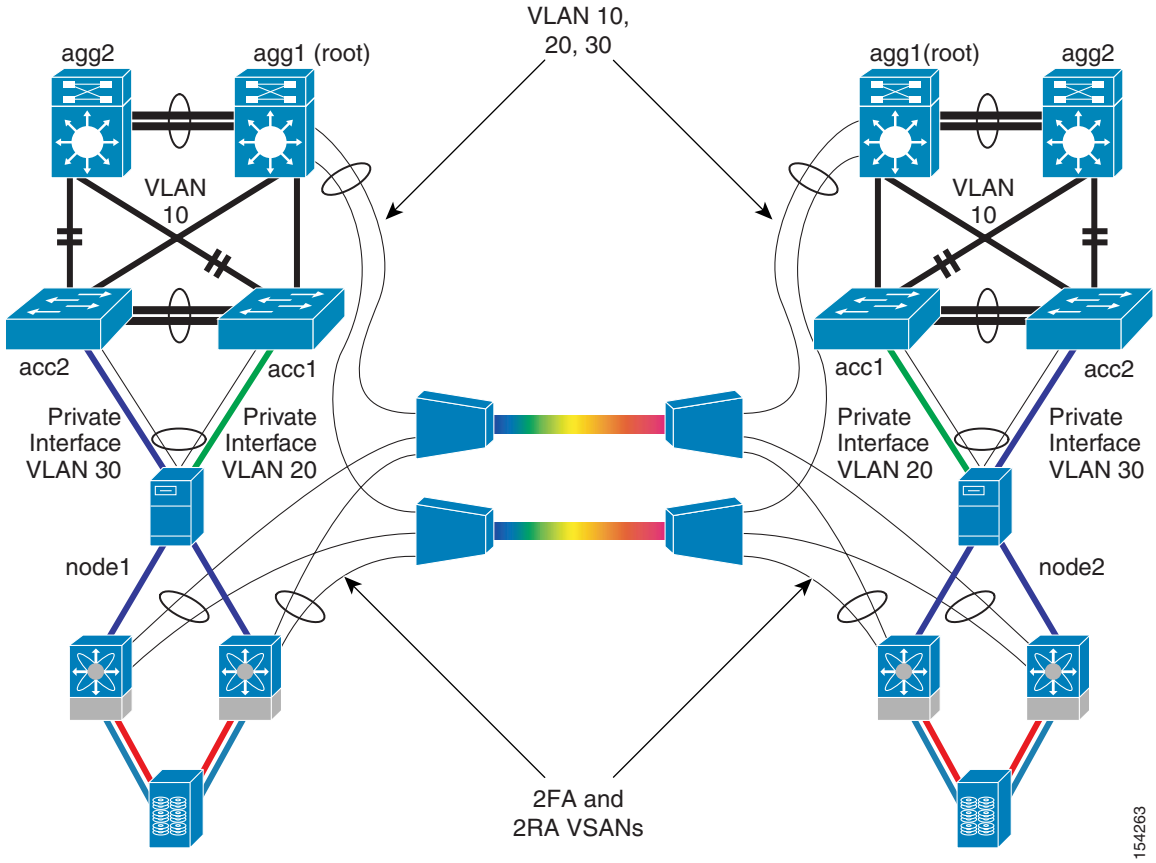
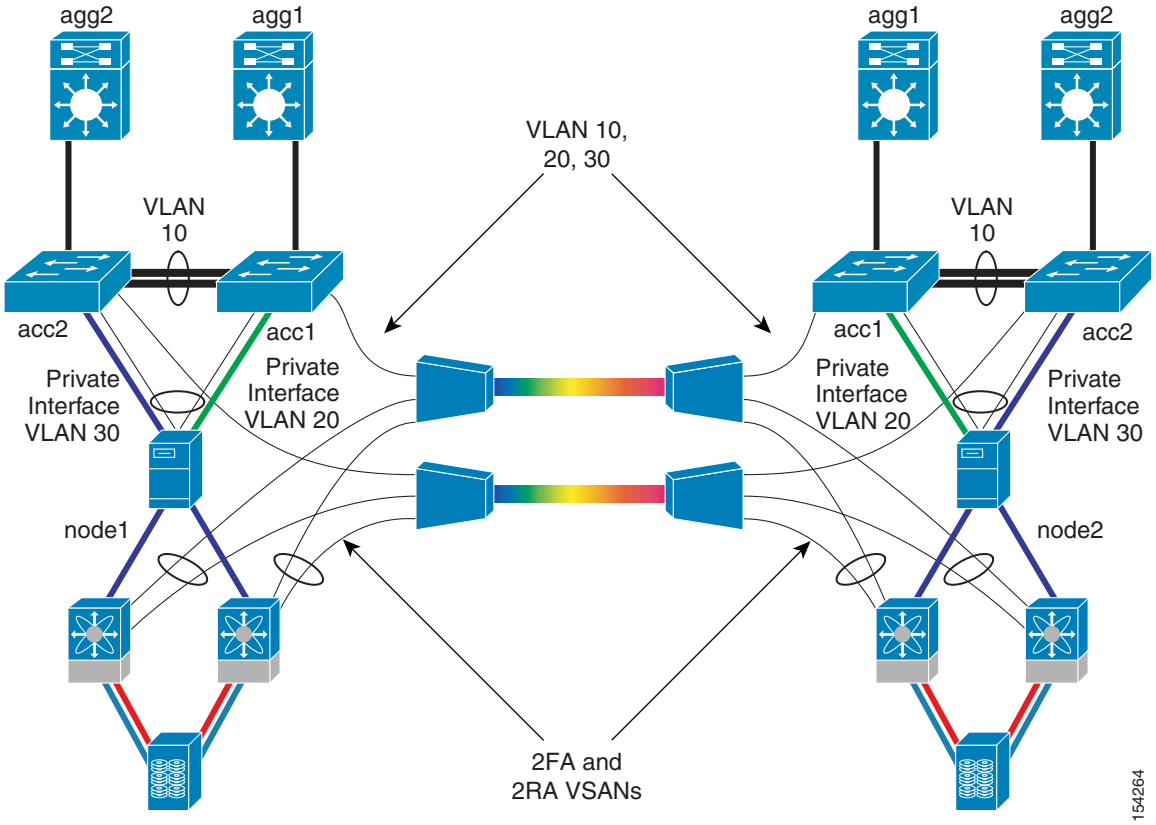


Figure 3-36 shows the cluster extension in the presence of a looped access.

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'.
```

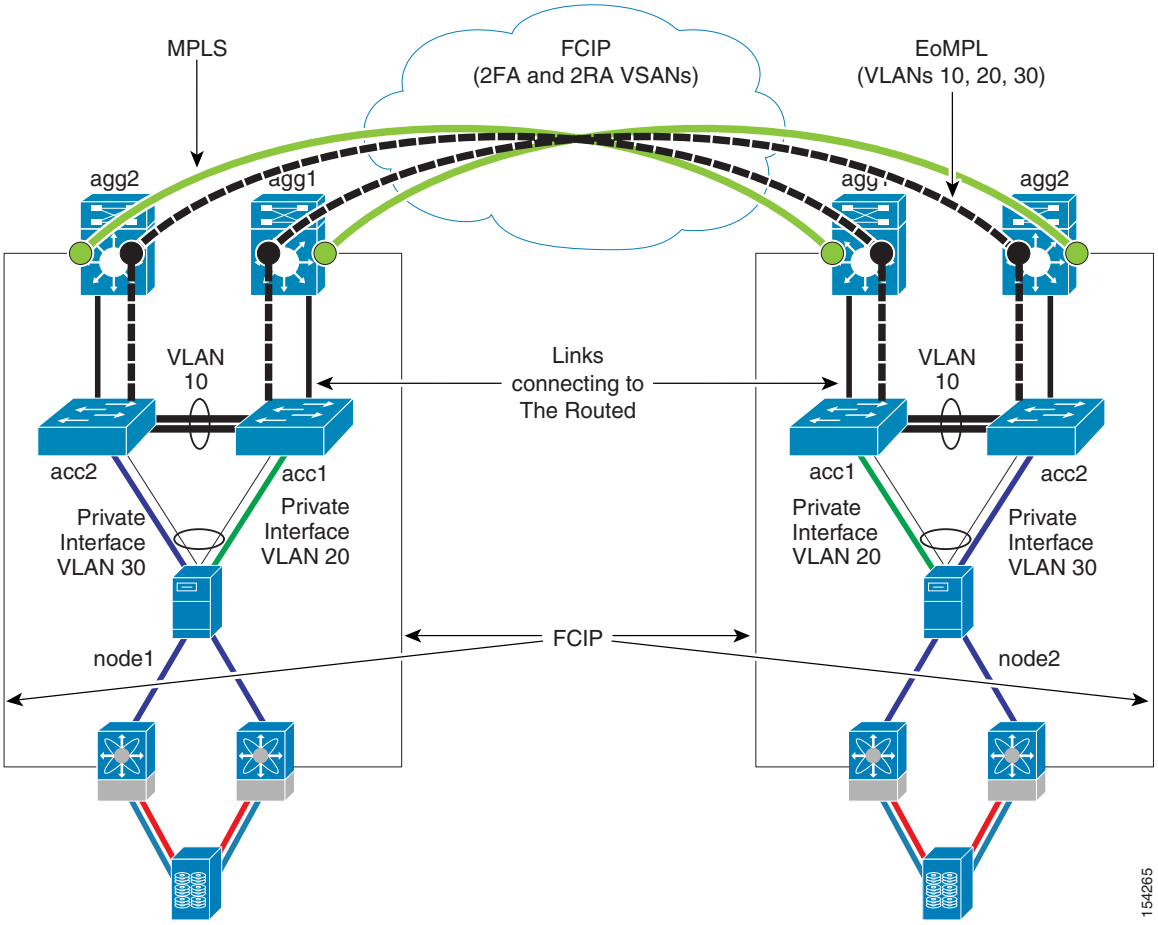
Note Note that by just failing over the disk, there is no automatic synchronous replication from site2 to site1 unless you also change the disk role of DiskArray2 from R2 to R1

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.

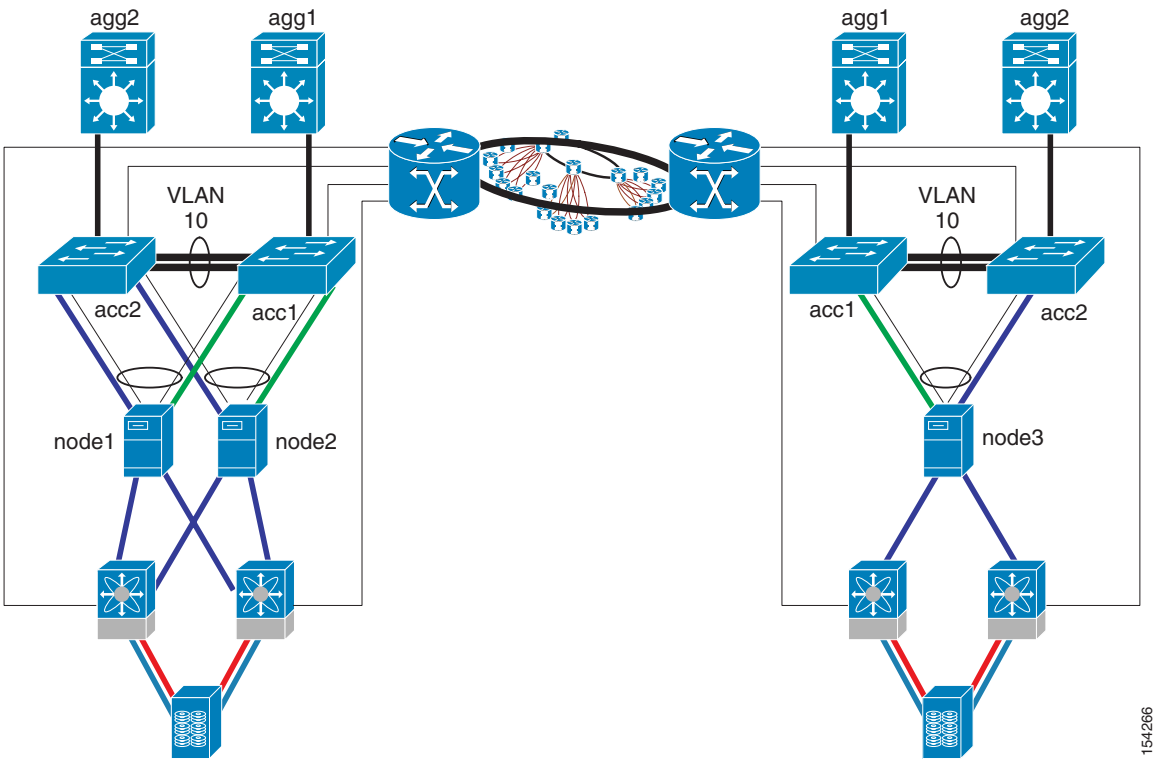
The following is an example of continental cluster configuration:

- 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.

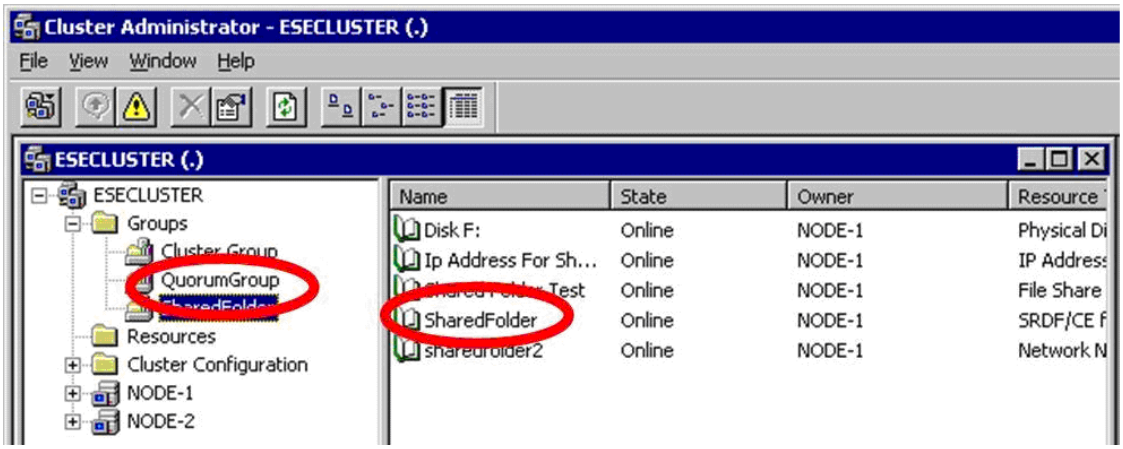
Software-Assisted Disk Failover

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



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

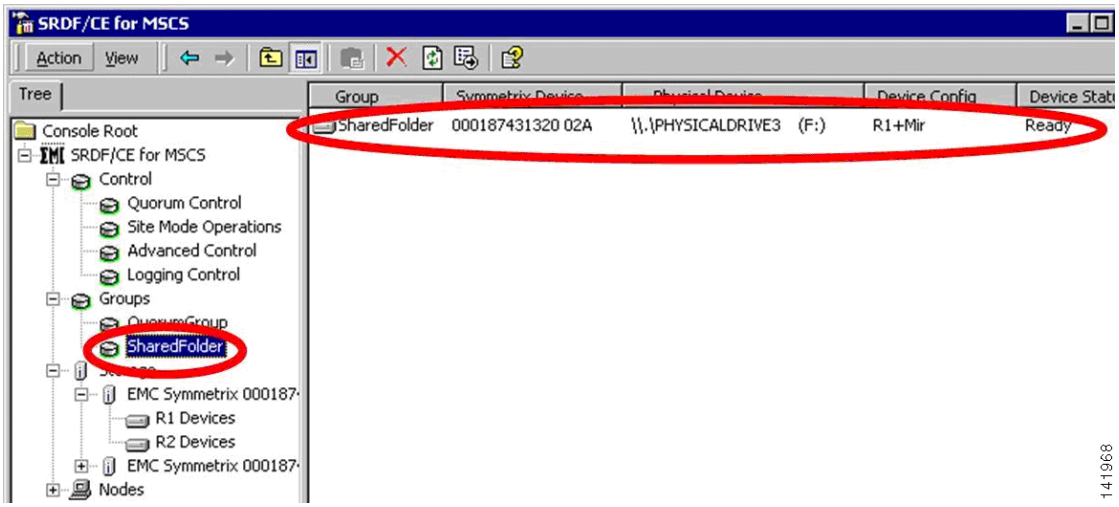
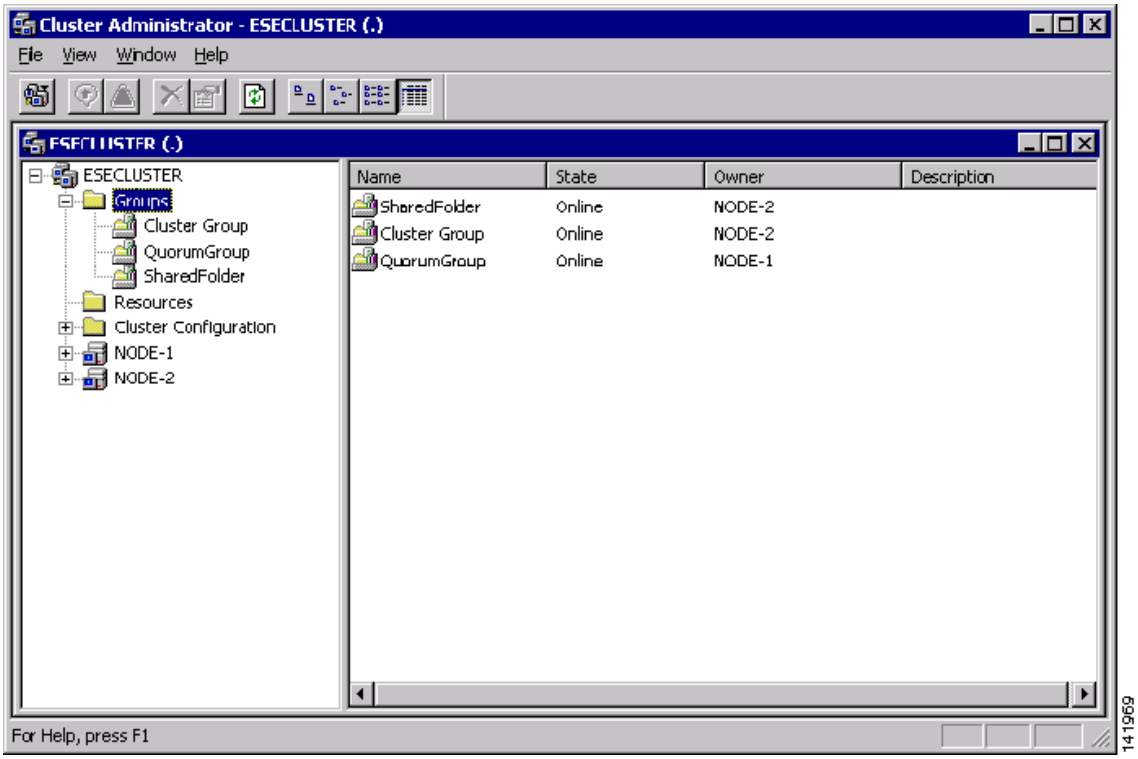


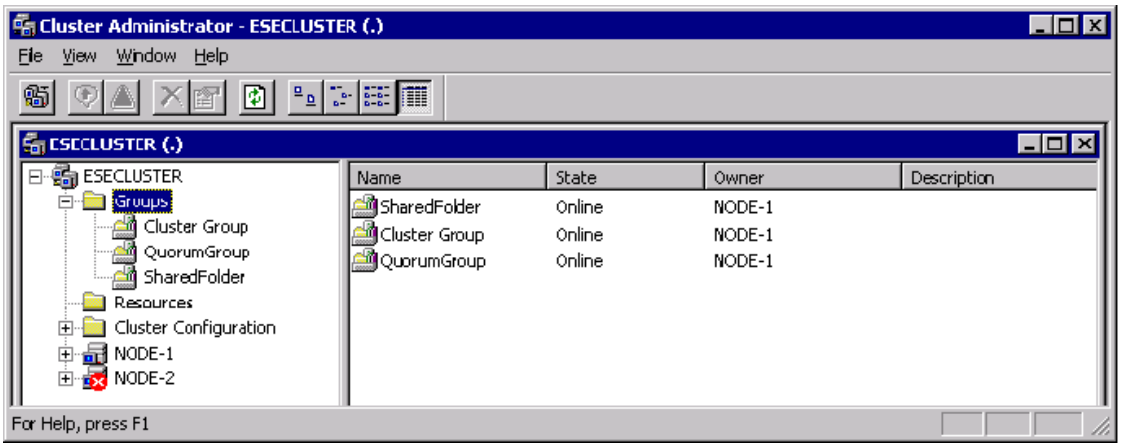
Figure 3-41 shows the failover of the Shared Folder disk when the NIC on the primary node fails, while the quorum group is still owned by the primary node. The owner for the shared folder disk is node2, while the owner for the quorum is still node1.

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



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



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:

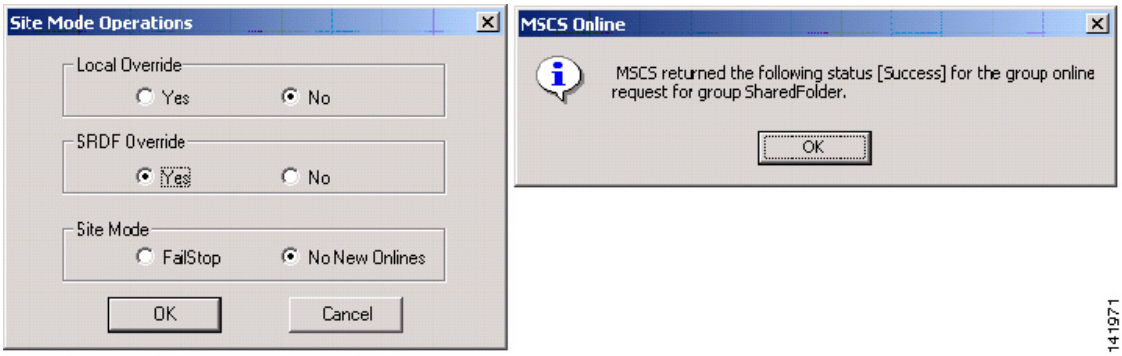
- Lost connectivity between the sites (in which case failover is not necessary)
- Complete failure of the primary site

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



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

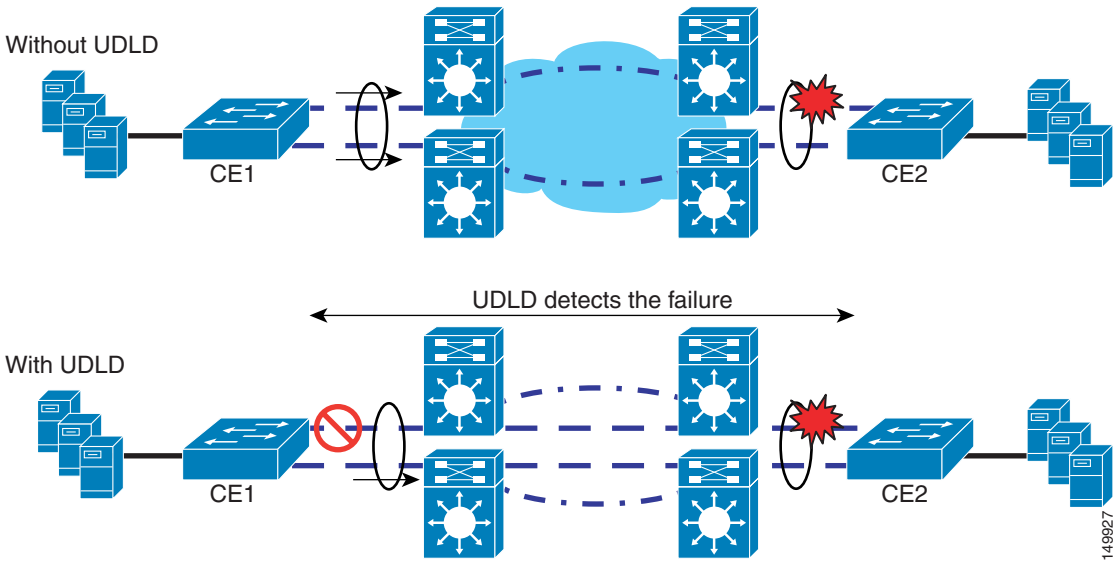
Layer 2 LAN extension means that either spanning tree or EtherChannels need to be used. Running spanning tree across sites works fine, but many customers prefer to design loop-free topologies from the very beginning to avoid dealing with Layer 2 loops caused by misconfigurations. Port channeling can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

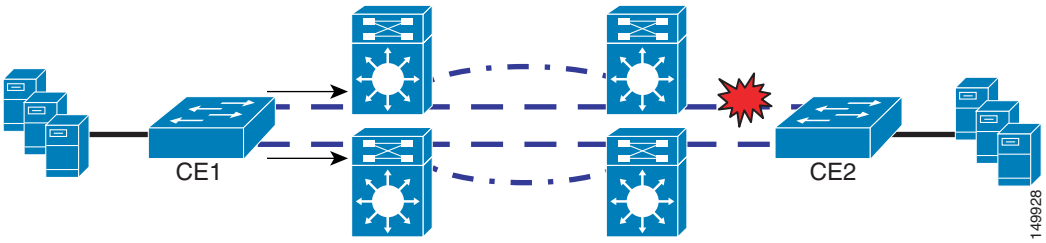
Note Recovery from a far-end failure is not automatic. When the link is recovered, you need to manually ensure that both ports on the local and remote site are shut/unshut to restore complete connectivity.

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.

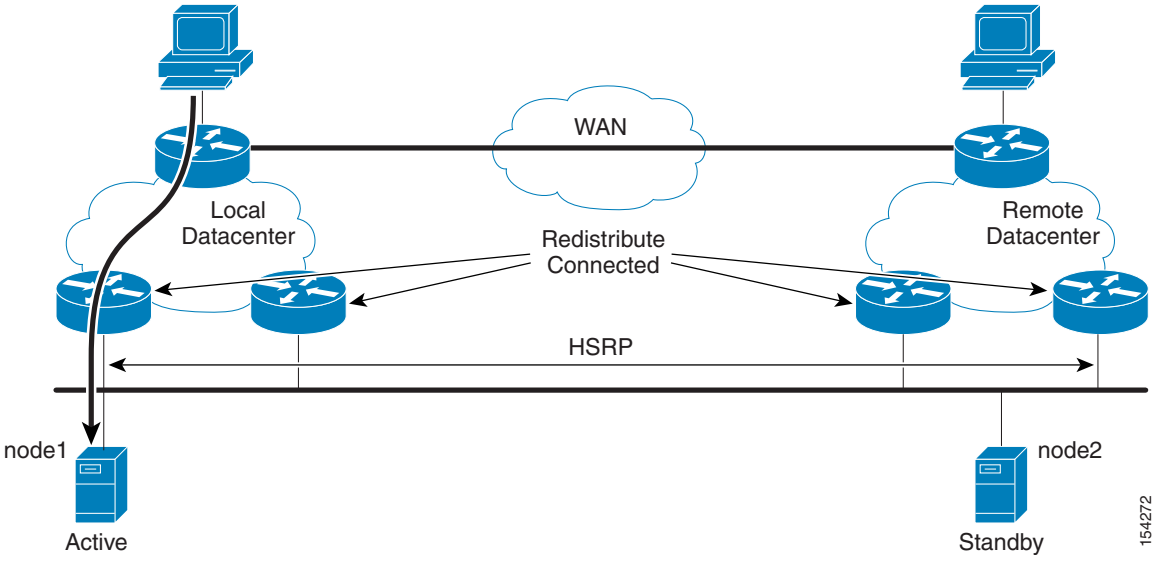
Routing Design

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
```

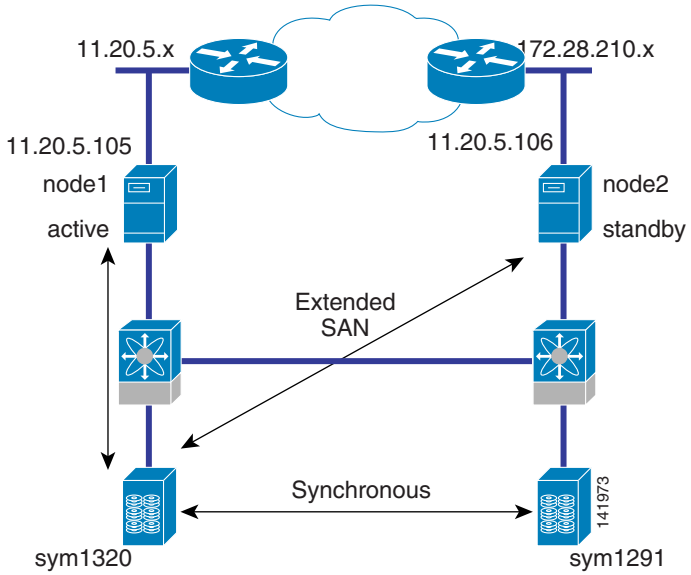
Local Area Mobility

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.

FCIP over IP/MPLS Core

This chapter discusses the transport of Fibre Channel over IP (FCIP) over IP/Multiprotocol Label Switching (MPLS) networks and addresses the network requirements from a service provider (SP) perspective. This chapter also describes service architectures and storage service offerings using FCIP as a primary storage transport mechanism.

Overview

Storage extension solutions offer connectivity between disparate storage "islands," and promote transport solutions that are specifically geared towards carrying storage area network (SAN) protocols over WAN and MAN networks. This emerging demand is providing a new opportunity for carriers. SPs can now deliver profitable SAN extension services over their existing optical (Synchronous Optical Network [SONET]/Synchronous Digital Hierarchy [SDH] and Dense Wavelength Division Multiplexing [DWDM]) or IP infrastructure. DWDM networks are ideal for high-bandwidth, highly resilient networks and are typically deployed within metro areas. Transporting storage traffic over the existing SONET/SDH infrastructure allows SPs to maximize the use of their existing SONET/SDH ring deployments. Some applications do not mandate stringent requirements offered by optical networks. These applications can be easily transported over IP networks using FCIP interfaces. The obvious advantage of transporting storage over IP is the ubiquitous nature of IP.

Disk replication is the primary type of application that runs over an extended SAN network for business continuance or disaster recovery. The two main types of disk replication are array-based (provided by EMC^2 SRDF, Hitachi True Copy, IBM PPRC XD, or HP DRM, and host-based (for example, Veritas Volume Replicator). Both disk replication types run in synchronous and asynchronous modes. In synchronous mode, an acknowledgement of a host-disk write is not sent until a copy of the data to the remote array is completed. In asynchronous mode, host-disk writes are acknowledged before the copy of the data to the remote array is completed.

Applications that use synchronous replication are highly sensitive to response delays and might not work with slow-speed or high-latency links. It is important to consider the network requirements carefully when deploying FCIP in a synchronous implementation. Asynchronous deployments of FCIP are recommended in networks with latency or congestion issues. With FCIP, Fibre Channel SAN can be extended anywhere an IP network exists and the required bandwidth is available. FCIP can be extended over metro, campus, or intercontinental distances using MPLS networks. FCIP may be an ideal choice for intercontinental and coast-to-coast extension of SAN.

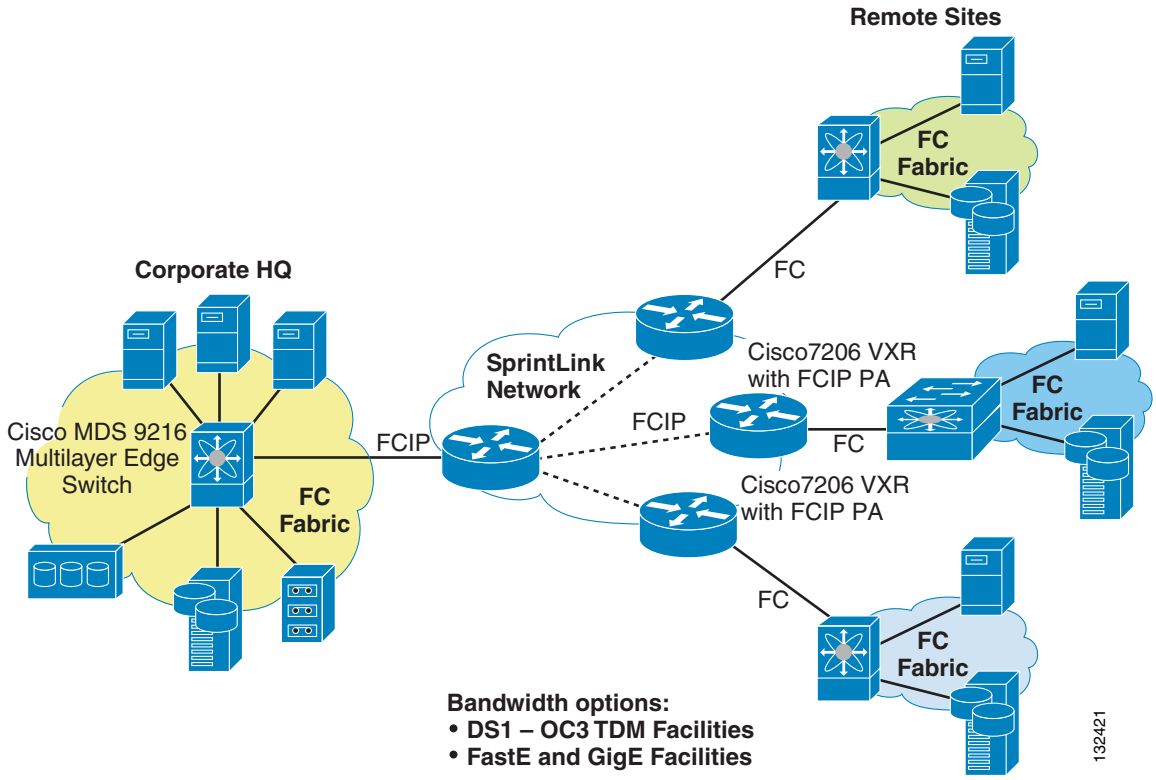
Typical Customer Requirements

Small-to-medium businesses (SMBs) represent about 90 percent of all companies in the United States. These companies typically employ a few hundred employees and are highly focused on their core services or products. They usually lack IT expertise and manpower to develop, deploy, and maintain LAN, WAN, and SAN infrastructures. A typical SMB may have one or two offices in multiple metro areas with one head office (corporate office) in one of the metro areas. The corporate office is considered as the *home office*, where the majority of the business activities occur, and other locations are usually designated as *satellite* offices where most of the activities are directed to the home office.

These SMBs use IP connectivity to connect between satellite and home offices. Connectivity varies from Frame Relay, TI, and fractional Ethernet, depending on the demand and size of the SMB. Currently, these networks are used to carry data and voice traffic. A similar connectivity is considered for storage, but is not currently installed because of cost constraints. Most of the data at the home office location may be consolidated into a local SAN, and the data at the satellite offices can be consolidated into small SAN islands. This introduces the problem of storage connectivity between SAN islands for disaster recovery and business continuance. There are several options to interconnect the SAN, but the IP network is the ideal choice because of its availability at the client site and its comparatively low cost.

Figure 4-1 shows a typical customer SAN extension through an SP network.

Figure 4-1 SAN Extension Through SP Network



In most cases, the SMB customers have connectivity that is less than DS3 speed and, in some cases, may be up to OC-3 speeds. Therefore, in some cases, compressing the Fibre Channel data *before* transporting might become a requirement. In any network, security is key to protect valuable data from being misused by intruders. FCIP traffic must be secured before transporting it across the SP network.

The requirements are as follows:

- FCIP transport over an optimized IP/MPLS network
- Some type of compression mechanism (software or hardware)
- Security mechanism (IPSec, encryption, and VPN networks)
- End-to-end management of FCIP traffic

Compression

The primary objective of compression is to reduce the amount of overall traffic on a particular WAN link. This is achieved when a data rate equal to the WAN link speed is compressed, thereby reducing the total amount of data on the WAN link. In this case, non-compressed storage data requires all of the 45 Mb/sec DS3 WAN connection. By enabling compression on the storage data (assuming an average of 2 to 1 compression), the effective utilization of the WAN link by storage traffic would be 22.5 Mb/sec. This allows the WAN link to be used by other IP traffic. The second objective for compression may be to carry more data over a WAN link than it is normally capable of carrying. An example of this is to compress a 90-Mbps Fibre Channel data stream and carry it over a 45-Mbps WAN link (still assuming an average of compression ratio of 2 to 1).

There are several types of compression algorithms. The most common type used in data networks is lossless data compression (LZS). This type of compression converts the original data into a compressed format that then can be restored into the original data. The service adapter modules (7600-SA-VAM, SA-VAM2) and the storage services module (MDS-IPS-8 IP) use the IP Payload Compression Protocol (IPPCP)/LZS (RFC 2395) algorithm for compressing data.

The LZS compression algorithm works by searching for redundant data strings in the input data stream and then replaces these strings with data tokens that are shorter in length than the original data. A table is built of these string matches, pointing to previous data in the input stream. The net result is that future data is compressed based on previous data. The more redundant the data in the input stream, the better the compression ratio. Conversely, the more random the data, the worse the compression ratio will be.

The compression history used by LZS is based on a sliding window of the last 2000 bytes of the input stream. When the data is transmitted, it contains both literal data and compressed tokens. Literal data are input data streams that cannot be compressed and are transmitted uncompressed. Compressed tokens are pointer offsets and data length that point to the compression history table. The remote side rebuilds the data from the compressed history table based on the pointers and length fields.

Note A full description of IPPCP and LZS are available in RFC 2395 and in ANSI X.3241-1994.

Compression Support in Cisco MDS

Both software- and hardware-based compression are supported by the Cisco MDS product line. Depending on the SAN-OS version and the hardware used, customers can determine which compression methods apply.

The software-based compression solution is available on the IPS-IP Storage Service Module for the Cisco MDS 9216/MDS 9216i fabric switch and the Cisco MDS 9500 series storage directors. This feature is available in SAN-OS version 1.3(2a) and later releases. The software-based compression is available on each of the eight IPS-8 Gigabit Ethernet ports. The number of Gigabit Ethernet ports used on the IPS does not affect the performance of the compression with this feature enabled.

Hardware-based compression is available with SAN-OS version 2.0 and with new hardware (MDS 9216i/MLS14/2). Compression is applied per FCIP interface (tunnel) with a variety of modes available. Beginning with SAN-OS 2.0, three compression modes are configurable with additional support for the MPS-14/2 module.

Compression Modes and Rate

In SAN-OS 1.3, the following two compression modes can be enabled per FCIP interface on the IPS-4 and IPS-8:

- High throughput ratio—Compression is applied to outgoing FCIP packets on this interface with higher throughput favored at the cost of a slightly lower compression rate.
- High compression ratio—Compression is applied to outgoing FCIP packets on this interface with a higher compression ratio favored at the cost of a slightly lower throughput.

In SAN-OS 2.0, three compression modes are available per FCIP interface on the IPS-4, IPS-8, and MPS-14/2:

- Mode 1—Equivalent to the high throughput ratio of SAN-OS 1.3. Use Mode 1 for WAN paths up to 100 Mbps on the IPS-4 and IPS-8; and WAN paths up to 1 Gbps on the MPS-14/2.
- Mode 2—Higher compression ratio than Mode1, but applicable only to slow WAN links up to 25 Mbps.
- Mode 3—Higher compression ratio than Mode 1 and slightly higher than Mode 2. Applicable to very slow WAN links up to 10 Mbps.

The following are the software-based compression options for FCIP for the Cisco MDS 9000 IP Storage Services Module:

- SAN-OS 1.3—Two algorithms: high throughput and high compression
- SAN-OS 2.0—Three algorithms: Modes 1–3

The following is the hardware- and software-based compression and hardware-based encryption for FCIP for the Cisco MDS 9000 Multi-Protocol Services module:

- SAN-OS 2.0—Three algorithms: Modes 1–3

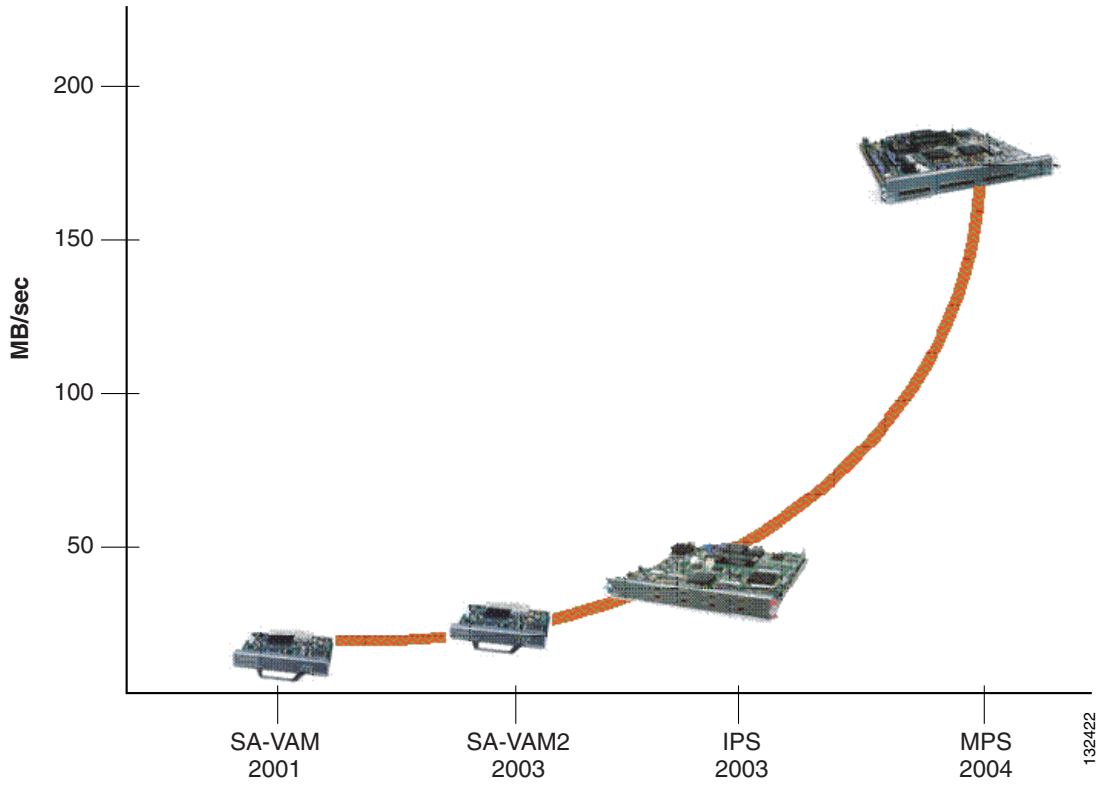
The choice between these solutions should be based on the following factors:

- Available link speed or bandwidth
- Choice of FCIP solution (IPS-8/4 FCIP, MPS-14/2, or PA-FC-1G port adapter)
- New or existing SAN fabric installations

Note For more information, see the following: LZS (RFC 1974), IPPCP with LZS (RFC 2395), Deflate (RFC 1951), and IPPCP with Deflate (RFC 2394).

Figure 4-2 shows a comparison of the Cisco compression solutions.

Figure 4-2 Cisco Compression Solutions



The following performance data applies to Figure 4-2:

- VAM—9.9–12 MB/sec –10.9 MB/sec average
- VAM2—19.7–25.4 MB/sec – 19 MB/sec average
- IPS—18.6–38.5 MB/sec – 24.6 MB/sec average
- MPS—136–192 MB/sec – 186.5 MB/sec average

Security

The security of the entire Fibre Channel fabric is only as good as the security of the entire IP network through which an FCIP tunnel is created. The following scenarios are possible:

- Unauthorized Fibre Channel device gaining access to resources through normal Fibre Channel processes
- Unauthorized agents monitoring and manipulating Fibre Channel traffic that flows over physical media used by the IP network

Security protocols and procedures used for other IP networks can be used with FCIP to safeguard against any known threats and vulnerabilities. FCIP links can be secured by the following methods:

- Using the IPSec Security Protocol Suite with encryption for cryptographic data integrity and integrity of authentication

- SPs providing VPN service to transport FCIP traffic to provide additional security
- Using an MPLS extranet for application-specific security

Cisco Encryption Solutions

For selecting compression solutions for FCIP SAN extension, a user needs to determine the requirements for the encryption solution. These requirements may include the speed of the link that needs encryption, the type of encryption required, and the security requirements of the network. Cisco offers three hardware-based encryption solutions in the data center environment. The SA-VAM and SA-VAM2 service modules for the Cisco 7200 VXR and 7400 series routers and the IPSec VPN Services Module (VPNSM) for the Catalyst 6500 switch and the Cisco 7600 router.

Each of these solutions offers the same configuration steps, although the SA-VAM2 and IPSec VPNSM have additional encryption options. The SA-VAM and SA-VAM2 are used only in WAN deployments, whereas the IPSec VPNSM can support 1.6 Gb/sec throughput, making it useful in WAN, LAN, and MAN environments.

The SA-VAM is supported on the 7100, 7200 VXR, and 7401 ASR routers with a minimum Cisco IOS version of 12.1(9)E or 12.1(9)YE. For use in the 7200 VXR routers, the SA-VAM has a bandwidth cost of 300 bandwidth points. The SA-VAM has a maximum throughput of 140 Mps, making it suitable for WAN links up to DS3 or E3 line rates.

The SA-VAM2 is supported on the 7200 VXR routers with a minimum Cisco IOS version of 12.3(1). The SA-VAM2 has a bandwidth cost of 600 bandwidth points. The SA-VAM2 has a maximum throughput of 260 Mps, making it suitable for WAN links up to OC-3 line rates.

The IPSec VPNSM is supported on the Catalyst 6500 switch and the Cisco 7600 router with a minimum Native IOS level of 12.2(9)YO. For increased interoperability with other service modules and additional VPN features, it is recommended that a minimum of 12.2(14)SY be used when deploying this service module.

The choice between these solutions should be based primarily on the following two factors:

- Available link speed or bandwidth
- Security encryption policies and encryption methods required

The Cisco MDS 9000 with MLS14/2 and the Cisco 9216i support encryption with no performance impact. The MPS Service Module and the Cisco 9216i support line rate Ethernet throughput with AES encryption.

The following are encryption methods supported per module:

- SA-VAM—DES, 3DES
- SA-VAM2—DES, 3DES, AES128, AES192, AES256
- VPNSM—DES, 3DES
- MDS MPS—DES, 3DES, AES192

Note An encrypted data stream is not compressible because it results in a bit stream that appears random. If encryption and compression are required together, it is important to compress the data before encrypting it.

Write Acceleration

Write Acceleration is a configurable feature introduced in SAN-OS 1.3 that enhances FCIP SAN extension with the IP Storage Services Module. Write Acceleration is a SCSI protocol spoofing mechanism that improves application performance by reducing the overall service time for SCSI write input/output (I/O) operations and replicated write I/Os over distance. Most SCSI Fibre Channel Protocol (FCP) write I/O exchanges consist of two or more round trips between the host initiator and the target array or tape. Write Acceleration reduces the number of FCIP WAN round trips per SCSI FCP write I/O to one.

Write Acceleration is helpful in the following FCIP SAN extension scenarios:

- Distance and latency between data centers inhibits synchronous replication performance and impacts overall application performance.
- Upper layer protocol chattiness inhibits replication throughput, and the underlying FCIP and IP transport is not optimally utilized.
- Distance and latency severely reduces tape write performance during remote tape backup because tapes typically allow only a single outstanding I/O. Write Acceleration can effectively double the supported distance or double the transfer rate in this scenario.
- Shared data clusters are stretched between data centers and one host must write to a remote storage array.

The performance improvement from Write Acceleration typically approaches 2 to 1, but depends upon the specific situation.

Write Acceleration increases replication or write I/O throughput and reduces I/O response time in most situations, particularly as the FCIP Round Trip Time (RTT) increases. Each FCIP link can be filled with a number of concurrent or outstanding I/Os. These I/Os can originate from a single replication source or a number of replication sources. The FCIP link is filled when the number of outstanding I/Os reaches a certain ceiling. The ceiling is mostly determined by the RTT, write size, and available FCIP bandwidth. If the maximum number of outstanding I/Os aggregated across all replication sessions (unidirectional) is less than this ceiling, then the FCIP link is underutilized and thus benefits from Write Acceleration.

Using FCIP Tape Acceleration

FCIP Tape Acceleration is a new feature introduced in SAN-OS 2.0 to improve remote tape backup performance by minimizing the effect of network latency or distance on remote tape applications. With FCIP Tape Acceleration, the local Cisco MDS 9000 IPS or MPS module proxies as a tape library. The remote MDS 9000, where the tape library is located, proxies as a backup server.

Similar to Write Acceleration, the MDS 9000 recognizes and proxies elements of the upper level SCSI protocol to minimize the number of end-to-end round trips required to transfer a unit of data and to optimally use the available network bandwidth. FCIP Write Acceleration achieves this by proxying the SCSI Transfer Ready and Status responses (in contrast, Write Acceleration proxies the Transfer Ready only). Write Filemarks and other non-write operations are not proxied and are passed directly to the remote tape library. The Write Filemarks operation corresponds to a checkpoint within the tape backup application. This is typically a tunable parameter but may default to 100 or 200 records depending upon the tape backup product.

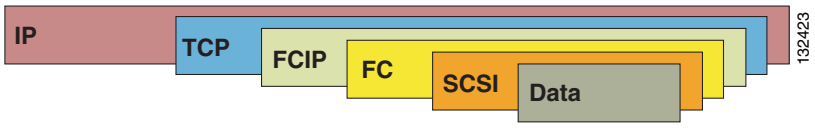
FCIP Tape Acceleration maintains data integrity in the event of a variety of error conditions. Link errors and resets are handled through Fibre Channel-tape Ethernet LAN services (ELS) recovery mechanisms. Should the remote tape unit signal an error for an I/O that the status has already been returned to "good", a Deferred Error is signaled to the tape backup application. The backup application either corrects the error and replays the command or rolls back to the previous file mark and replays all I/Os from that point.

You can enable FCIP Tape Acceleration on any FCIP interface on the Cisco IPS-4, IPS-8, and MPS-14/2 modules, or the Gigabit Ethernet interfaces on the Cisco MDS 9216i.

FCIP

FCIP encapsulates Fibre Channel frames and transports these frames within TCP packets. The FCIP tunnel acts as an Inter-Switch Link (ISL) between two fabric switches. The endpoint devices detect each other as they would between two local switches interconnected with standard ISL. FCIP endpoints are associated to virtual e-ports and these ports communicate with themselves and exchange information such as reconfigure fabric (RCF), Fabric Shortest Path First (FSPF), build fabric (BF), and so on. FCIP relies on the TCP/IP protocol to provide contention control and orderly delivery of packets. Figure 4-3 shows the FCIP encapsulation process.

Figure 4-3 FCIP Encapsulation



TCP Operations

TCP implemented on traditional servers or hosts tends to overreact to packet drops. The throttling back that occurs in the traditional TCP implementation is not acceptable to storage traffic. The TCP stack implemented for FCIP (in the Cisco MDS 9000) is optimized for carrying storage traffic by reducing the probability of drops and increasing the resilience to drops when they occur.

Fibre Channel traffic can be highly bursty, and traditional TCP can amplify that burstiness. With traditional TCP, the network must absorb these bursts through buffering in switches and routers. Packet drops occur when there is insufficient buffering at these intermediate points. To reduce the probability of drops, the FCIP TCP implementation reduces the burstiness of the TCP traffic that leaves the Gigabit Ethernet interface.

In the FCIP TCP stack, burstiness is limited through the use of variable rate, per-flow shaping, and by controlling the TCP congestion window size. After idle or partially idle periods, the FCIP interface does not send large packet bursts at Gigabit interface speeds. If not controlled, large Gigabit Ethernet bursts can overflow downstream routers or switches and speed mismatches can occur. For example, a Gigabit Ethernet feeding into a DS3 (45 Mbps) link through a router may overflow the router buffers unless the traffic is controlled or shaped in a way that the router can handle the transmission.

TCP Parameters

TCP parameters may require adjustments when implementing SAN extension that uses FCIP. This section provides general information and recommendations for key TCP parameters that require adjustments. The following parameters are considered:

- TCP window size
- TCP maximum bandwidth
- TCP minimum available bandwidth
- Round Trip Time (RTT)

TCP Window Size

TCP uses a sliding window to control the flow of data from end to end. The TCP maximum window size (MWS) is the maximum amount of data the sender allows to be outstanding without acknowledgment at one time. The minimum MWS is 14 KB; the maximum is 32 MB.

The sender can use a larger window size to allow more outstanding data and to make sure that the pipe remains full. However, sending too much data at once can overrun intermediate routers, switches, and end devices. The TCP congestion control manages changes to the window size.

You cannot configure the TCP window size directly. This value is automatically calculated from the product of the maximum bandwidth x RTT x 0.9375 + 4 KB. In SAN-OS 1.3 and later, the RTT can dynamically adjust up to four times the configured value in the FCIP profile according to network conditions. The TCP sender dynamically changes the maximum window size accordingly.

TCP Maximum Bandwidth

The TCP maximum bandwidth is the maximum amount of bandwidth an FCIP link consumes from the point of view of the TCP sender. The maximum bandwidth settings for an FCIP link can be asymmetric. Set the TCP maximum bandwidth to the maximum amount of bandwidth you want the FCIP link to consume. Set it no higher than the bandwidth of the slowest link in the FCIP link path. For example, if the FCIP link is mapped over a dedicated DS3 WAN link, set the maximum bandwidth to 45 Mbps.

The TCP maximum bandwidth value is used as the **bandwidth** value in the **bandwidth-delay** product calculation of the TCP MWS.

Observe the following guidelines when selecting a value for TCP maximum bandwidth:

- Set the TCP maximum bandwidth value no higher than the maximum path bandwidth available to the FCIP.
- If deploying FCIP over a shared link with critical traffic, lower the maximum bandwidth to a level that allows the other traffic to coexist with minimal retransmissions. Quality of service (QoS) should be considered in these situations.
- When using the Cisco MDS 9000 software compression, set the maximum bandwidth value as though compression is disabled. The Cisco MDS 9000 uses a dynamic moving average feedback mechanism to adjust the TCP window size according to the compression rate.

TCP Minimum Available Bandwidth

The value should represent the minimum amount of bandwidth in the FCIP path that you expect to be always available. This value determines the aggressiveness of FCIP—a higher value is more aggressive, a lower value is less aggressive. A value that is too high can cause congestion and packet drops for any traffic traversing the shared network links.

Bandwidth allocation strongly favors the FCIP traffic when mixed with conventional TCP traffic, which recovers from drops more slowly. To cause FCIP to behave more fairly, use a lower value for the *min-available-bw* parameter. FCIP starts at a lower rate and increments the send rate every RTT, just like classic TCP slow-start.

The *min-available-bw* parameter provides the necessary control. Even in the presence of drops, the sender tries aggressively to reach the value configured for this parameter. Even if the congestion window is decreased because of drops, it is increased again on every send so that it is not less than the configured minimum bandwidth.

Round Trip Time

RTT is a measure of the latency or delay back and forth over the FCIP tunnel. RTT is typically twice the end-to-end or one-way delay. Note that IP packets can take different paths each way through a network, so the unidirectional delays are not always equal.

You must configure an appropriate estimate for the RTT. An underconfigured RTT may cripple FCIP throughput. In SAN-OS 1.3 and later releases, the RTT is automatically adjusted from its initial setting.

The configured RTT value is used to calculate the initial TCP MWS and the appropriate TCP window scaling factor. Based on the dynamic RTT, the actual TCP MWS is adjusted dynamically within the bounds of the initially chosen TCP scaling factor.

Customer Premises Equipment (CPE)—Cisco 9216/9216i and Cisco 7200

Cisco 9216

The Cisco MDS 9216 Multilayer Fabric Switch brings new functionality and investment protection to the fabric switch market. Sharing a consistent architecture with the Cisco MDS 9500 Series, the Cisco MDS 9216 combines multilayer intelligence with a modular chassis, making it the most intelligent and flexible fabric switch in the industry. Starting with 16 2/1-Gbps auto-sensing Fibre Channel ports, the MDS 9216 expansion slot allows for the addition of any Cisco MDS 9000 Family modules for up to 48 total ports.

The modular design of the Cisco MDS 9216 allows it to support any Cisco MDS 9000 Family switching or storage services module. The available modules include the following:

- 16-port and 32-port 2/1-Gbps auto-sensing Fibre Channel switching modules
- IP Storage Services Module supporting iSCSI and FCIP over eight ports of 1-Gbps Ethernet
- Advanced Services Module providing in-band storage virtualization and switching services
- Caching Services Module supporting fabric-based storage virtualization with integrated data caching

Optionally configurable, these modules give the Cisco MDS 9216 Multilayer Fabric Switch unparalleled functionality and versatility.

IPS Module

The IP Storage (IPS) Services Module is the heart of FCIP service. The flexible IP storage services (4-port and 8-port configurations) deliver both FCIP and iSCSI IP storage services. The iSCSI functionality is software configurable on a port-by-port basis.

The IPS module offers the following:

- Simplified business continuance and storage consolidation—Uses widely known IP to cost-effectively connect to more servers and more locations over greater distances than previously possible.

- Simplified management—Provides a unified management environment independent of whether servers use FCIP to connect to the storage network.
- Comprehensive security—Combines the ubiquitous IP security infrastructure with Cisco virtual SANs (VSANs), hardware-based zoning, and hardware-based access control lists (ACLs) to provide robust security.

Cisco FCIP

Cisco FCIP offers the following functions:

- Simplifies data protection and business continuance strategies by enabling backup, remote replication, and disaster recovery over WAN distances using open-standard FCIP tunneling
- Improves utilization of WAN resources for backup and replication by tunneling up to three virtual ISLs on a single Gigabit Ethernet port
- Reduces SAN complexity by eliminating the need to deploy and manage a separate remote connectivity platform
- Preserves the Cisco MDS 9000 Family enhanced capabilities including VSANs, advanced traffic management, and security across remote connections

Cisco MDS 9216i

The Cisco MDS 9216i Multilayer Fabric Switch is designed for building mission-critical enterprise storage networks that take advantage of the cost-effectiveness and ubiquity of IP for more robust business continuance services, leveraging both Fibre Channel and IP in a single module. The Cisco MDS 9216i brings new capability to the fabric switch market. Sharing a consistent architecture with the Cisco MDS 9500 Series, the Cisco MDS 9216i integrates both Fibre Channel and IP storage services in a single system to allow maximum flexibility in user configurations.

With 14 2-Gbps Fibre Channel ports, two Gigabit Ethernet IP storage services ports, and a modular expansion slot, the Cisco MDS 9216i is ideally suited for enterprise storage networks that require high performance SAN extension or cost-effective IP storage connectivity. This level of integration gives Cisco MDS 9216i users the benefits of a multiprotocol system without sacrificing Fibre Channel port density. The expansion slot on the Cisco MDS 9216i allows for the addition of any Cisco MDS 9000 Family module, so users can add additional Fibre Channel ports and additional IP ports. Alternatively, the expansion slot may be used for a variety of Cisco MDS 9000 Family services modules, thereby providing an unparalleled level of storage services in a single, highly available 3-rack unit system.

As the storage network expands further, Cisco MDS 9000 Family modules can be removed from Cisco MDS 9216i switches and migrated into Cisco MDS 9500 Series Multilayer Directors, providing smooth migration, common sparing, and outstanding investment protection.

The modular design of the Cisco MDS 9216i allows it to support current and future Cisco MDS 9000 Family switching or services module.

Currently available modules include the following:

- 16-port and 32-port 2-Gbps Fibre Channel switching modules
- IP Services Module supporting iSCSI and FCIP over both four and eight ports of Gigabit Ethernet
- Multiprotocol Services Module supporting 14 ports of 2-Gbps Fibre Channel and 2 ports of Gigabit Ethernet that provide iSCSI and FCIP storage services
- Advanced Services Module and Caching Services Module supporting integrated network-hosted application services

Multiprotocol Services Module

The Cisco MDS 9000 Family 14/2-port Multiprotocol Services Module delivers the intelligence and advanced features required to make multilayer SANs a reality, by integrating in a single module the functions offered by the Cisco 16-Port Fibre Channel Switching Module and the Cisco IP Storage Services Module. The Cisco MDS 9000 Family 14/2-port Multiprotocol Services Module doubles both the Fibre Channel and port density of the Cisco MDS 9216i when used in the switch expansion slot.

Cisco 7200

The Cisco PA-FC-1G PAM, when used with a Cisco 7200 VXR series router or a Cisco 7401 router, provides a method for extending the performance of a SAN by providing Fibre Channel bridge port (B port) functionality. The PA-FC-1G is based on an FCIP ASIC that provides the Fibre Channel and FCIP functions for the port adapter. The FCIP ASIC provides direct memory access (DMA) for complete packets across the Peripheral Component Interconnect (PCI) bus.

When designing for maximum SAN performance with the PA-FC-1G, consider the effect of the PCI bus structure on the Cisco 7200 series router. The Cisco 7200 series router with the NSE-1 or NPE-400 network processors provides the following two internal PCI buses:

- PCI bus mb1 controls slot 0 (the I/O controller) and PA slots 1, 3, and 5
- PCI bus mb2 controls PA slots 2, 4, and 6.

Each PCI bus has a raw throughput of approximately 1.6 Gbps. This effectively limits the PA-FC-1G to a theoretical limit of approximately 500 Mbps if there are two other port adapters on the same PCI bus. Additionally, when data is fast-switched between port adapters, it must traverse to and from the 7200 system memory over the PCI bus. If the data is transmitted between two port adapters on the same bus, the data must traverse the PCI bus twice, effectively reducing the performance of the data path by three-fourths.

Also consider the PCI bandwidth utilization of the other modules installed in the Cisco 7200 router, including port adaptor modules (PAMs), service adapters (SAs), and the I/O controller. Each module has an assigned bandwidth value that reflects the maximum PCI bus utilization of that module. Each PCI bus (mb1 and mb2) has a maximum recommended bandwidth point value of 600 points per bus. When configuring the Cisco 7200 VXR series router, you must also take these bandwidth points into consideration when placing the modules into the chassis. For more information, see the *Cisco 7200 Series Port Adapter Installation Requirements* documentation at the following URL:

http://www.cisco.com/en/US/products/hw/modules/ps2033/products_configuration_guide_chapter09186a008014cf5c.html

CPE Selection—Choosing between the 9216i and 7200

The Cisco MDS 9216i is used when the line rate is a requirement. The Cisco MDS 9216i also supports VSANs and it is optimized for TCP operations. For typical new installations where line rate is a requirement, then the Cisco MDS 9216i with IPS is the correct option.

Most SMBs may already have invested in the Cisco 7200 to transport their data. In these scenarios, installing a Fibre Channel port adaptor on the Cisco 7200 provides a cost effective solution. This solution with the VAM/VAM2 modules supports compression and encryption. The Cisco 7200-based solution is ideal if the traffic demand on the WAN is less than OC-3 speed.

QoS Requirements in FCIP

Currently, most of the FCIP links are dedicated for pure Fibre Channel traffic. But in most cases if QoS is enabled, most of the SAN applications can be transported across the same traffic engineered infrastructure shared with traditional IP traffic. Because there are several QoS models, make sure the right Differentiated Services Code Point (DSCP) is applied to get the assumed results.

To ensure that one QoS model does not conflict with another application (that is, they can run on the same network), it is important for a common set of QoS priority markings to be used across the entire SP network that is in agreement with enterprise traffic. Following is a recommended set of class of service (COS) and MPLS EXP encodings for a typical application:

- 0—Best effort
- 1—Peak information rate (PIR)
- 2—Committed information rate (CIR)/priority
- 3—Cisco AVVID call control
- 4—FCIP (this is just a recommendation; a full analysis of the network and application is required)
- 5—Voice transport
- 6—Video
- 7—SP network management and control

Note The above DSCP values are just a recommendation; a network administrator may choose another consistent set of numbers if desired.

Cisco MDS supports DSCP values for marking all IP packets in the type of service (TOS) field in the IP header. You can specify different values for control and data:

- The control DSCP value applies to all FCIP frames in the control TCP connection.
- The data DSCP value applies to all FCIP frames in the data connection.

If the FCIP link has only one TCP connection, that data DSCP value is applied to all packets in that link.

The following command configures the control TCP connection and data connection to mark all packets on that DSCP value.

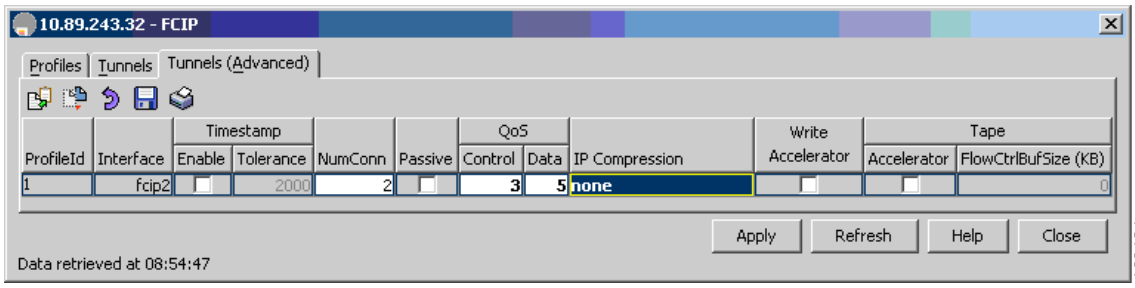
```
switch(config-profile)# tcp qos control 4 data 5
```

The following command reverts the switch to its factory default:

```
switch(config-profile)# no tcp qos control 4 data 5
```

QoS can also be applied using the Fabric Manager/Device Manager GUI. Figure 4-4 shows a QoS value of 3 applied to Control and a value of 5 applied to Data using the management GUI.

Figure 4-4 Using the GUI to Apply QoS



Applications

Disaster recovery and business continuance plans drive the need for solutions that protect critical business information and provide continuous access to important data in case of disaster. Disaster recovery applications are intended to replicate data to a remote backup location. The backup site can be located in the same metro area, such as New York and New Jersey, or at transcontinental distances. The more stringent requirements of business continuance emphasize real-time restoration; when disaster occurs, failover is nearly immediate, providing for faster recovery. Business continuance is put in place to protect business applications at times when downtime is not an option. Common applications for replicating and protecting critical information include synchronous and asynchronous replication and tape backup.

Synchronous Replication

Synchronous replication protects data and applications that have stringent availability requirements. Some applications, such as online trading, must be designed and implemented so that no data is lost in case of a disaster. To achieve this, transactions must be written on both the main and backup sites synchronously to keep the databases consistent. When an application writes data to disk, that data is being replicated to the remote site before a write acknowledgement is sent back to the application. The write I/O is acknowledged on the server only when a block of data has been written on both sites. Therefore, the latency introduced in the network directly affects the application performance.

To limit the impact of replication, storage array vendors are imposing distance limitations for synchronous replication. The distance is typically around 100 kilometers.

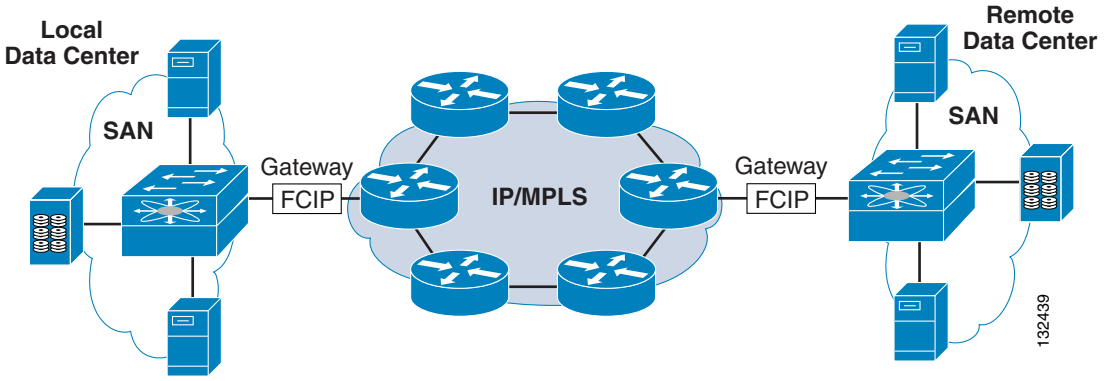
Asynchronous Replication

If a limited amount of business information can be lost in case of disaster, asynchronous replication can be used. Asynchronous replication provides very good protection, but some transactions can be lost in case of disaster. With asynchronous replication, write I/O is completed after being written on the main storage array. The server does not wait until the I/O is completed on the other storage array. There is no distance limitation and typical asynchronous replication applications can span thousands of kilometers or more. See Application Requirements, page 4-29 for more on application requirements.

Service Offerings over FCIP

Figure 4-5 shows a typical service architecture for deploying FCIP over IP/MPLS.

Figure 4-5 FCIP over IP/MPLS Architecture



The FCIP gateway is the key component of the overall architecture.

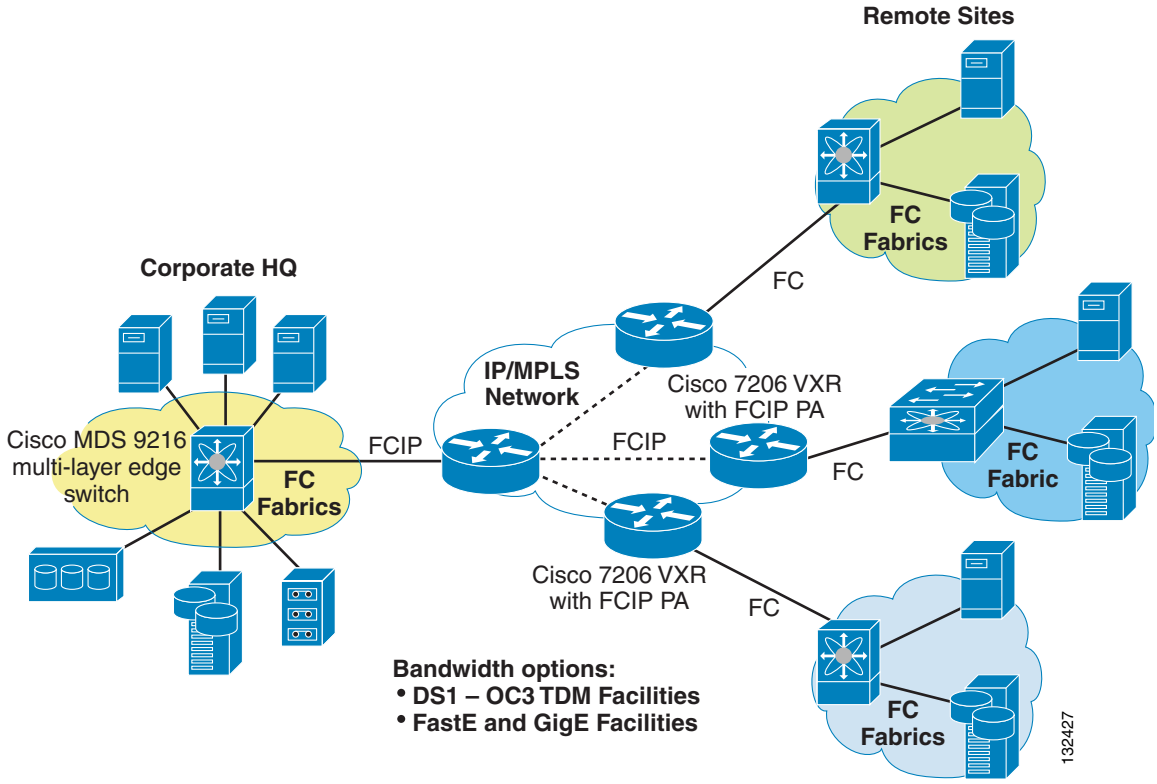
Some typical uses of FCIP to provide SAN extension services are as follows:

- Asynchronous data replication—Enables low recovery point objective (RPO) applications between intelligent storage arrays using proprietary replication software. Network latency does not affect application performance the way it does with synchronous replication. You may need to tune the replication software or upper-layer protocol to ensure optimum use of the FCIP link.
- Remote tape vaulting—Enables remote backup for disaster recovery using tape or disk. Tape applications typically allow a single outstanding I/O operation, which limits throughput on long distance links. Write Acceleration and optionally compression techniques can help to optimize throughput in these situations.
- Host initiator to remote pooled storage—Enables access to FC-attached pooled storage arrays in another site or data center.

Service Offering Scenario A—Disaster Recovery

A customer wants to use FCIP to implement disaster recovery solutions for their remote sites. Corporate HQ is used as a primary site and data is replicated across the SP IP/MPLS network for business continuance and disaster recovery. The same setup can be used to implement backup and restore applications. Figure 4-6 shows a typical hub-and-spoke setup where customer SAN traffic can be transported over the SP IP/MPLS network.

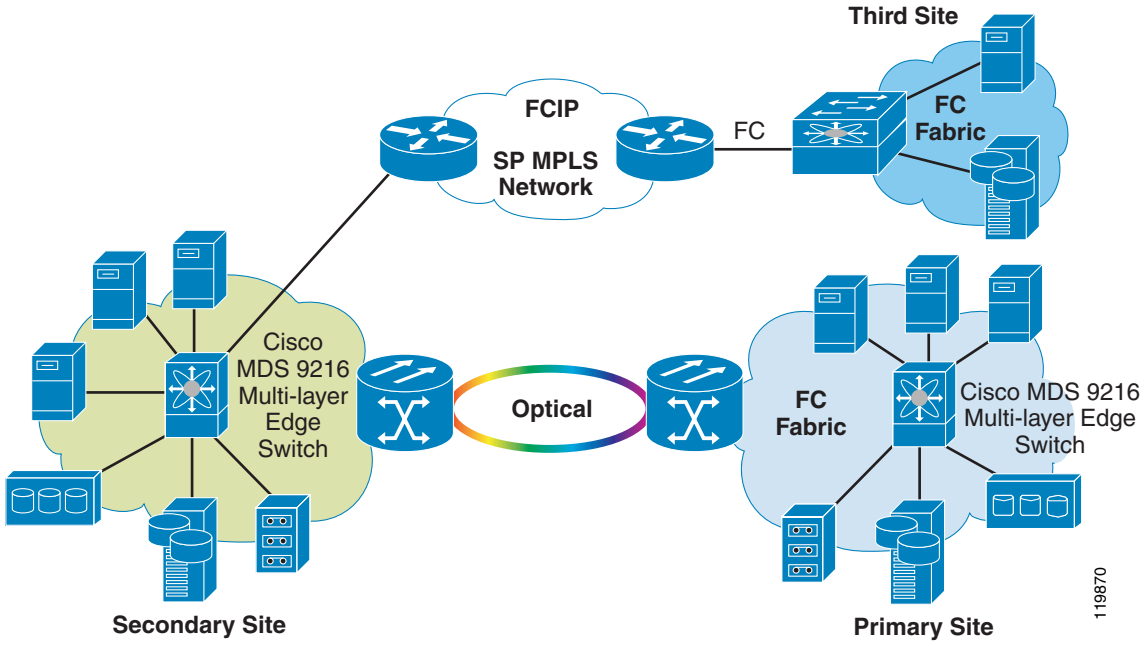
Figure 4-6 FCIP over SP IP/MPLS Core for Disaster Recovery Solutions



Service Offering Scenario B—Connecting Multiple Sites

In certain cases, customers prefer to have primary and secondary sites connected by optical networks such as DWDM or SONET/SDH for high density and reliability. However, because of disaster recovery requirements, corporations might need a third site to protect all their business needs in case of a disaster. FCIP is preferred to connect the secondary site to the third site as shown in Figure 4-7.

Figure 4-7 FCIP Connectivity between Second Site and Third Site



Service Offering Scenario C—Host-based Mirroring

IP/MPLS networks can be used to implement host-based mirroring based on iSCSI. A typical network setup is shown in Figure 4-8.

Figure 4-8 Host-based Mirroring using iSCSI

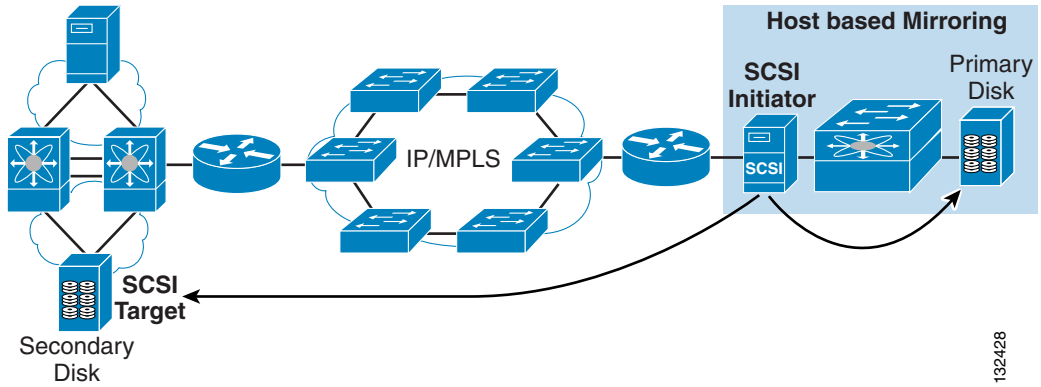


Table 4-1summarizes the possible service offerings an SP can provide to its customers:

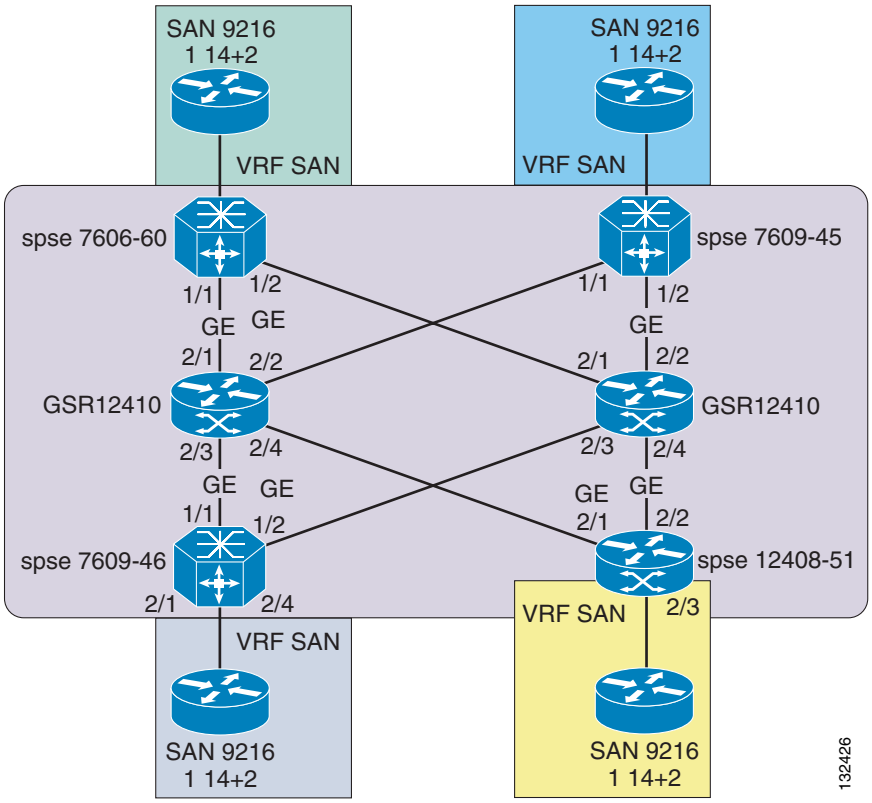
Table 4-1 Possible Service Provider Offerings

Storage Service	Target Customers	Storage Platform	Protocol	Transport Options	CPE
• Synchronous • Data replication • (Real-time ext. distance data mirroring)	• Require no data loss • High volume/rev. impact • Finance/banking/brokerage • Telecom/federal defense • Disaster recovery providers	• CLARiiON • Symmetrix • Hitachi • And so on	• Ethernet • Fibre Channel • FICON	• DWDM • SONET	• DWDM • ONS 15530 • ONS 15540 • ONS 15454
• Asynchronous • Data replication • (Near real-time ext. distance mirroring)	• Larger market • Healthcare • Life Sci/Biomedical • Engineering	• Symmetrix • CLARiiON	• Ethernet • Fibre Channel • FICON	• SONET • FCIP • WDM	• ONS 15454 • MDS 9xxx • 7200 VXR
• High speed • Remote database backup	• Retail • Service Organizations • Airlines	• Symmetrix • CLARiiON • /UltraNet	• Ethernet • Fibre Channel • FICON	• SONET • T3 • IP	• 15310 • MDS 9xxx, 7200VXR
• Low speed • Remote backup	• All Tier 2/Medium-size businesses	• CLARiiON • /UltraNet	• Fibre Channel • FICON	• T1 • IP	• 15310 • MDS 9xxx, 7200VXR
• Email archival	All businesses that use email and instant messaging	Custom	Ethernet	• T1 • SONET/IP	• Router • CSU/DSUs

MPLS VPN Core

MPLS provides an efficient mechanism for supporting VPNs, which offer performance guarantees and security. Using a VPN, customer traffic passes transparently through the Internet in a way that effectively segregates the storage traffic from other traffic on the backbone network. Figure 4-9 shows a sample architecture for an MPLS VPN for storage.

Figure 4-9 *MPLS VPN for Storage Architecture*



Multiple storage customers can be supported on the same MPLS network. Customer 1 cannot see the customer 2 network because there are separate VPN routing/forwarding (VRF) tables for each customer.

Using VRF VPNs

A VRF VPN tunnel is built to provide a secure, managed network between the storage devices. In addition, MPLS VRF VPNs provide distinct advantages for transporting multicast FCIP. VRF VPNs also provide scalability, performance, and stability of the system.

MPLS VPNs use Multiprotocol Border Gateway Protocol MP-BGP) between the provider edge (PE) routers to facilitate the routes between storage VPN areas. MPLS forwarding is used to carry the packets across the backbone. PE routers can use multiple routing and forwarding instances. BGP propagates reachability information for VPN-IPv4 prefixes among PE routers using MP-BGP. This ensures that the routes for a given VPN are learned only by other members of that VPN, enabling members of the VPN to communicate with each other.

When a VPN route learned from a customer edge (CE) router is injected into BGP, a list of VPN route-target extended-community attributes is associated with it. Typically, the list of route-target community values is set from an export list of route targets associated with the VRF from which the route was learned. Based on the routing information stored in the VRF IP routing table and the VRF Cisco Express Forwarding (CEF) table, packets are forwarded to their destination using MPLS. A PE router binds a label to each customer prefix learned from a CE router and includes the label in the network reachability information for the prefix that it advertises to other PE routers.

When a PE router forwards a packet received from a CE router across the provider network, it labels the packet with the label learned from the destination PE router. When the destination PE router receives the labeled packet, it pops the label and uses it to direct the packet to the correct CE router. Label forwarding across the provider backbone is based on either dynamic label switching or traffic engineered paths.

A VRF contains the routing information that defines the customer VPN site that is attached to a PE router. It consists of the following elements:

- An IP routing table
- A derived CEF table
- A set of interfaces that use the forwarding table
- A set of rules and routing protocols that determine what goes into the forwarding table

Testing Scenarios and Results

Test Objectives

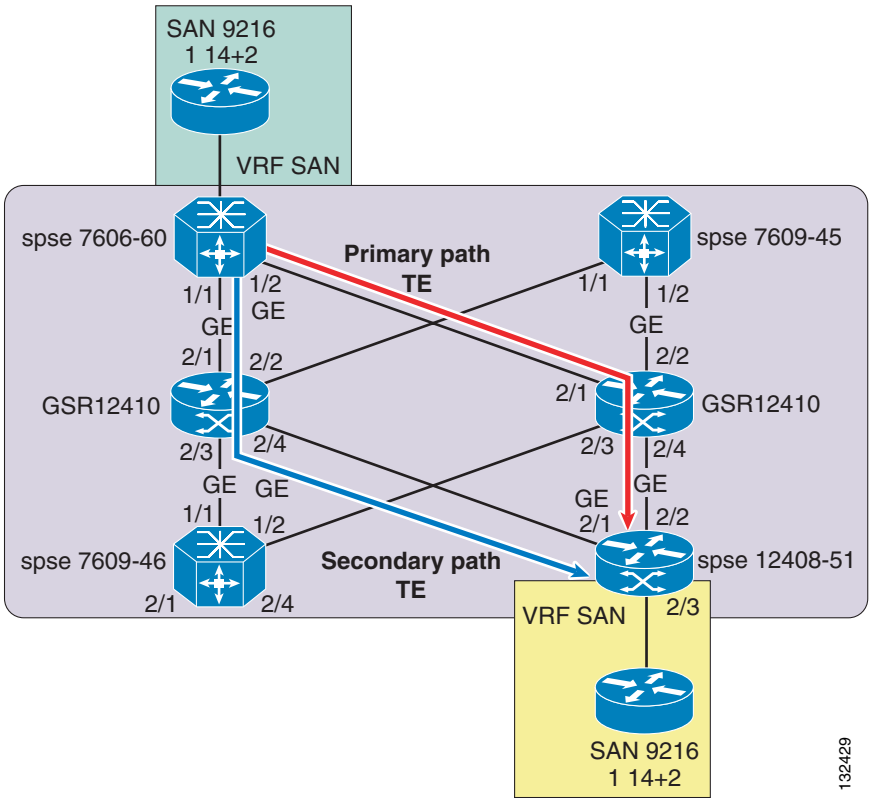
This section describes the testing performed to simulate an IP/MPLS network and to transport FCIP traffic across the simulated SP network. The test objectives were as follows:

- Transporting FCIP traffic (from the customer location) using the Cisco MDS 9216i as the FCIP gateway across IP/MPLS
- Verifying throughput across IP/MPLS
- Verifying whether the traffic can be passed without any errors, including VSANs
- Assessing the impact of core failover
- Assessing the impact of maximum transmission unit (MTU) size

Lab Setup and Topology

Figure 4-10 shows the test lab setup, which consists of two Cisco MDS 9216s connected to the PE routers (Cisco 7500 and GSR) running MPLS. The PEs connect to the core GSR boxes running MPLS.

Figure 4-10 Test Lab Setup and Topology



VPN VRF—Specific Configurations

MP BGP Configuration—PE1

The following is a sample configuration for PE1 in Figure 4-10. The MP-BGP between the two PEs carries the CE routing information. The PE router learns the IP prefix from a CE router through a BGP session with the CE router.

```
router bgp 65001
 no synchronization
 bgp log-neighbor-changes
 neighbor 10.200.0.105 remote-as 65001            ⇓Remote PE
 neighbor 10.200.0.105 update-source Loopback0
 no auto-summary
 !
 address-family vpnv4
 neighbor 10.200.0.105 activate
 neighbor 10.200.0.105 send-community extended
 exit-address-family
 !
 address-family ipv4 vrf storage
 redistribute connected          ⇓redistribute the CE routes onto the storage VRF.
 no auto-summary
 no synchronization
 exit-address-family
!
```

Gigabit Ethernet Interface Configuration—PE1

The following sample configuration is for the Gigabit Ethernet interface for PE1 in Figure 4-10.

```
interface GigabitEthernet0/0/0
 ip vrf forwarding storage
 ip address 11.11.11.2 255.255.255.0
 no ip directed-broadcast
 load-interval 30
 negotiation auto
```

VRF Configuration—PE1

The following are the VRF definitions on the PE1(7500-105) router:

```
ip vrf storage
 rd 105:106
 route-target export 105:106
 route-target import 105:106
!
```

MP BGP Configuration—PE2

MP-BGP between the two PEs carries the CE routing information. The PE router learns an IP prefix from a CE router through a BGP session with the CE router.

```
router bgp 65001
 no synchronization
 bgp log-neighbor-changes
 neighbor 10.200.0.106 remote-as 65001          ⇓Remote PE
 neighbor 10.200.0.106 update-source Loopback0
 no auto-summary
 !
 address-family vpnv4
 neighbor 10.200.0.106 activate
 neighbor 10.200.0.106 send-community extended
 exit-address-family
 !
 address-family ipv4 vrf storage
 redistribute connected     ⇓redistribute the CE routes onto the storage VRF.
 no auto-summary
 no synchronization
 exit-address-family
!
```

Gigabit Ethernet Interface Configuration—PE2

The following sample configuration is for the Gigabit Ethernet interface of PE2.

```
!
interface GigabitEthernet0/0/0
 ip vrf forwarding storage
 ip address 12.12.12.2 255.255.255.0
 no ip directed-broadcast
 load-interval 30
 negotiation auto
!
```

VRF Configuration—PE2

The following are the VRF definitions on the PE2(7500-106) router:

```
ip vrf storage
 rd 105:106
 route-target export 105:106
 route-target import 105:106
```

This test assumes the Cisco MDS 9216/9216i as the CPE. MPLS VPN VRFs allow the SPs to leverage a common backbone to offer shared transport services. To facilitate these services, the provider gets the added security mechanisms of VRFs. The VPN VRFs provide an address space separation; therefore, the use of VRFs on the PE devices and MP-BGP between them achieves address separation not only among the different VPNs but also with the SP core network. Thus Customer 1 cannot see any boxes or interfaces of Customer 2, even though they are on the same transport network and may also share the same PE device.

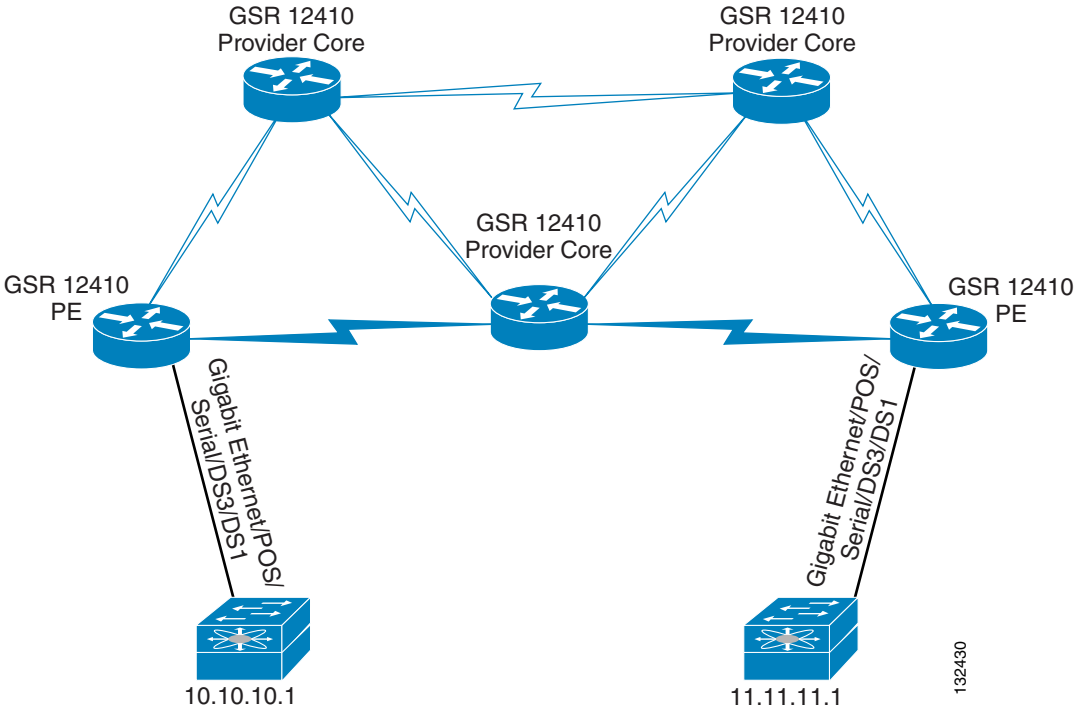
There is no visibility of the core network to the end storage customer, which means that the core network infrastructure including addressing and topology is not visible to the VPN customers. Customer VPN routes that originate from other PE routers across the core network are associated with the BGP next-hop address of the originating PE router. The BGP next-hop address of the PE router is not visible or reachable in the customer address space.

The use of the **traceroute** command can potentially reveal the addresses in the core topology. The core network address can be hidden from view in a VPN by configuring the **no mpls ip propagate-ttl forwarded** command. Therefore, the storage customer can be stopped from seeing the routers in the core network that are carrying the storage traffic.

Scenario 1—MDS 9216i Connection to GSR MPLS Core

In this scenario, GSR is assumed to be the provider (P) and PE device (see Figure 4-11). FCIP traffic is passed across the MPLS network. Tests were performed with different packet sizes. The MPLS networks with proper configurations of MTU size and the TCP parameters on the CPE were able to carry line rate traffic.

Figure 4-11 Cisco MDS 9216i Connection to GSR MPLS Core



Configuring TCP Parameters on CPE (Cisco MDS 9216)

A simple **ping** command from the Cisco MDS 9000 CLI, provides the RTT between the two IP addresses. RTT is specified as part of the following configuration command. It may be specified in either microseconds (-us suffix) or in milliseconds (-ms suffix).

The following command shows RTT set for 20 milliseconds:

```
tcp max-bandwidth-mbps XXXX min-available-bandwidth-mbps xxxx round-trip-time-ms 20
```

Configuring the MTU

The MTU is the maximum payload the Gigabit Ethernet interface will handle. The default MTU for the Gigabit Ethernet interface is 1500, which does not include Ethernet headers and trailers.

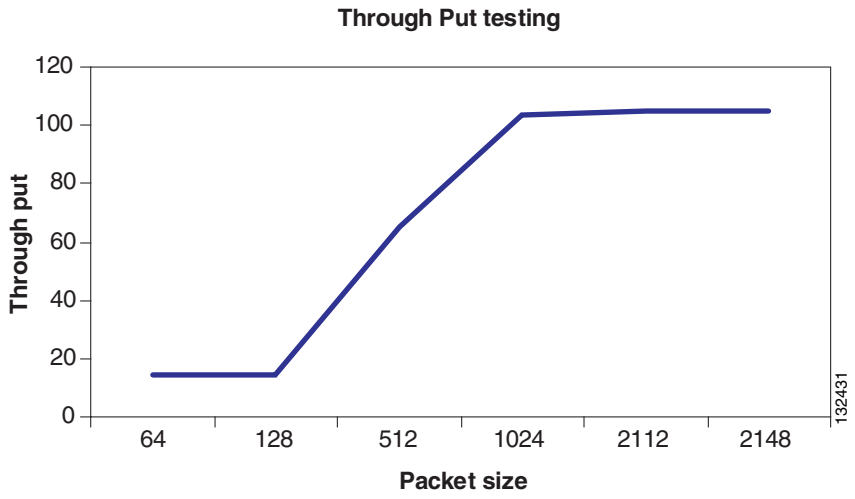
The maximum Fibre Channel frame size including headers and trailers is 2148 bytes. Fibre Channel data frames in typical storage applications have a payload of 2048 bytes plus 36 bytes in headers and trailers, leaving a frame size of 2084 bytes. The Cisco MDS 9000 optionally adds two headers to the Fibre Channel frame. The EISL header is an eight-byte field carrying VSAN tagging information that is only added if the FCIP interface is defined as a TE_Port. If the EISL header is present, it is located immediately after the four-byte start-of-frame (SOF) sequence. There is also an optional header (up to 16 bytes) that is reserved for future use.

With the inclusion of EISL and optional headers, the maximum Fibre Channel frame size is 2172 bytes. An FCIP packet over Ethernet includes 94 to 98 bytes of headers, plus a four-byte Ethernet CRC32 trailer. When carrying the maximum size Fibre Channel frame, the maximum Ethernet frame size is 2274 bytes (2172 + 98 + 4 = 2274).

Where sustained Gigabit throughput is not required (for example, over an OC-12 or slower WAN link), an MTU of 1500 bytes is adequate. Otherwise, use jumbo frames if possible and set the MTU on the Gigabit Ethernet interface to 2300. Also, set the MTU size on the CPE to 2300 bytes. You also need to consider VPN and MPLS headers when configuring MTU size across the path. The MTU size needs to be configured on all routers and switches, including the CPE in the path. The MTU size of 4700 was configured on all PE and P routers to accommodate VPN and MPLS specific headers.

Selective acknowledgement (SACK) is enabled by default in the FCIP profile and should not be turned off. SACK enables the TCP receiver to identify to the sender the contiguous TCP blocks that have been successfully received. The sender can then *selectively* retransmit only the missing blocks. Figure 4-12 shows the result for the throughput testing of the MTU. A full line rate is achieved with packet sizes larger than 1024.

Figure 4-12 Full Line Rate is Achieved with Packet Size Larger than 1024



Scenario 2—Latency Across the GSR MPLS Core

Figure 4-13 shows the average latency with a packet size of 2112.

Figure 4-13 Average Latency—Packet size 2112

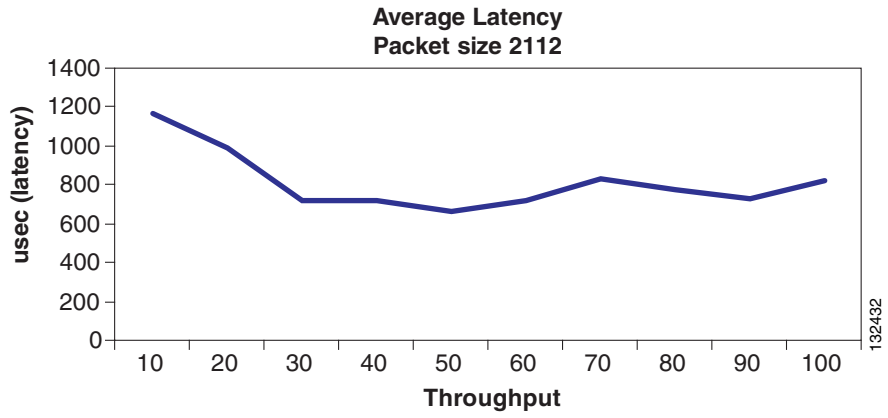
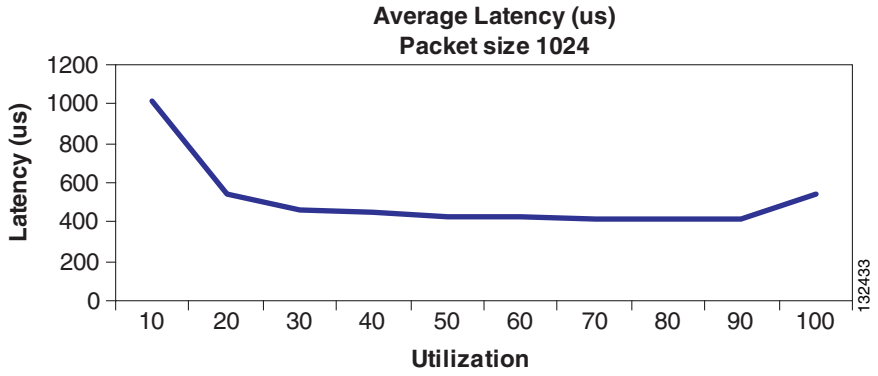


Figure 4-14 shows the average latency with packet size of 1024.

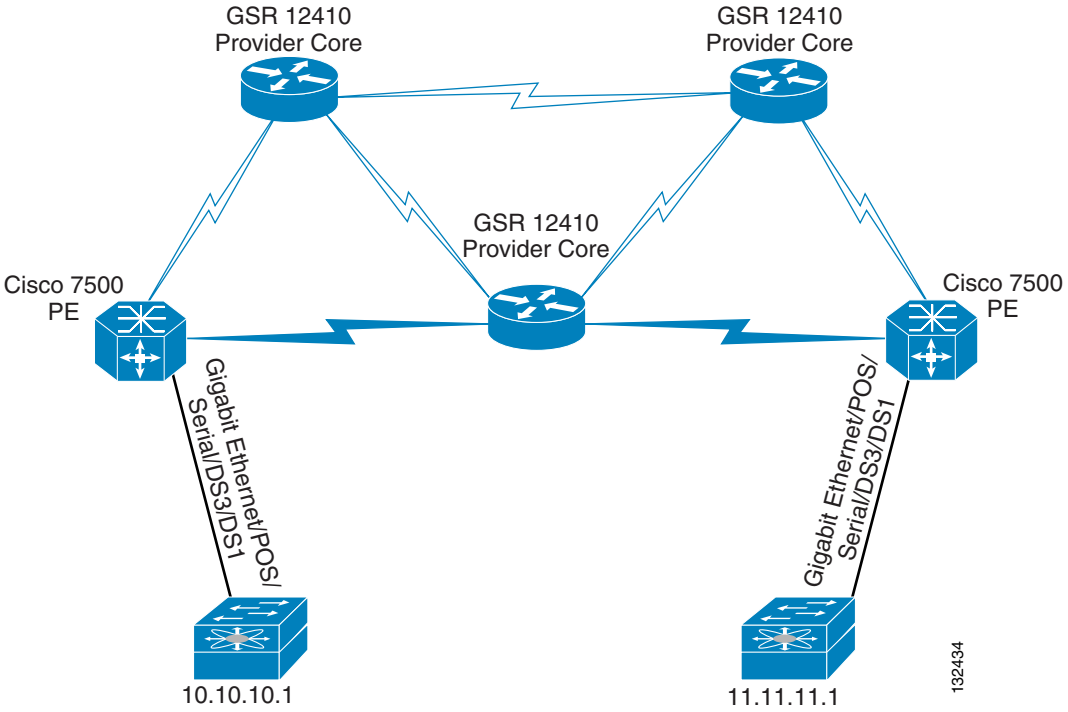
Figure 4-14 Average Latency—Packet size 1024



Scenario 3—Cisco MDS 9216i Connection to Cisco 7500 (PE)/GSR (P)

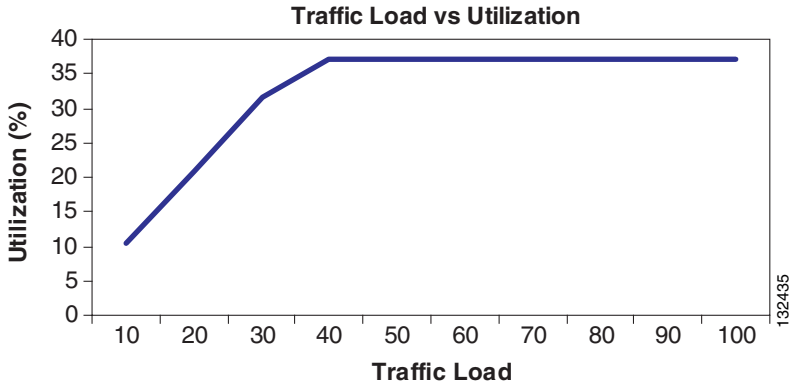
In this scenario, the Cisco 7500 is used as a PE and the FCIP traffic is passed over the GSR (P) routers (see Figure 4-15). Typically, the Cisco 7500 is used as a PE router where traffic demand is minimal. The Cisco 7500 cannot transport line rate for Gigabit Ethernet traffic and is limited to around 35 percent of Gigabit Ethernet bandwidth. This is ideal when the customer traffic is compressed and the requirement does not go beyond the capabilities of the Cisco 7500. As the traffic demand increases, the Cisco 7500 can be replaced by high performing routers like the GSR.

Figure 4-15 Scenario 3—MDS 9216i Connected to 7500 (PE)/GSR (P)



The test results reveal that the maximum traffic that can be transported across the Cisco 7500 as PE is around 35 percent, as shown in Figure 4-16.

Figure 4-16 Traffic Load versus Utilization Test Results



Scenario 4—Impact of Failover in the Core

No convergence testing was done with this FCIP testing. SP backbone convergence depends on the different protection mechanisms deployed in the network at different layers. In general, the following numbers are valid about convergence speed:

- Optical layer—Less than 50 ms
- SONET/SDH—Less than 60 ms
- IP (IGP convergence) variable(s)—With fine IGP tuning, sub-second is achievable for deployment
- MPLS Fast Reroute—Less than 50 ms

In the case of an MPLS backbone, Label Distribution Protocol (LDP) convergence also has to be taken into consideration. The convergence of this protocol depends on the particular mode of operation that is being used: frame mode or cell mode.

Scenario 5—Impact of Core Performance

MPLS provides an efficient mechanism for supporting VPN VRFs. With a VRF, the traffic of a given enterprise or group passes transparently through the Internet in a way that effectively segregates that traffic from other packets on the internet, offering performance guarantees and security.

The total number of routes and VPN supported is dependent on a number of factors, including platform, linecards, and topology. Note the following conditions:

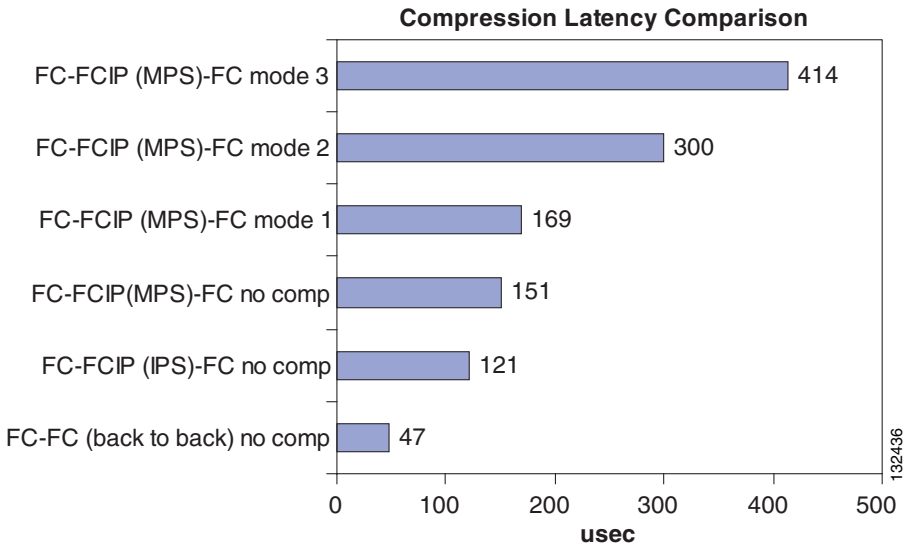
- VRF limits are constrained mainly by CPU
- VPN and global route limits are constrained mainly by available memory

Scenario 6—Impact of Compression on CPE (Cisco 9216i) Performance

You can compress the data stream to reduce WAN link utilization. Some WAN deployments may require compression to obtain adequate throughput. For example, with a 2 to 1 compression ratio, you can obtain 90 Mb/sec of storage data throughput over a 45-Mb/sec DS3 WAN connection. You may also need encryption to protect the confidentiality and integrity of the storage data stream.

Figure 4-17 shows the MDS FCIP and compression latency.

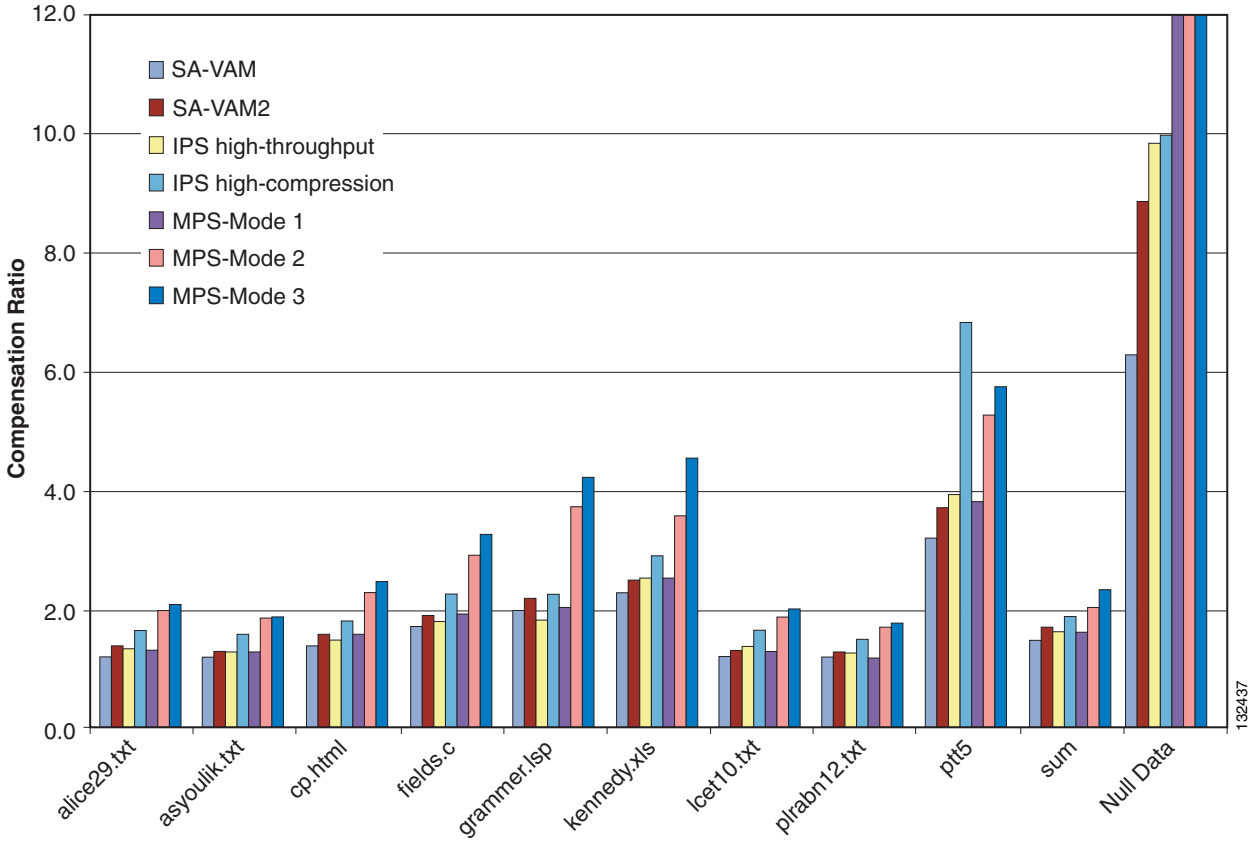
Figure 4-17 MDS FCIP and Compression Latency



Another variable that affects compression performance is the IP MTU size on the FCIP connection. SAN Extension data packets are usually the maximum Fibre Channel size of 2148 bytes. With an IP MTU of 1500 bytes, the IPS must fragment large Fibre Channel frames into two IP frames, which decreases overall performance. Also, the data compression occurs after the Fibre Channel frame is encapsulated into IP, and compression works better with a larger frame to examine. Therefore, the use of IP jumbo frames is recommended for IPS FCIP connections if the network can support it. This increases the efficiency of both Fibre Channel frame encapsulation and data compression.

Figure 4-18 shows the different compression ratio of IPS and MPS as compared to other modules.

Figure 4-18 Compression Ratio Comparisons



Application Requirements

Before designing a WAN, it is important to understand the requirements of each application and how it performs on the network. There are numerous applications available; Table 4-2 generalizes some of the applications. The MAN/WAN should be able to provide the requirements that each application demands to obtain maximum efficiency.

Table 4-2 Application Requirements

Application	Bandwidth	Latency	Async or Sync	Comments
Tape backup	Typically 10 to 15 MB per tape drive Up to 40 MB per tape drive (Super DLT tapes).	< 1–5 ms	Synchronous or asynchronous.	Sensitive to delay. Rely on SCSI protocol for timeouts and error recovery. Note—Once a session is interrupted, the all-backup session is lost.

Table 4-2 Application Requirements

Disk mirroring	Varies depending on storage array. Typically maximum 50 MB per storage array.	< 1–5 ms for synchronous. Asynchronous replication tolerates higher latency (100 ms).	Synchronous or asynchronous.	Synchronous applications are very sensitive to delay. Asynchronous are less sensitive to delay.
File access	OS dependent.			Depends on the OS and application above it.

In addition to those listed in Table 4-2, other requirements include provisioning, error monitoring, and end-to-end management.

Remote Tape-Backup Applications

In an enterprise network, certain sites (remote branches or small offices) can have a small SAN that connects a few servers to the storage arrays. Backing up and restoring these servers over the WAN is a fundamental component of disaster recovery operations. Extending tape backup over a wide area imposes stringent requirements for efficient tape backup and recovery operations. These requirements include no data loss, low latency and jitter, monitoring of the link, and high security.

Slow wide area links can increase backup time and can make it impossible to complete backup within the allocated time period (or "window"). Distance is not a major limitation for backup to tape applications as long as it is possible to predict delay requirements. For backup to tape to be as efficient as possible, Cisco recommends sustaining a certain speed so that a continuous stream of data is sent to tape. Backup performance has been found to be best when the tape can accept a continuous stream. Backup to tape transfer over the WAN is asymmetrical in nature. The asymmetrical nature of tape-backup data transfer creates unique challenges when designing SAN extension networks.

Tape "pipelining" technology helps to extend tape drives thousands of kilometers, thus making remote tape backup an essential component of business continuance and disaster recovery applications. The efficiency is achieved by implementing buffering and error-recovery mechanisms. The concept is similar to spoofing; even though the server and tape controller are separated by a large distance, they behave as if they are co-located. The tape pipelining technique relaxes the design constraints of SAN extension technologies.

A typical solution includes transport over MPLS, which provides all the necessary QoS requirements required by tape backup applications. The Cisco solution provides necessary provisioning, management, bandwidth optimization, and performance parameters that are critical to implement tape backup applications. The Cisco solution can scale as bandwidth requirements increase and still maintain the QoS requirements required to support this application.

Conclusion

IP is becoming a protocol of choice to transport storage traffic across WANs and MANs. The IP/MPLS networks of SPs can be used to transport FCIP and iSCSI efficiently for disaster recovery and business continuance solutions. SPs can leverage their current infrastructure with out much modification to the network elements to transport FCIP and iSCSI. By providing storage services, SPs can increase the utilization of their network while providing value-added services.

Extended Ethernet Segments over the WAN/MAN using EoMPLS

This chapter assists system engineers understand the various options available to extend an Ethernet segment using Ethernet over Multiprotocol Label Switching (EoMPLS) on the Cisco Sup720-3B. Special focus is placed on designs for geographically dispersed clusters.

Introduction

Several technologies can be used to extend Ethernet segments among multiple sites for different profiles of enterprises and networks.

This guide specifically focuses on enterprises that possess an MPLS metropolitan area network (MAN).

For disaster recovery purposes, data centers are hosted in multiple sites that are geographically distant and interconnected using a WAN or a MAN network, which in turn relies on dark fiber, dense wavelength division multiplexing (DWDM), and so on. The entire data center can be deployed using active/backup or active/active load distribution, ensuring that the critical applications such as CRM, ERP, and e-commerce are always available.

High availability (HA) clusters can often be geographically dispersed between various data centers. There are often two data centers; one active, and the other serving as a backup, with at least one member of the extended cluster in each data center. Most HA clusters require one or multiple VLANs to interconnect the nodes hosted in the various data centers. Generally, one VLAN is required for the heartbeat, also known as the private network, and another VLAN is used to carry the virtual IP address (VIP) managed by the application cluster itself, also known as the public network. One VLAN can provide both functions.

For more information, see Chapter 1, "Data Center High Availability Clusters," and Chapter 4, "FCIP over IP/MPLS Core."

EoMPLS can be used to carry Ethernet frames (native or dot1Q) across long distances. This chapter compares the available design options with EoMPLS that allow extending VLANs on top of an existing routed network.

Hardware Requirements

The following hardware is required:

- Aggregation layer—Cisco Catalyst 6000 Series switch or 7600 router with supervisor sup720-3B and Cisco Native IOS (tested with version 12.2(18) SX2)
- Line cards that support configurable MTU at aggregation or in the core
- Access layer—Any Layer 2 switch
- WAN-MAN edge (PE)—EoMPLS-aware devices (7x00, Cat65xx, 7600, 12000)

Note ISR routers do not yet support EoMPLS, but this feature will be available in the future. Layer 2 TPv3 can be an alternative to EoMPLS if ISR is used. Any type of links, such as Gigabit Ethernet, Fast Ethernet, POS, ATM, leased line, Frame Relay, and more can be used.

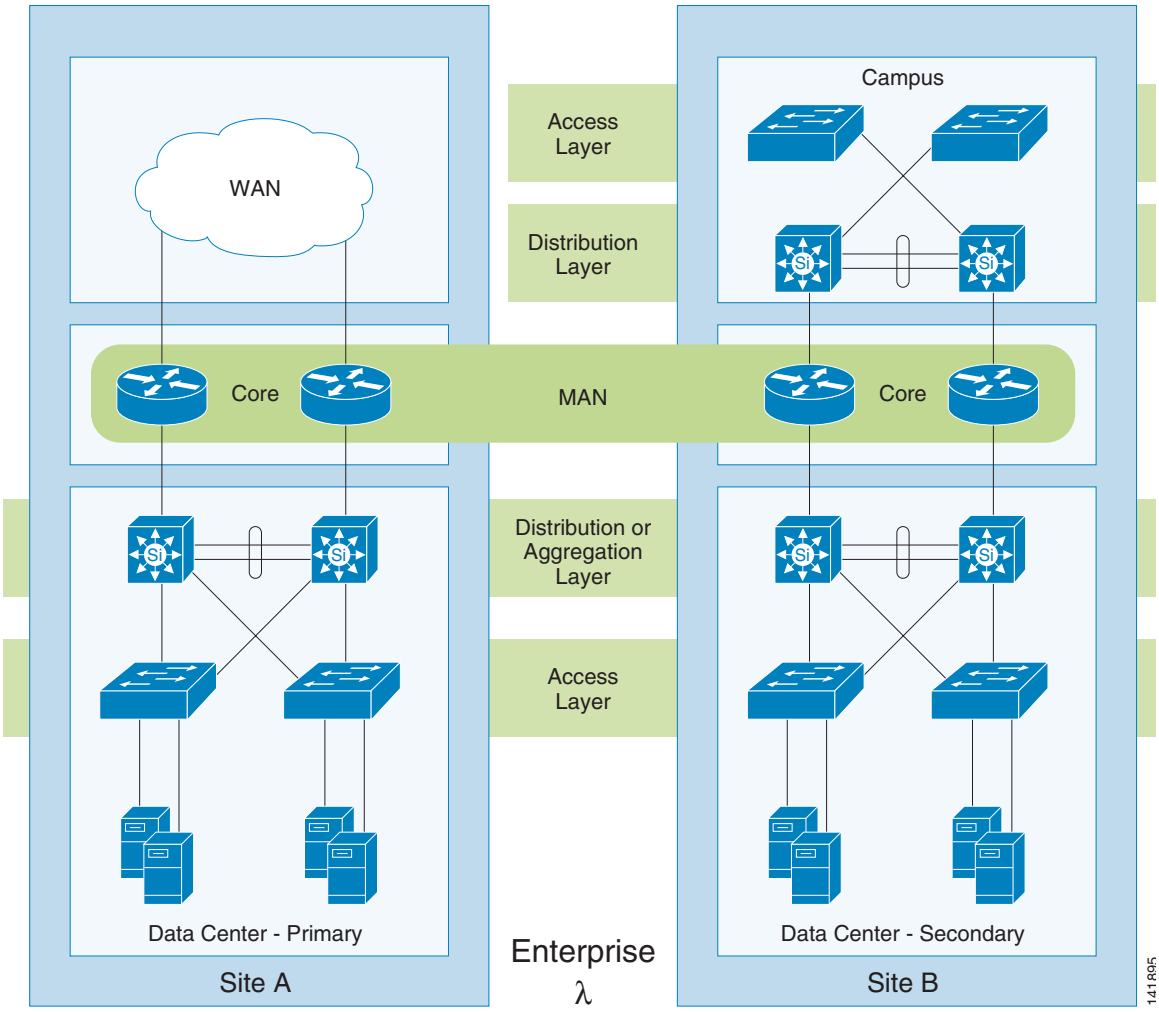
- WAN-MAN core (P)—MPLS router (7x00, 7600, 12000)

Note Sup720 with OSM or SIP cards is not within the scope of testing for this chapter.

Enterprise Infrastructure

Figure 5-1 shows a large enterprise network with two sites.

Figure 5-1 WAN MAN for Enterprise Core



The enterprise network is built around a hierarchical architecture of building blocks. Each building block supports specific functions, such as WAN access, campus, core, and data centers. The Intranet of large enterprises is often extended between multiple buildings or main offices using a MAN. The MAN is the extension of the core that exists at each site. For disaster recovery purposes, cluster members can be hosted by two or more data centers, which often requires that a Layer 2 network be deployed on top of a MAN.

EoMPLS Designs for Data Center Interconnectivity

During the last five years, MPLS has increasingly been the main WAN technology in the service provider arena. Multiple drivers have been key for this success. The two main ones are as follows:

- Switched IP-VPN—MPLS has given service providers the capability to create virtual IP networks over their physical high-speed backbone without the need to use encryption. Currently, approximately one million customer sites worldwide are connected through an MPLS-VPN network.

- Fast Reroute (FRR)—As a switching technology, MPLS offers an option to either logically replace optical fast rerouting techniques using pre-computed protected logical paths, or to replace the optical physical alternate path, ensuring a 50 ms backup at a low cost.

Large enterprises have recently started to adopt MPLS IP-VPN for their own needs. More than one hundred enterprises worldwide are currently deploying MPLS. The main interest for these enterprises is the virtualization capability of MPLS that facilitates the creation of private VPNs in the enterprise.

VLAN bridging can be easily done over an MPLS MAN. MPLS allows a core label to go from an ingress node to an egress node, which can then be used by any edge services.

The EoMPLS ingress node (PE) adds a label into the MPLS stack for any packet coming from a port or VLAN. This label has been previously negotiated with the egress node (PE) to point toward the egress port or VLAN. Thus, bridged packets are transported transparently over a Layer 3 core. This is a key benefit of EoMPLS because no Layer 2 technology, such as MAC address switching or spanning tree, has to be implemented in the MAN. In addition, core links and nodes are Layer 3 protected with a higher stability and faster convergence than any Layer 2 solution. However, packets are not routed, but instead are bridged between sites.

There are three types of EoMPLS. In two of them, attached-circuits are physical interfaces or sub-interfaces and are supported natively in the Sup720-3B. The third one, where the attached circuit is an internal VLAN, requires the core-facing card to be either an OSM or a SIP card.

The three types of EoMPLS are the following:

- Interface Xconnect, also called port-mode Xconnect (cross-connect)

  Interface Xconnect transports any packet getting into the physical port as is, transparently toward the egress associated port. This simulates a cross-connect cable (with an infinite length, as it is transported through the MPLS network). This approach allows flexibility, smooth integration, and full function transparency.

- Sub-interface Xconnect, also called VLAN-edge Xconnect

  Sub-interface Xconnect differs from Interface Xconnect because the ingress port is a trunk (such as dot1Q). The switch removes the 1Q header, extracts the VLAN number, determines the associated sub-interface, and performs the cross-connect with the associated VLAN at the other side of the network. This simulates VLAN-to-VLAN switching (at long distance through the MPLS network). VLAN renumbering is possible, but it adds complexity to the design.

- Internal VLAN Xconnect—This option is not covered in this document. Internal VLAN Xconnect requires OSM or SIP cards.

EoMPLS Termination Options

EoMPLS appears as a tunnel technology that connects ports, or sub-interfaces, or even VLANs located on both sides of a network. In the following diagrams, the EoMPLS tunnels are represented by a virtual link, shown in red, which is transported transparently by an MPLS core.

Note The red dots in the following figures represent the attachment point of the EoMPLS pseudowire.

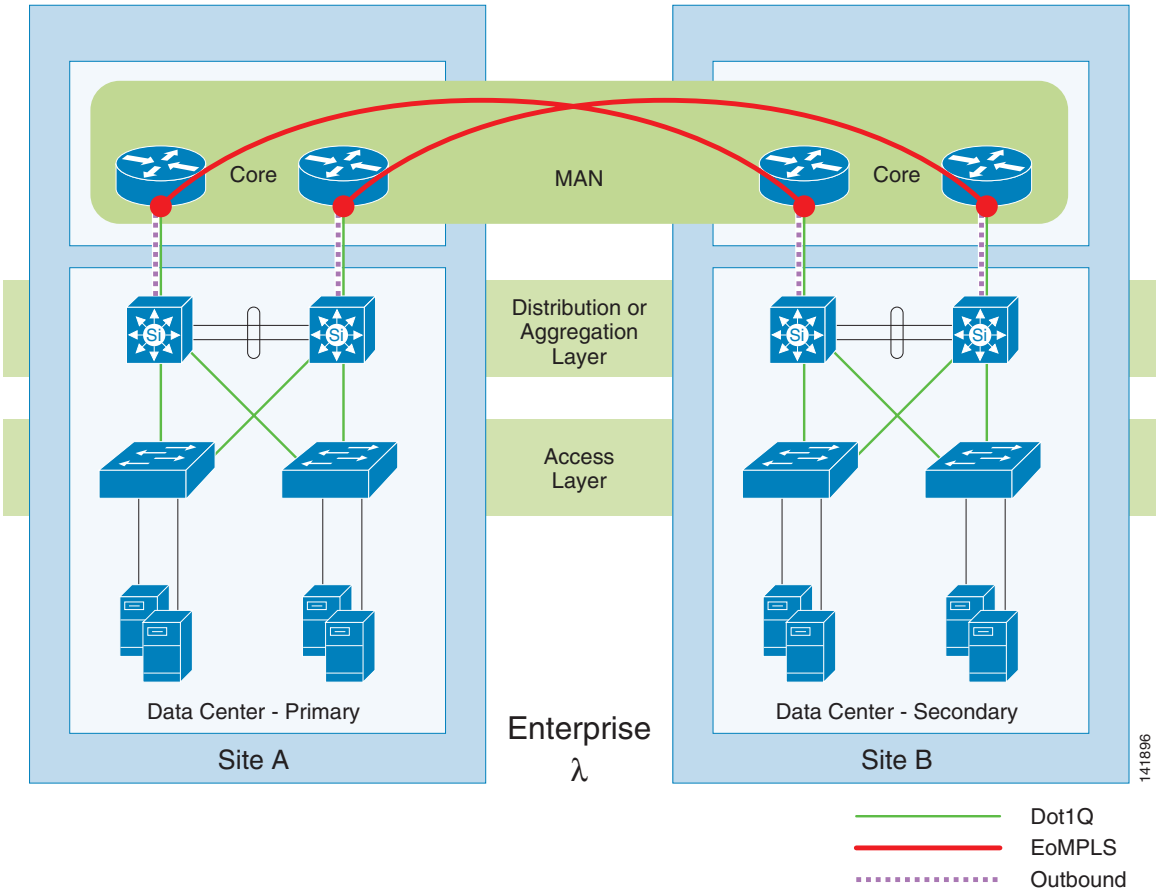
This MPLS core is composed of multiple sites, with a ring or a partially meshed topology. This design guide categorizes the designs based on where the MPLS MAN capability is located: in the MAN, at the data center premise, or within the aggregation layer.

This design guide describes the following four termination options, all based on a port-based Xconnect:

- EoMPLS termination on the MAN routers (see Figure 5-2)

This design relies on an MPLS-enabled core and does not require the data center routers to be configured for MPLS. This design typically relies on a managed Layer 2 service from a service provider. It is quite rare and difficult to use the same fiber from the data center to the POP to provide both Layer 3 VPN and Layer 2 VPN services. This would require deploying another technology, such as VLAN-based Xconnect or VPLS, which are not within the scope of this guide.

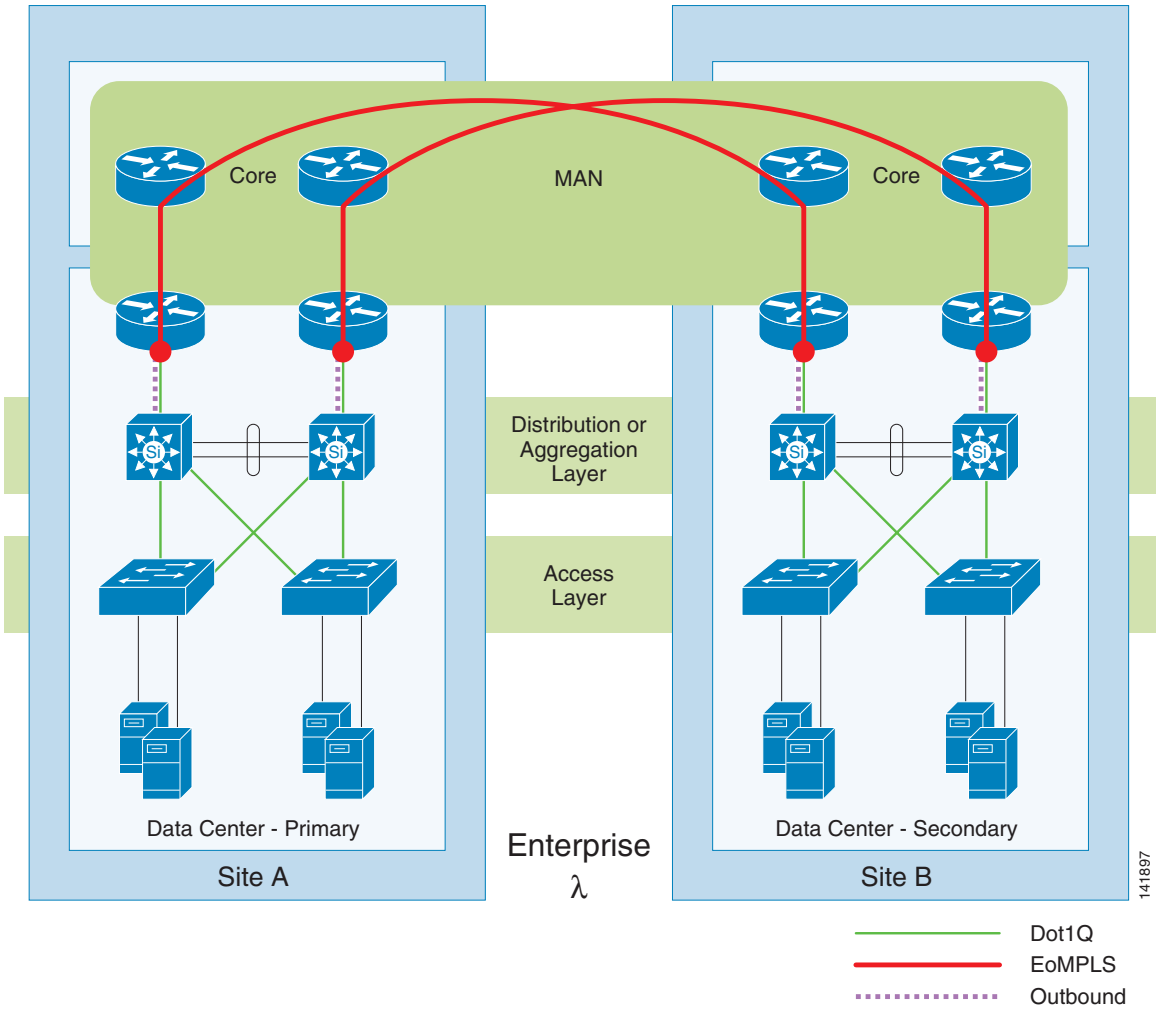
Figure 5-2 EoMPLS Termination on the MAN Routers



- EoMPLS termination on the WAN edge routers (see Figure 5-3)

  This design applies mainly to enterprise self-deployed MANs. With this design, a MAN PE is physically located at the data center boundary and terminates the EoMPLS pseudowire (indicated by the red dot).Two links exist between the aggregation and the PE; one link is used for the Layer 2 traffic between data centers, and one link is for Layer 3 traffic.

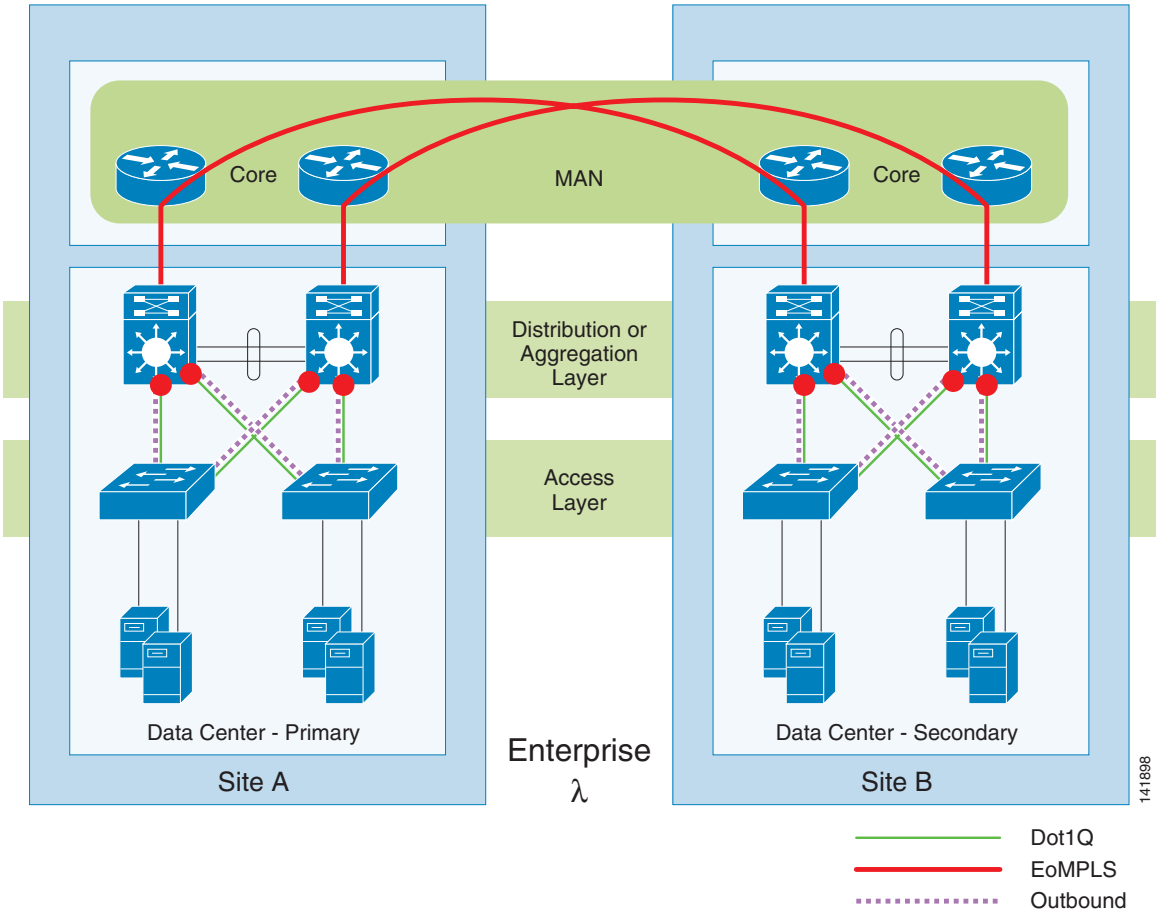
Figure 5-3 EoMPLS Termination on the WAN Edge Routers



- EoMPLS termination in the data center aggregation switches using a dedicated link to the access layer (see Figure 5-4)

  This design leverages the MAN capability of the aggregation layer devices. The aggregation switches in this design provide port-based Xconnect. The access switches need to connect to each aggregation switch with two cables: one cable going to the port cross-connect (which tunnels the VLANs between the two sites), and one cable providing regular routing and switching connectivity to the rest of the network.

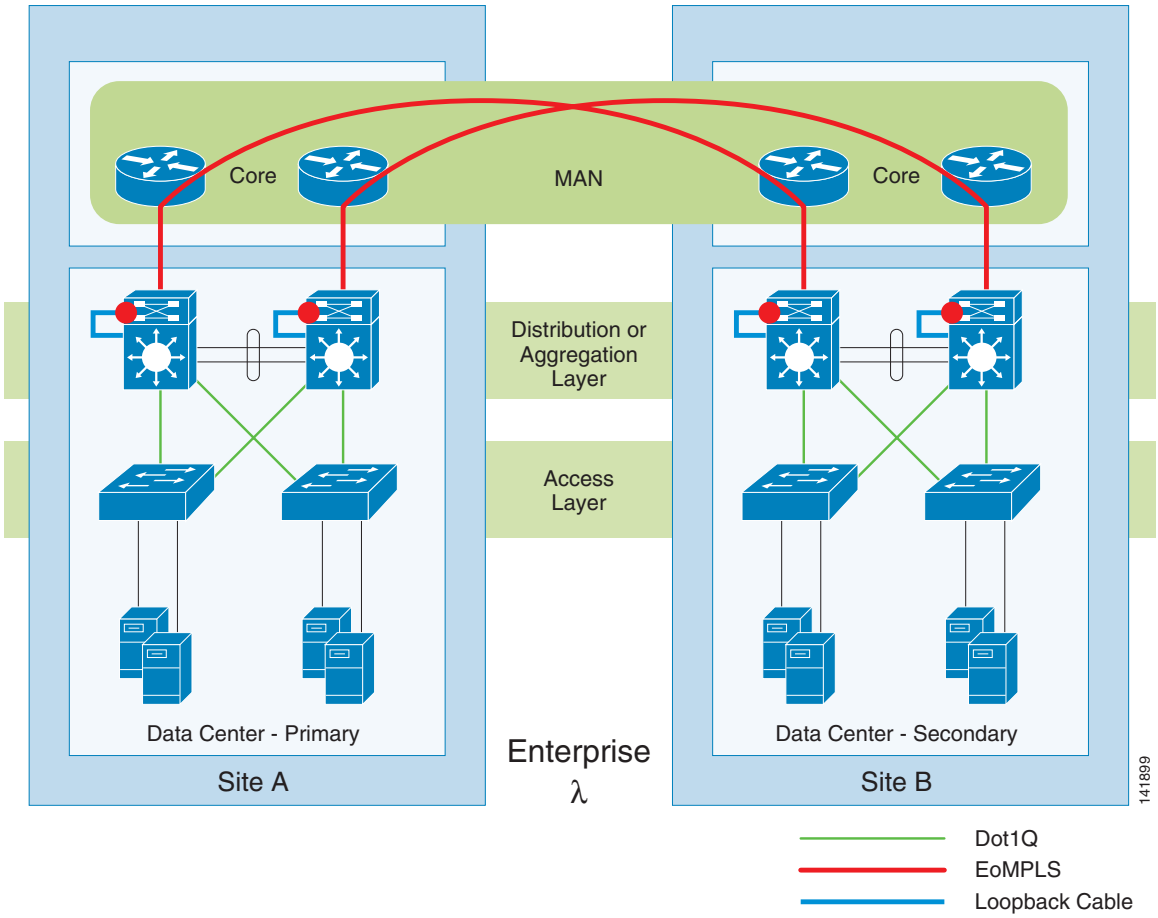
Figure 5-4 EoMPLS Termination at the Data Center Aggregation Using Dedicated Link to the Access Layer



- EoMPLS termination in the data center aggregation switches using a loopback cable (see Figure 5-5)

  To allow transport of the server farm VLAN of the aggregation switches through EoMPLS, a loopback cable is used to re-inject the internal VLAN on a physical port. From an EoMPLS point of view, this approach is very similar to the third option. The main difference is that it does not require multiple physical links from the access switch to the aggregation switch. Figure 5-5 presents a port-based Xconnect and requires a loopback cable at the aggregation layer to carry the VLAN independently.

Figure 5-5 EoMPLS Termination at the Data Center Aggregation using Loopback Cable



Several VLANs from the access layer can be carried through the Layer 2 VPN. Some of these VLANs are limited to data center-to-data center communication, such as the VLAN used for the heartbeat of the cluster. Other VLANs are used for the access to the outside of the data center (outbound), such the VLAN for the virtual IP address.

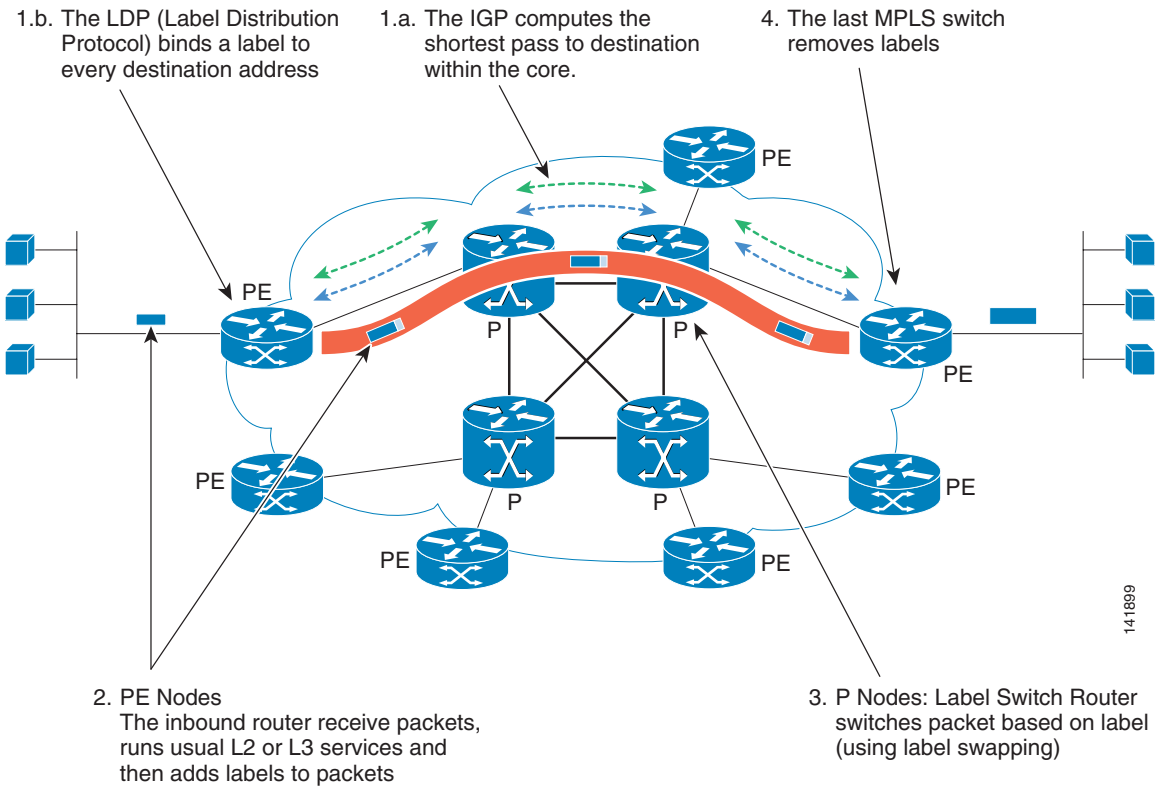
MPLS Technology Overview

The fundamental concepts in MPLS are the following:

- MPLS relies on IP connectivity.

  Figure 5-6 illustrates an MPLS-IP switching service.

Figure 5-6 MPLS-IP Switching Service



Before any labeling of transport and services, MPLS requires that the core be IP-enabled. Therefore, a core IGP must be selected first.

Any IGP is supported by MPLS, MPLS Layer 3VPN, MPLS Layer 2VPN (from static to ISIS, including RIPv2, IGRP, EIGRP, OSPF), but OSPF and ISIS are required to enable the most services, such as traffic-engineering, Fast Reroute, and fast convergence.

- MPLS is a label-based technology.

  At the first stage, labels are negotiated between peers, then subsequent packets are tagged by the ingress device and, at the egress layer, the labels are treated according with the negotiation.

  There are several protocols that are used to negotiate labels, depending on the architecture layer.

  Core label types are the following:

  - LDP (Label Distribution Protocol)
    Used to negotiate best path between to adjacent core nodes.
  - eBGP + labels
    Used to negotiate best path at an Autonomous System interconnect.
  - RSVP
    Used to negotiate deterministic path (with or without bandwidth reservation) along core nodes.

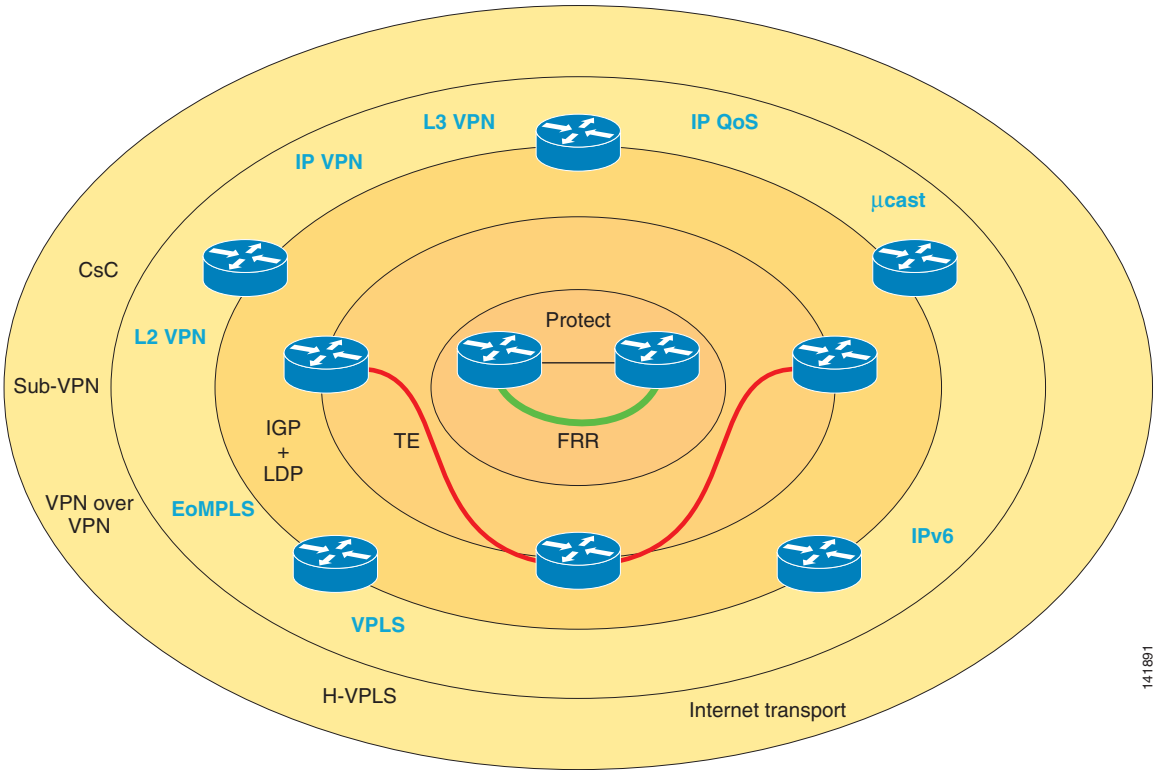
  Edge label types are the following:

  - Directed-LDP
    Used to negotiate Layer 2 virtual-circuit edge to edge.

  - MP-BGP
    Used to negotiate Layer 3 multi-points connection between virtual routing instances between edge nodes.

- MPLS supports the stacking of labels (see Figure 5-7).

Figure 5-7 MPLS Recursivity



MPLS supports the overlay of architecture, each of them being independent of the others, and transported through a stack of labels.

Typical stacking includes the following:

  - A Fast Reroute label
  - A Traffic-Engineering label
  - A Core IGP label
  - A VPN label
  - A Sub-VPN label

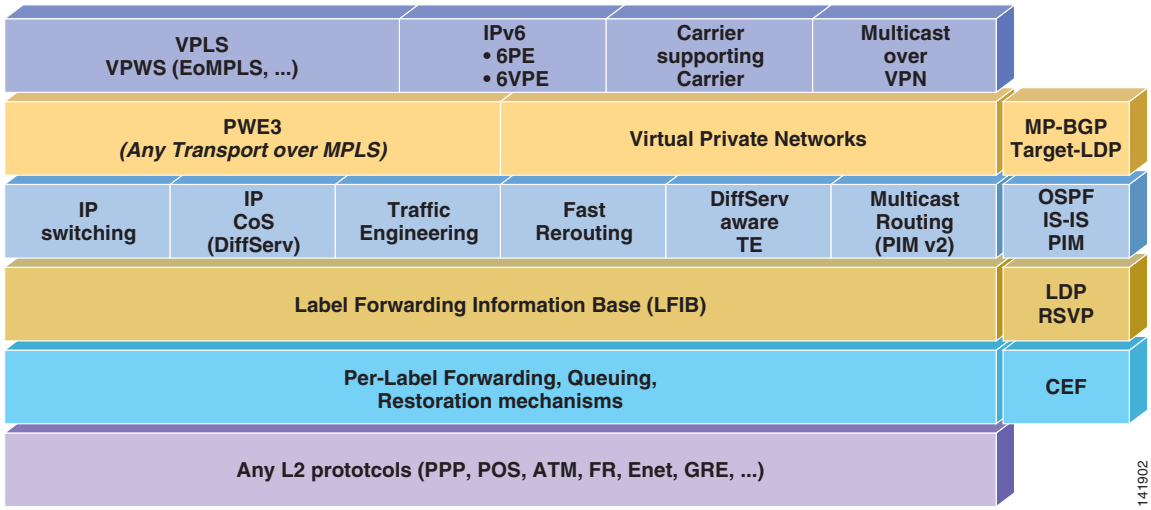
Within an enterprise, such a depth of stack is quite rare, and roughly a Core-IGP label and a VPN label is the most common, while a Fast Reroute label can also be useful.

- MPLS is constructed using three types of nodes:
  - A P node is a Layer 3 node that performs swapping of labels in-between interfaces.
  - A PE node is an edge node (Layer 3 or Layer 2) that imposes labels to plain ingress packet (and removes them at egress).
  - A CE node is any Layer 2 or Layer 3 node attached to a PE, performing IP routing, VLAN bridging, or any other switching technology.

The same device can be a pure P, a pure PE, a pure CE, or it can be some mixture of the three.

- MPLS is a layered technology (see Figure 5-8) consisting of the following:
  - A Data Link layer (Layer 2)—Can be any type of link and, in a MAN, very often an Ethernet transport
  - A core label-swapping layer—IGP or RSVP generated labels
  - Switching services—Core layer induced services
  - Routing virtualization—Layer 3 VPN or Layer 2 VPN virtualization over Layer 3 transport
  - Edge virtualization—Layer 2 or Layer 3 point-to-point/multipoint, IP multicast, IPv6 tunneling, sub-VPN, and so on

Figure 5-8 MPLS Layered Architecture

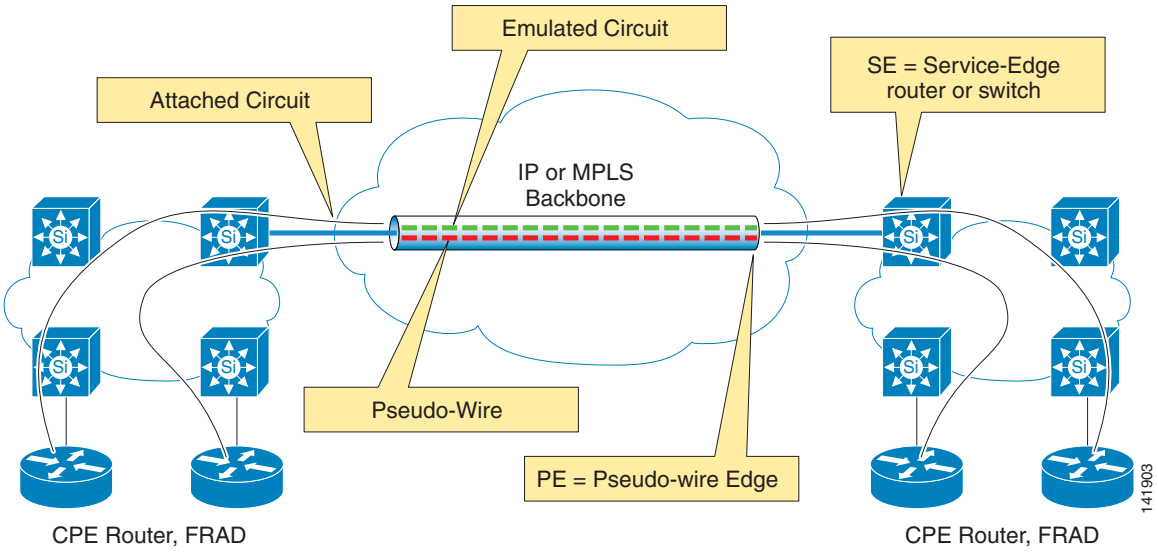


EoMPLS Design and Configuration

EoMPLS Overview

EoMPLS is a virtual circuit technology over an MPLS core (see Figure 5-9).

Figure 5-9 Pseudowire Emulation Edge-to-Edge—PWE3



The elements shown in Figure 5-9 are described as follows:

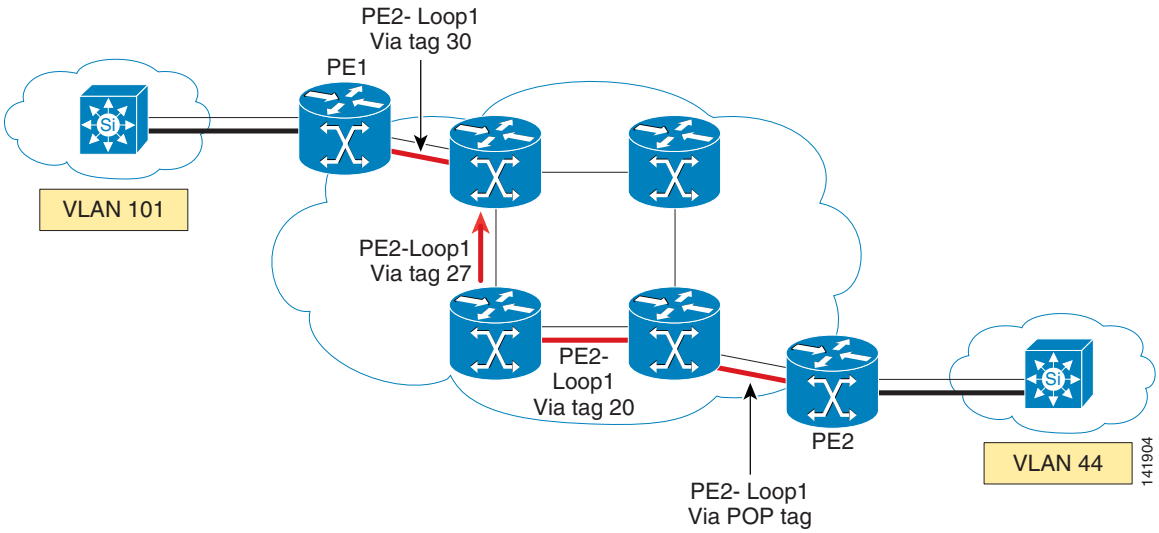
- Attached Circuit

  The purpose of EoMPLS is to transparently remotely interconnect attached circuits at both sides of the network.

- Type of Attachment Circuit with EoMPLS
  - Edge VLAN.
  - Edge Port.
- Emulated Circuit
  - The edge-to-edge transport virtual-circuit; associated with an edge Layer 2 label in the MPLS stack
- Pseudowire
  - The core PE-to-PE transport path (label switched path)
  - In general, established by LDP (sometimes RSVP) all along the core.
- Service edge
  - The Layer 2 device directly attached to the PE.

Figure 5-10 shows the core transport labels.

Figure 5-10 PWE3—Stage 1

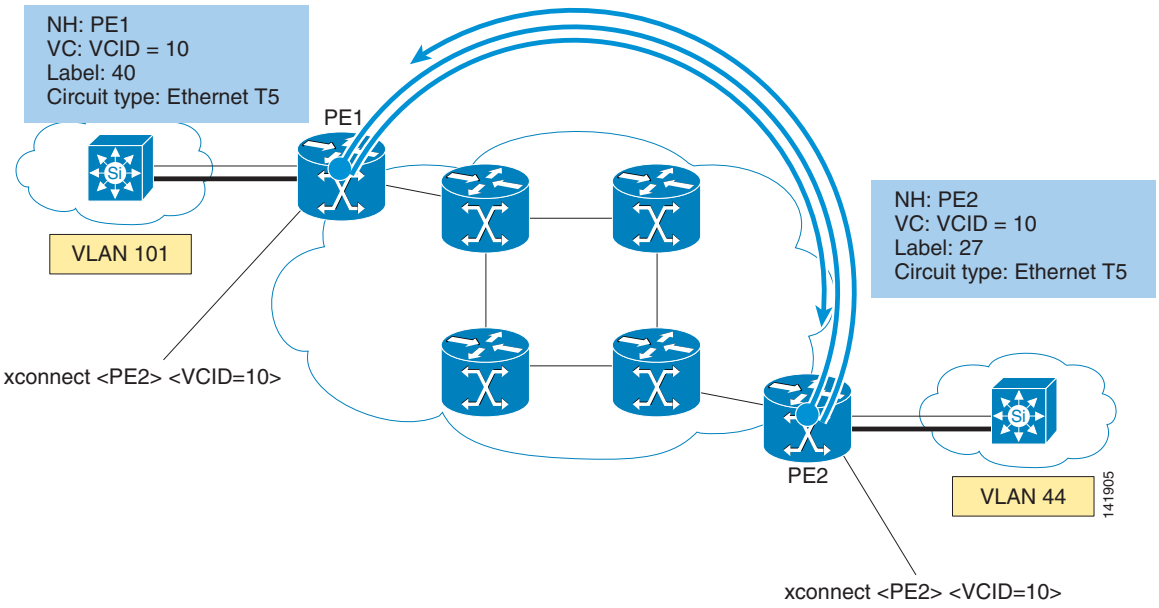


As with any MPLS service, the key element is the MPLS stack. The transport stack layers are used to interconnect PEs (with edge devices providing the upper layer services) and whatever the service is. The service stack layers then provide the labeling of the services themselves (Layer 2VPN, Layer 3VPN, and so on).

Stage 1 is common to any MPLS service, Layer 3 VPN, or Layer 2 VPN. LDP is the common way to distribute these core labels; RSVP can be another one. At this point, edge PEs are able to reach each other through a label switched path. This path can be traced and monitored using MPLS OAM services.

Figure 5-11 shows the labels being distributed through a directed LDP session.

Figure 5-11 PWE3—Stage 2



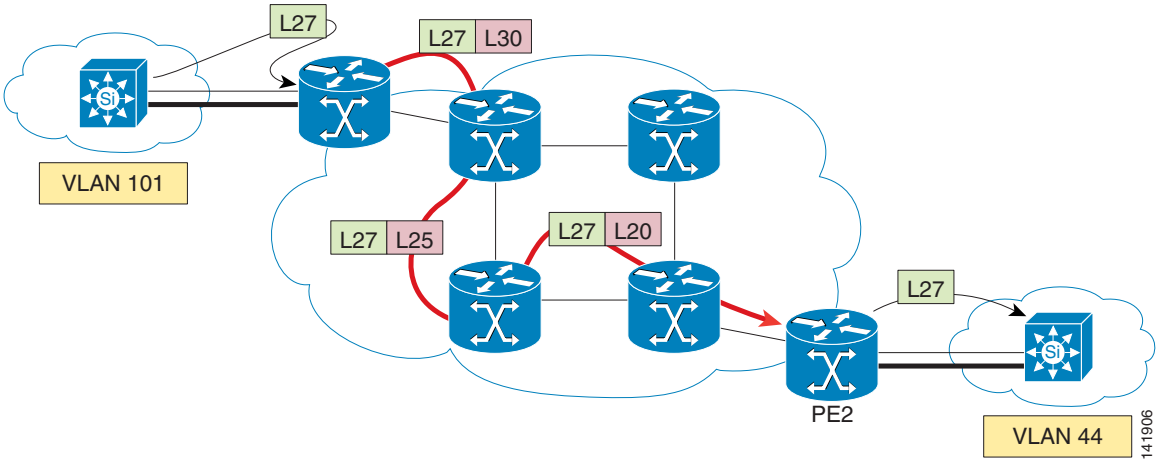
On overlay of the core, PEs are establishing a direct LDP session, completely independent from the core LDP. Each PE is advertising asynchronously to all of the VCs that have been configured.

Towards the specified PE:

- VC-ID number
  - Identifies the Emulated-circuit.
  - The key parameter that allows the other side PE to tie-up the emulated circuit with its local attachment circuit.
  - The VC-ID of each side must be the same.
- Label
  - Every packet at ingress on the attachment circuit of the other side is encapsulated with this label.
- Next-hop
  - Every packet at ingress on the attachment circuit of the other side receives a second label that is the core label leading to the next-hop.

Figure 5-12 shows label forwarding.

Figure 5-12 PWE3—Stage 3



The pseudowire (virtual path) from PE-to-PE is established. Any packet at ingress of the attachment circuit is encapsulated with two labels (the upper one in the stack is the Core label, the second one is the Emulated Circuit label, as shown in Figure 5-13.

The packet is then switched into the core using the top-most label, until it reaches the penultimate core device, (the last P). This one removes the top-most label, which has no further purpose, and the packet, along with only the Emulated Circuit label, is passed to the egress PE. The packet is then pushed toward the attachment circuit using the Emulated Circuit label, which is eventually removed. The removal of the Core label by the last P is an option that is called Penultimate Hop Popping. It is enabled by default.

When two PEs are directly connected, because of the Penultimate Hop Popping, the packet exchange on the direct link is always encapsulated with only the Emulated Circuit label, because the Core label is always empty.

EoMPLS—MTU Computation

This section describes how MTU is configured.

Core MTU

In EoMPLS, the full Ethernet packet is transported, except for the FCS (and the preamble and SFD, which would have no purpose). The maximum PDU size is then 1514.

As it is plain switching, the source and destination MAC address are key to be transported as is, but it is important to note that in Port-mode Xconnect, no bridging of the Ethernet frame is performed by any PE or P. No bridging function at all is performed on the PE ingress or egress ports, and therefore not in the core. That means that none of the MAC addresses of the customers, or the spanning tree, or other bridging features are handled by the PE, which is an important aspect for the stability of the MAN.

Figure 5-13 Ethernet v2—802.3 Encapsulation

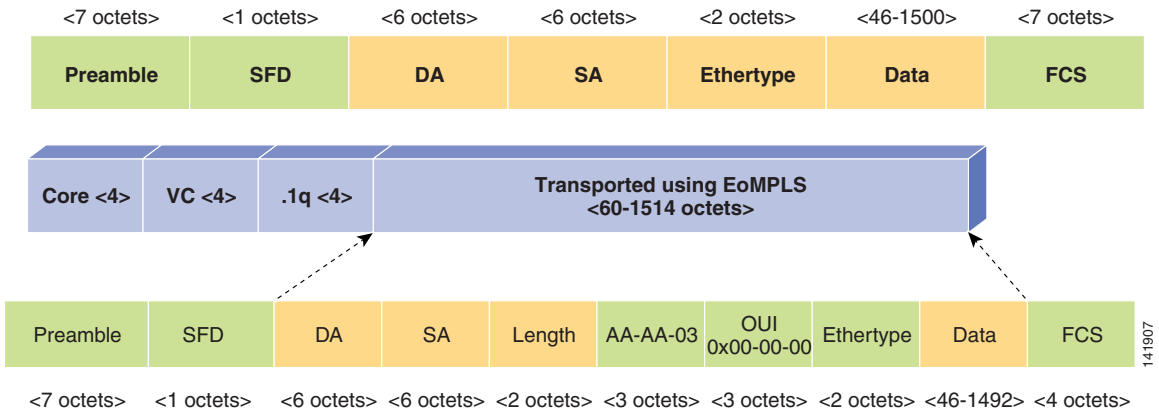
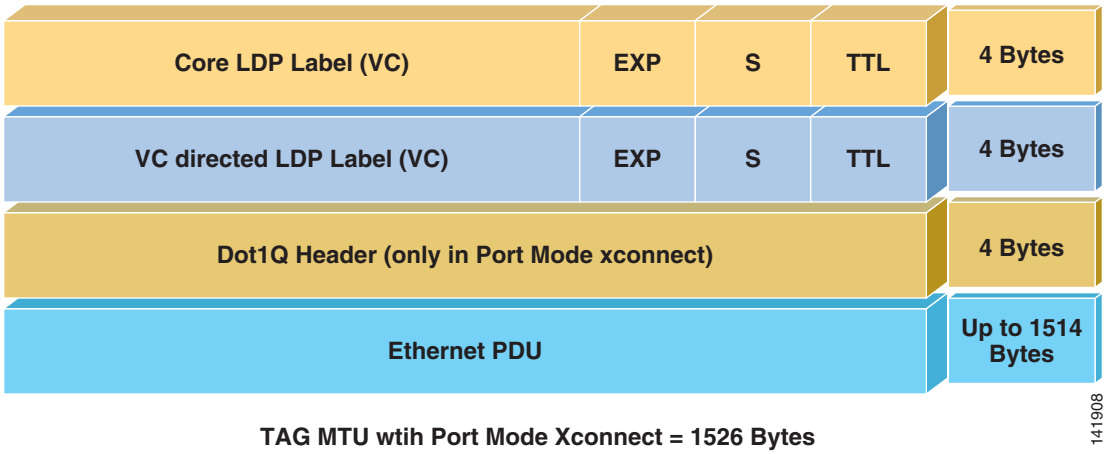


Figure 5-14 shows the MTU settings.

Figure 5-14 TAG MTU with Port-Based Xconnect



TAG MTU wtih Port Mode Xconnect = 1526 Bytes

As in any encapsulation technology in the Ethernet world, MTU is an important aspect to consider. The following is often the most common configuration failure:

Ethernet max PDU = 1514 (as the FCS is not transported).

In VLAN-edge Xconnect, there is no additional header; the Dot1Q header has been removed to determine the right subinterface.

In Port-based Xconnect, any ingress packet is encapsulated without any kind of processing:

- If the Service-edge device is sending plain packets, then the encapsulated PDU max size is 1514.
- If the Service-edge device has defined the link going to the PE as Trunk, then the PDU max size is 1518, as the Dot1Q header is transported as is.

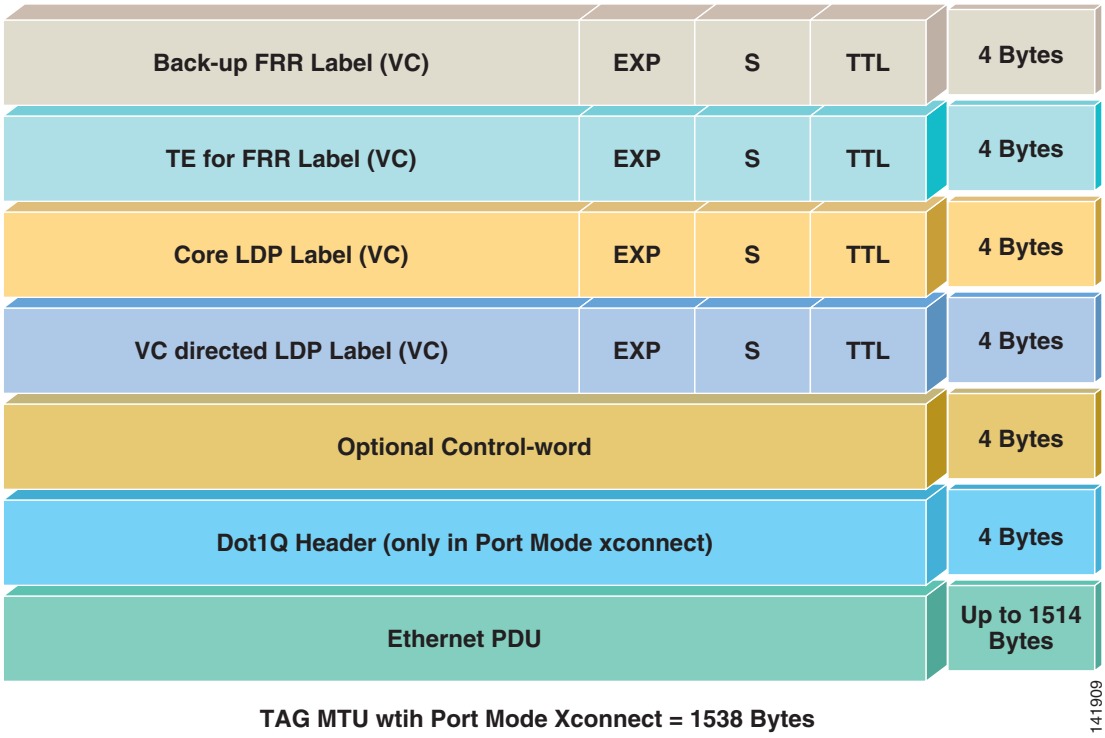
Then, four bytes are added for the Emulated Circuit label (VC-directed LDP label), and 4 more bytes for the core label (core LDP label), if any (remember that in back-to-back, the core label is empty).

The Tag MTU is the only active MTU setting for labeled packets. They are not checked against the Interface MTU. In addition, the Tag MTU includes the label stack.

In MPLS, you do not have to modify the interface MTU, the Emulated Circuit MTU is derived from the MPLS MTU minus the stack size (in this case, 1526 - 8 = 1518).

Figure 5-15 provides recommended MTU settings.

Figure 5-15 MPLS Links Recommended MTU Setting



TAG MTU wtih Port Mode Xconnect = 1538 Bytes

If the Fast Reroute option can be used in the MAN core to reach sub-50 ms backup time, two additional labels must be allowed.

Optionally, an additional control word can be added to the EoMPLS header to allow for features, such as mis-ordering detection. This would increase the size of the encapsulated PDU by 4, but in a common EoMPLS it is unused. A best practice would be to plan for these capabilities, even if they are not used today. The recommended core links MTU is then 1538.

Setting an MTU is always highly dependent on the physical card, or the physical connection used. In Gigabit Ethernet, this constraint is usually light, but in Fast-Ethernet, or with a service provider offering, physical limitation might be encountered.

If the core is transporting only MPLS, which means that no application traffic is being sent using the global routing table, then it is a good practice to increase the physical MTU to the tag value, to ensure transport.

Also, be careful when using a giant or jumbo frame, the previous recommendation is assuming that you are using only a plain Ethernet frame, or the Dot1Q frame has to be transported. If you transport larger Ethernet frames, the core link and tag MTUs must be increased as well (always with 24 additional bytes for EoMPLS).

Edge MTU

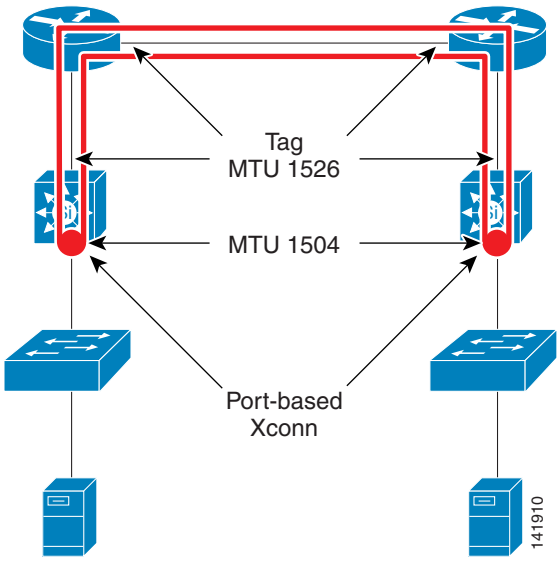
In a Port Cross-Connect configuration, the full traffic coming in is cross-connected toward the egress port and this port is declared as a routed port. If any 802.1Q frames have to be cross-connected, then this means that the port receives MTU frames at 1504 bytes. Therefore, the edge port MTU must be set to 1504. Keep in mind that some physical Fast-Ethernet or Ethernet cards cannot support configurable MTU.

Following are some examples of implementation MTU sizes:

- The physical Port-based Xconn Interface (Gig5/2) requires the 1500 + 4 bytes for tagged frames = 1504 bytes.
- The Uplink physical interface connected to the upstream router (MAN) requires the physical Ethernet max frame to be increased from 1518 by 4 bytes = 1522 bytes. This is required, as the PDU transported by EoMPLS is not 1500 bytes, but the full Ethernet frame (except for FCS), plus the 802.1Q header.
- On every core link, the TAG MTU should be set to 1514 + 4 bytes Dot1Q + 4 Bytes VCID + 4 bytes LDP = 1526 bytes. The reason for 1514 is that the encapsulated frame excludes the FCS (4 bytes).
- The Uplink physical interface of the upstream router (MAN) that is connected to the downstream aggregation switch requires that the TAG MTU be set to 1526 bytes and the physical interface connected to the MAN (remote sites) should have the TAG MTU set to 1526 bytes.

Figure 5-16 illustrates MTU size.

Figure 5-16 MTU Size



EoMPLS Configuration

This section provides some best practices to follow when setting up EoMPLS configurations.

Using Core IGP

To build a Label-Switched-Path, MPLS requires the IP connectivity into the MAN to be set. MPLS can use the current core IGP rather than a specific parallel IGP. If the choice is to use the current MAN WAN IGP to add the MPLS service, then the only real task is to enable MPLS on core links.

Note Ensure that the IGP path in-between edge nodes always cross links with MPLS enable.

Another choice that Cisco has made for the purpose of this best practice example is to use another instance of an IGP for MPLS. Therefore, on every core link that must support MPLS, ISIS has been enabled in addition to MPLS.

One loopback per PE per service is best:

```
interface Loopback99
 ip address 10.99.65.5 255.255.255.255
 ip router isis
 isis circuit-type level-1


router isis
 net 49.0001.0000.6500.5555.00
 is-type level-1
 metric-style wide
 passive-interface Loopback99
 advertise passive-only
```

Set MPLS Globally

```
MPLS label protocol ldp
MPLS ldp router-id Loopback99
```

Limit label advertisement to the only useful address, there is no need to transport other IP addresses in label mode.

For Layer 2VPN (as in Layer 3VPN), the only useful addresses are the PE loopback addresses, which are used for the targeted-LDP exchange.

Use a dedicated IP addressing class for all of the MPLS services loopbacks.

```
Best practice: Use different PE loopbacks for Layer 2 VPN service and Layer 3 VPN service
for a clear separation.

no MPLS ldp advertise-labels
MPLS advertise-tags for 1
access-list 1 permit 10.99.0.0 0.0.255.255
```

Enable MPLS on Core Links

On every link that comprises the MAN (or the WAN, if any), enable ISIS and MPLS transport.

```
interface FastEthernet1/1
 ip address …
 ip router isis
 tag-switching mtu 1526
 tag-switching ip
```

For the MTU setting, see Appendix A, “MTU Considerations.”.

Verify MPLS Connectivity

Without the use of the Traffic-Engineering capability, the MPLS path is the same as the IP path. Therefore, a plain IP-ping or IP-traceroute would still be useful for checking the data path. MPLS-ping and MPLS-traceroute will precisely check the label switched path for LDP.

In addition, MPLS-pseudowire-ping will allow packets to generate directly into the pseudowire to verify connectivity. These new MPLS OAMs are also able to generate packet in sub-second fashion, with a tunable tempo.

```
#ping mpls ipv4 10.99.65.5 255.255.255.255
Sending 5, 100-byte MPLS Echos to 10.99.65.5/32,
     timeout is 2 seconds, send interval is 0 msec:

Codes: '!' - success, 'Q' - request not transmitted,
       '.' - timeout, 'U' - unreachable,
       'R' - downstream router but not target,
       'M' - malformed request

Type escape sequence to abort.
!!!!!

#traceroute mpls ipv4 10.99.65.5 255.255.255.255 ttl 7
Tracing MPLS Label Switched Path to 10.99.65.5/32, timeout is 2 seconds
  0 10.0.0.6 MRU 1526 [Labels: 26 Exp: 0]
R 1 10.0.0.5 MRU 1526 [Labels: 16 Exp: 0] 4 ms
```

```
R 2 192.168.25.2 MRU 1530 [Labels: implicit-null Exp: 0] 1 ms
! 3 10.10.0.6 2 ms
```

Create EoMPLS Pseudowires

As previously stated, there are many ways to create pseudowires. In a data center environment, without any of the 6500/7600 OSM/SIP cards, the standard use is Port mode.

```
Cross-connect at port-level (physical interface level)
interface GigabitEthernet5/2
 mtu 1504
 no ip address
 Xconnect 10.99.65.2 100 encapsulation mpls
```

In Port mode, every packet ingress is transported through the pseudowire without any analysis. Therefore, if the interface might receive a DOT1Q packet, the PDU MTU must be sized accordingly.

Another option, that is not used in these designs, is to perform the cross-connect at the sub-interface level.

Verify EoMPLS Pseudowires

```
#show mpls l2 vc

Local intf     Local circuit        Dest address    VC ID      Status
-------------  -------------------- --------------- ---------- ------
Gi5/2          Ethernet             10.99.65.5      100        UP

#show mpls l2 vc detail
Local interface: Gi5/2 up, line protocol up, Ethernet up
  Destination address: 10.99.65.5, VC ID: 100, VC status: up
    Tunnel label: 26, next hop 10.0.0.5
    Output interface: Fa3/2, imposed label stack {26 16}
  Signaling protocol: LDP, peer 10.99.65.5:0 up
    MPLS VC labels: local 16, remote 16
    MTU: local 1504, remote 1504
  Sequencing: receive disabled, send disabled
  VC statistics:
    packet totals: receive 92471, send 349251
    byte totals:   receive 10962950, send 29963199
    packet drops:  receive 0, send 5

#ping mpls pseudowire 10.99.65.5 100
Sending 5, 100-byte MPLS Echos to 10.99.65.5,
     timeout is 2 seconds, send interval is 0 msec:
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 2/3/5 ms
```

Optimize MPLS Convergence

Optimize to a sub-second or less. In a data center interconnection, the convergence time on a failure becomes very important.

In general, the weakest part of the MAN is the long distance link, but still it is a requirement that the data center be dual-connected to the MAN through two edge switches. One very good approach to increase the high-availability is to rely on the IGP capabilities in terms of convergence.

These days, in a network of a reasonable size a very few hundred of milliseconds is perfectly reachable in term of convergence after a link or node failure has been detected. If the network is more complex, then less than a second is a perfectly reasonable expectation.

```
router isis
 net 49.0001.0000.6500.5555.00
 is-type level-1
 metric-style wide! Allows Traffic-Engineering attribute propagation
 spf-interval 20 100 20! ISIS SPF fast reaction, but backoff protected, see below
 prc-interval 20 100 20! same for IP addresses changes
 lsg-gen-interval 1 1 20! Same for LSP advertisement
 fast-flood 15! Fast flooding for first LSP
```

Backoff Algorithm

The fast reaction of the IGP in terms of path re-computation is controlled by a backoff algorithm that prevents any instability.

In the previous setting, the first SPF is run right after the failure detection (20 ms), but a subsequent computation occurs only after a pace of 100 ms, a third one occurs after an interval of 200 ms, then 400 ms, then 800 ms, up to a max of 20 s.

Return to stability is considered okay after 20 s of perfect stability, and all timers are reset.

Carrier Delay

Fast computation of a new path can occur only after connection failure detection, the next step is to optimize link or node detection failure.

If the failure leads to a physical link down, which is clearly the case of fiber interconnection or even often in an Ethernet over SONET service, then the reaction can be almost instantaneous.

Carrier-delay must be set to 0. Carrier-delay is useful for links that are already protected with another Layer 1 technology, such as SONET or WDM protection.

When Ethernet is used as a point-to-point link, there is no need to waste time in a multi-points negotiation. To protect against any instability versus a fast convergence, dampening will control flapping of the link.

```
interface FastEthernet1/1
 ip address …
  carrier-delay msec 0
 ip router isis

 isis network point-to-point
 dampening
```

When the failure does not lead to a physical detection, the IGP timers will detect the loss of neighbor-ship in one second:

```
interface FastEthernet1/1
 ip address …
 ip router isis
 …
 isis circuit-type level-1
 isis hello-multiplier 10 level-1
 isis hello-interval minimal level-1
```

Following, is an example of convergence testing on link failure. To measure the convergence time, use a **ping mpls pseudowire** command with a frequency of 100 ms:

```
#ping mpls pseudowire 10.99.65.5 100 interval 100 repeat 600 size 1500 timeout 1
Sending 600, 1500-byte MPLS Echos to 10.99.65.5,
     timeout is 1 seconds, send interval is 100 msec:

Codes: '!' - success, '.' - timeout, 'U' - unreachable,

!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
!!!!!!!!!!!!!!.!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
!!!!!!!!!!!!!!!UUUUUUU...!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
Success rate is 98 percent (589/600), round-trip min/avg/max = 1/1/20ms
```

When the link goes down physically, the IGP convergence time is less than a few 200 ms (one pseudowire ping is missing), but the link down detection can be longer, depending on the failure. The worst case is IGP hello detection + path re-computation, which should be below 1,5s. if no BFD is used.

It is important to notice that when the link came back up, the connectivity went down for roughly one second. This strange behavior is caused by a feature that is not yet supported on the Catalyst 6500 or Cisco 7600: MPLS LDP-IGP Synchronization.

The IGP converged well without any loss, but it took one second for LDP to recalculate the new labels and set them up into the line cards. Future enhancements will improve this recovery behavior.

A good way to decrease this back to normal synchronization problem is to use targeted-LDP in-between direct neighbors, instead of using automatic discovery. This allows the LDP session to not be dropped on neighboring failure, and to not wait for an LDP label advertisement when the neighbor comes back to normal.

So, at the global configuration level, you can optimize performance by entering the list of direct neighbors:

```
mpls ldp neighbor 10.99.65.2 targeted ldp
..
```

Consider that the MAN always recover s in less than 2s in an upper-layer convergence design.

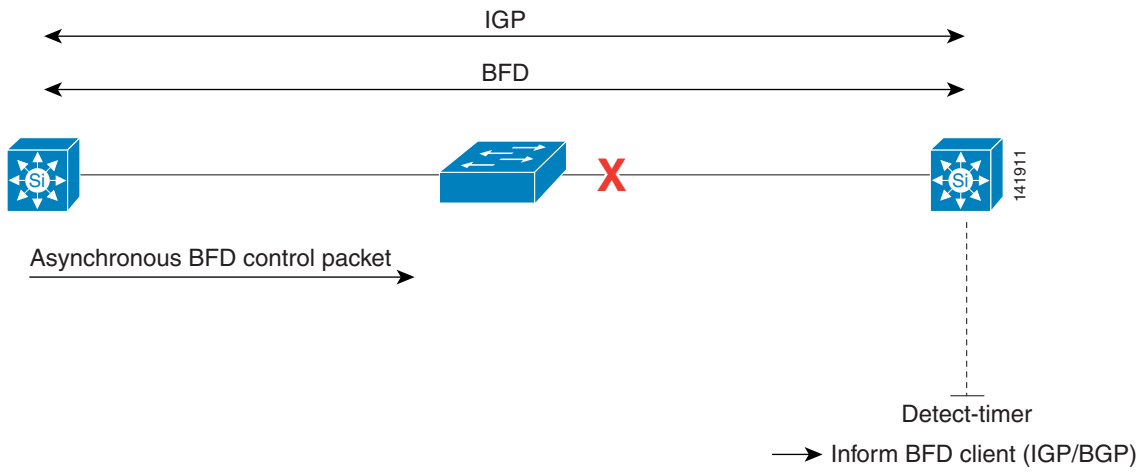
BFD (Bi-Directional Failure Detection)

In some specific environments, the connectivity failure might be difficult to detect in less than one second, and this is typically the case if the core links are provided using a plain Ethernet switching technology, such as QinQ, or even sometimes with some Ethernet over SONET devices.

As previously described, the IGP will not detect this kind of failure in less than one second. If sub-second detection is mandatory, then a new Layer 3 neighboring failure detection mechanism, BFD, can be used. BFD is a scalable asynchronous hello mechanism, which substitutes itself to the IGP hellos. As with any failure detection mechanism, timers must be tuned to avoid false detection. A few 100 ms would be, in many cases, adapted to enterprise requirements, while sub-100 ms detection is also feasible, it requires care.

As BFD is associated to IGP, flapping is managed by the back-off capabilities of the IGP, with BFD being just a link/Layer 2-path detection failure mechanism. Figure 5-17 illustrates the bi-directional failure detection process.

Figure 5-17 Bi-Directional Failure Detection




```
   interface Vlan600
 ip address …
 ip router isis
 bfd interval 10 min_rx 10 multiplier 3
 bfd neighbor 10.10.0.18
 dampening                ! protection against link flapping

router isis
 bfd all-interfaces

ping mpls pseudowire ipv4 10.99.65.1 255.255.255.255 time 1 int 100 repeat 1000
Sending 1000, 100-byte MPLS Echos to 10.99.65.1/32,
     timeout is 1 seconds, send interval is 100 msec:

!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
3w0d: %BFD_TEST-5-TRANSITION: Neighbor 10.10.0.17 DOWN - admin is FALSE
3w0d: %CLNS-5-ADJCHANGE: ISIS: Adjacency to Cat6k-DC2-left (vlan600)
Down, BFD !
3w0d: %LDP-5-NBRCHG: LDP Neighbor 10.99.65.4:0 is DOWN!!!!!!!!!!!
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
    Codes: '!' - success, 'Q' - request not transmitted,
           '.' - timeout, 'U' - unreachable,
           'R' - downstream router but not target,
           'M' - malformed request

Type escape sequence to abort.
```

Note As you can see, no ping was dropped, meaning a sub-100 ms convergence time.

It is considered best practice to protect against any instability by setting BFD to 100 ms at the beginning:

```
bfd interval 100 min_rx 100 multiplier 3
```

Improving Convergence Using Fast Reroute

Another approach to fast convergence would be to use the capacities of MPLS in terms of Traffic-engineering. This approach is more complex than a plain IGP tuning, and, in general, it requires more design work.

Note IGP fast-convergence is often adequate to meet requirements, so it is best to start simple, before implementing Traffic-engineering Fast Reroute (FRR).

MPLS Traffic-Engineering FRR is based on a similar concept to Optical backup. It supposes that a pre-set alternate path as been prepared to diverge local traffic as soon as a failed link or node as been detected. This minimizes the propagation and new path recalculation time.

In a complex network, this time can be quite large, specifically in a MAN where the common topology is a dual ring including all sites. FRR reduces the convergence time down to a few tenths of milliseconds after failure detection. The failure detection might be a physical link down or RSVP fast-hello detection.

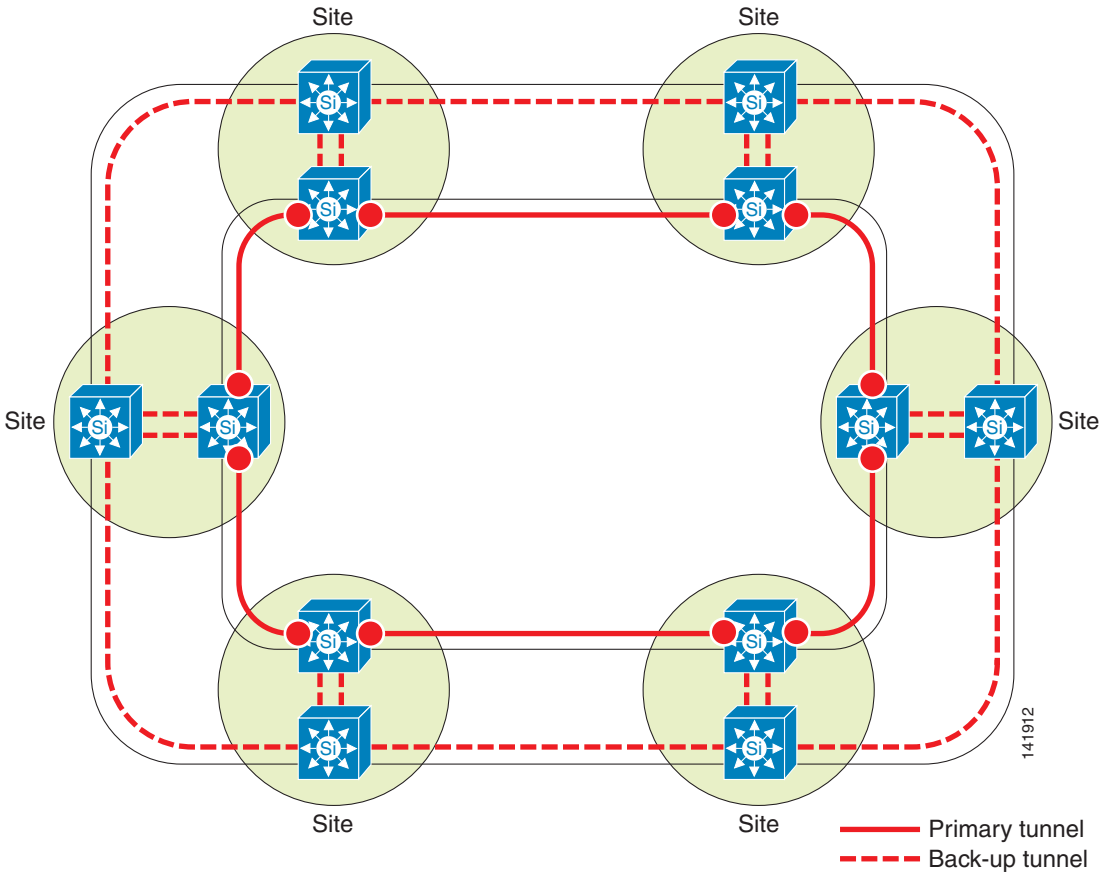
This guide does not focus on the usage of FRR that would require a separated analysis. FRR can protect link failures or even P-nodes failures. In a MAN design, where all nodes are PE, the FRR node protection can be useless. However, the weak part of a MAN is very often the link connection between sites.

The following three possible FRR configurations are possible:

- Manual settings:
  - Manual setting of primary tunnel.
  - Manual setting of the alternate tunnel.
- Automatic protection of each link:
  - Auto-setting of one-hop primary tunnels on all links.
  - Auto-building of back-up paths.
  - Not yet available on the Catalyst 6500 or the 7600 router.
- Automatic PE-PE full-meshing:
  - All PE are tunneled end-to-end to every other PE (automatic setting).
  - Auto-building of backup paths.
  - Not yet available on the Catalyst 6500 or the 7600 router.

Only one approach, the manual approach, is applicable to a network with only PE, such as in a MAN. Figure 5-18 illustrates the FRR design.

Figure 5-18 FRR Link-Protection Design - Dual Ring



Note that FRR protects TE tunnels, but not the plain IP or LDP packets. That means that traffic is protected only when pushed into Traffic-Engineering tunnels first, only then will these tunnels be FRR-protected.

Example of FRR setting:

Globally enable MPLS Traffic-Engineering:

```
mpls traffic-eng tunnels
```

Set traffic-engineering attributes on every core link:

```
Int …
 mpls traffic-eng tunnels
 ip rsvp bandwidth
```

Enable routing protocol transport of Traffic-engineering attributes:

```
router isis
     mpls traffic-eng router-id Loopback99
     mpls traffic-eng level-1
```

Manually create a Primary tunnel on the main path:

```
interface tunnel 25
     ip unnumbered Loopback99
     mpls ip                  ! Allows LDP over TE
     tunnel destination 10.99.72.5
     tunnel mode mpls traffic-eng
```

```
    tunnel mpls traffic-eng bandwidth 0
    tunnel mpls traffic-eng path-Option 1 dynamic
    tunnel mpls traffic-eng autoroute announce
    tunnel mpls traffic-eng fast-reroute
```

Create the backup path:

(Here, on the shortest path that does not use a primary link)

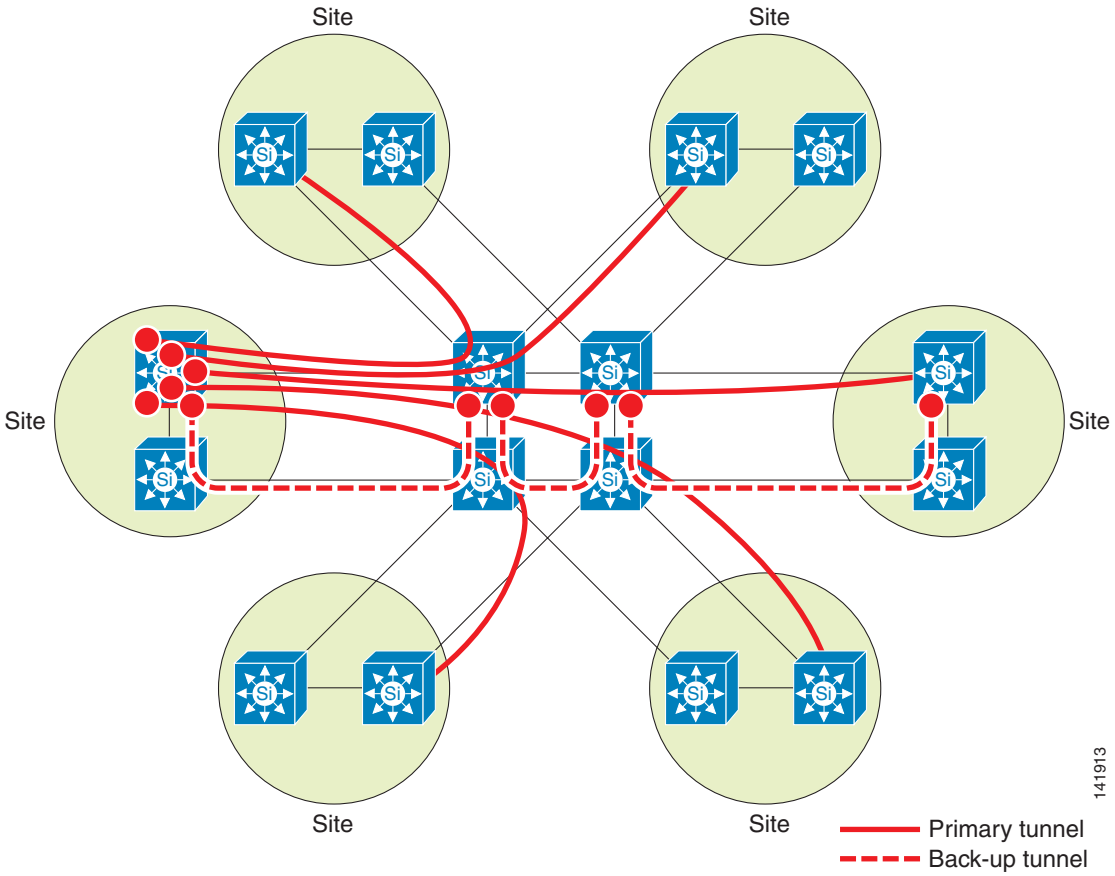
```
interface Tunnel 251
    ip unnumbered Loopback99
 tunnel destination 10.99.72.5
    tunnel mode mpls traffic-eng
 tunnel mpls traffic-eng path-Option 1 explicit name POS2

ip explicit-path name POS2 enable
    exclude-address 192.168.25.2
 exit
```

Associate a backup path to the Primary link:

```
interface pos 2/0
    mpls traffic-eng backup-path tunnel 251
```

An alternative design, especially applicable to a partially-meshed network, consists in the setting of a full-mesh of TE tunnels between PEs that must benefit from FRR, and then protect links or even core nodes along the path. Figure 5-19 illustrates an FRR link-protection design with a partially-meshed MAN.

Figure 5-19 FRR Link-Protection Design—Partially Meshed MAN



High Availability for Extended Layer 2 Networks

It is common practice to add redundancy to the interconnect between two data centers to avoid split-subnet scenarios and interruption of the communication between servers.

Note Proper design can address both problems. For example, the split-subnet is not necessarily a problem if the routing metric makes one site preferred over the other. Also, if the servers at each site are part of a cluster and the communication is lost, other mechanisms (such as the quorum disk) avoid a split-brain scenario.

Adding redundancy to an extended Ethernet network typically means relying on spanning tree to keep the topology free from loops. STP domains should be reduced as much as possible and limited inside the data center.

Cisco does not recommend that you deploy the legacy 802.1d because of its old timer-based mechanisms that make the recovery time too slow for most applications including typical clustering software. Consider using Rapid PVST+ or MST instead.

Etherchannel (with or without 802.3ad, Link Aggregation Control Protocol) provides an alternative to STP when using multiple links between sites for redundancy. With an EtherChannel, you can aggregate multiple physical links to shape a logical link while the traffic is load distributed over all available physical links. If one physical link fails, the switch redistributes all the traffic through the remaining active links.

EoMPLS Port-based Xconnect Redundancy with Multiple Spanning Tree Domains

This design uses the Multiple Spanning Tree (MST) protocol. The main enhancement introduced with MST is to allow several VLANs to be mapped into a single spanning tree instance. It is also easier to control a geographic region of a spanning tree domain containing multiple spanning tree instances (similar to an Autonomous System with a Layer 3 protocol such as Border Gateway Protocol (BGP).

With 802.1s, the bridges exchange a table containing the information that built the MST region:

- MST name
- Revision number
- VLAN mapping to instances

When a bridge receives this information, it compares it to the parameters of its MST region. If only one parameter differs from its database, the port receiving this digest is at the boundary of its MST region, which means that the remote spanning tree domain does not belong to the same MST region. Any communication to and from this interface outside of the MST region uses the IST-0 (Instance 0).

When connecting two MST regions together, Rapid Spanning Tree Protocol (RSTP) is used to control the spanning tree between two boundary ports of each MST region. IST (Instance 0) maps all VLANs to communicate outside of the MST region. Implicitly, the IST root is located inside the MST region. When two MST regions communicate with each other, they use the Instance 0; therefore, they use 802.1w BPDU (RSTP) to control any Layer 2 loop topology. All VLANs are mapped to the RSTP instance (IST).

The MST region, at its boundary interfaces, should be able to interoperate with any STP or RSTP devices as follows:

- RSTP (802.1w) to communicate with another MST region
- RSTP (802.1w) to communicate with a RSTP device connected to a boundary interface
- STP (802.1d) to communicate with an STP device connected to a boundary interface

Note The topology changes should not be affected by any WAN /MAN failure because they are masked by the MPLS convergence. Therefore, Instance 0 should only converge on aggregation switch failure.

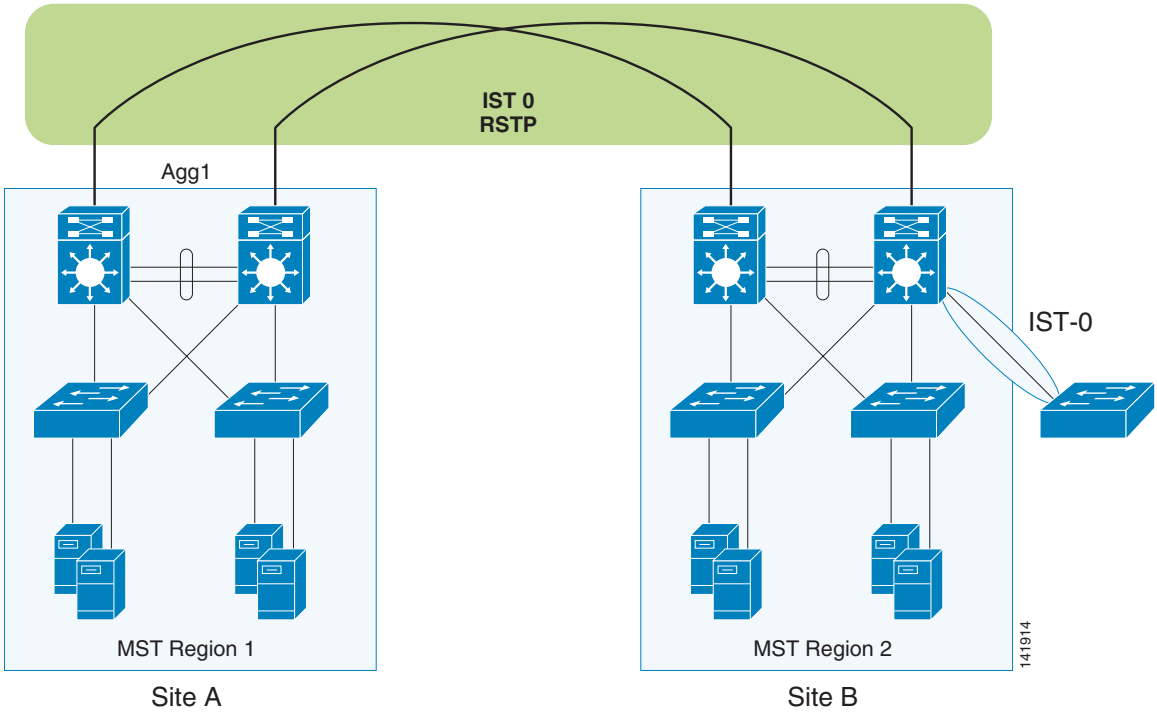
IST Everywhere

Only one single root bridge is elected for the IST - Instance 0 - for the whole Layer 2 network. In addition, any Layer 2 switch without MST enabled, or with MST configured to belong to different MST regions, will use the same Instance 0 to communicate with the original MST region, regardless of to which switch in the MST region it is connected.

This means that two data centers that belong to different MST regions are interconnected using the same IST. One of the data centers will support the primary root bridge for the whole extended Instance 0.

As shown in Figure 5-20, an additional switch, not configured to be in the same MST region (DC 2), is connected to one of the aggregation switches of an MST region. As the Instance 0 shares the same IST root bridge, the root bridge for that Instance 0 is the aggregation switch of DC 1.

Figure 5-20 MST and Instance 0



Interaction between IST and MST Regions

Root Bridges are defined per instance of spanning tree within a well delimited MST region. Any change in topology within the MST region is not propagated outside of its spanning tree instance, except for the Instance 0.

With IST (Instance 0) existing on all ports inside and outside of the MST region, there are some situations where a topology change can block the ports of the switches in a different MST region. For example, if the secondary root bridge is located in a different MST region and an alternate port in the same region begins forwarding traffic because of the topology change, then all downstream ports will go through the blocking stage (~1sec). To prevent this behavior, Cisco recommends enabling the secondary root switch inside the same MST region where the primary root bridge seats. In case of the IST root bridge failing, it will perform a flush on the remote CAM tables. However, the traffic is not disrupted.

First, a specific MST region is created on each data center: MST DC1 and MST DC2. Only the Instance 0 is used to communicate between two different MST regions using RSTP.

Both data centers are interconnected using two pseudowires (Layer 2 VPN) as shown in Figure 5-5.

Between two MST regions, the IST (Instance 0) cannot use PVST+ or Rapid PVST+, but only RSTP and eventually 802.1d in some rare cases. Therefore, it is not possible to load-share the traffic per instance of STP on different physical links, with the IST-0 being present on all interfaces. Figure 5-21 and Figure 5-22 illustrate MST that is dedicated to an external cluster.

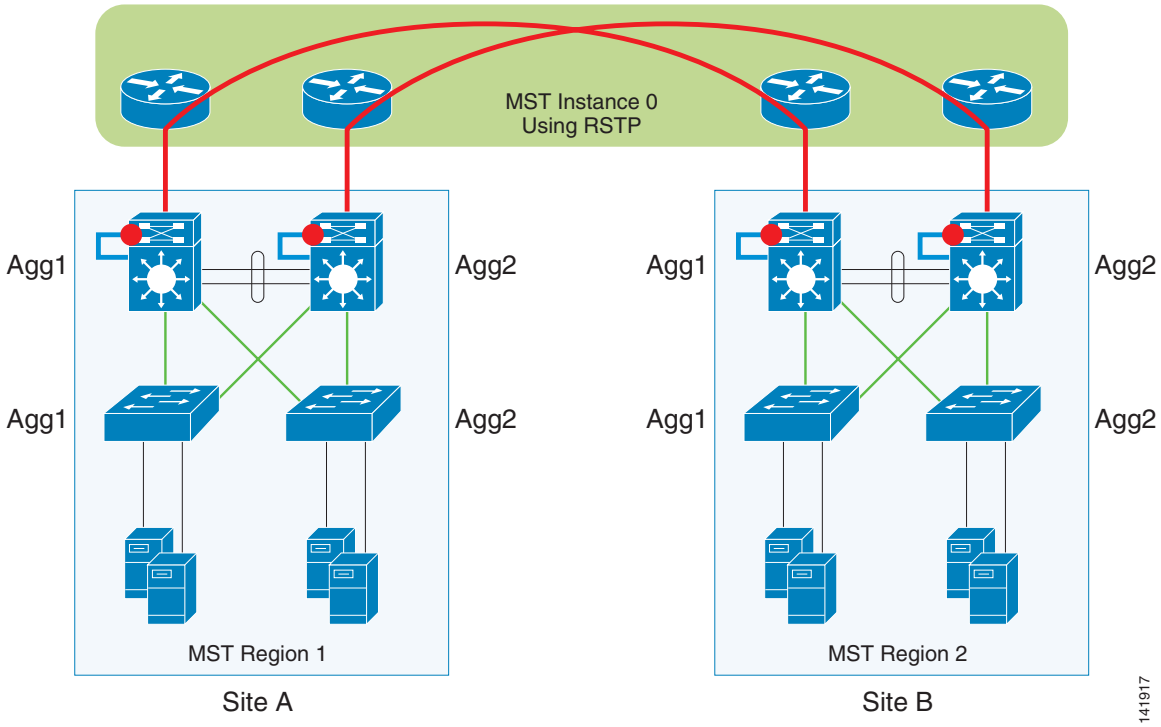
Figure 5-21 MST Dedicated for An Extended Cluster

From outside, the MST region is seen as a single logical switch, regardless the number of switches on each MST region, or boundary ports that belong to different devices.

Because of RSTP being used on the Instance 0, if the forwarding logical Layer 2 connection from site A to site B fails, the backup Layer 2 link will take a sub-second or so to failover - a sub-second failover in case of a direct physical link failure, otherwise it takes three times the BPDU hello timers before a timeout occurs (by default, 6 seconds).

Note The term "logical Layer 2 connection" is used in this guide because the extended Layer 2 is built on top of MPLS (EoMPLS), so any physical failure is fully controlled by the Layer 3 fast convergence protocol (IGP and BFD), as explained in MPLS Technology Overview, page 5-8.

Figure 5-22 EoMPLS at the DC Aggregation layer



This design assumes the aggregation switch 1 (Agg1) on DC1 is the root bridge for the MST region DC1 Instance 1 and Instance 2 and aggregation switch 1 (Agg1) on DC2 forms the root bridge for the MST region DC2 Instance 1 and Instance 2.

Therefore, you have a root bridge on each site for the Instance 1 and Instance 2; note that this is for the Instances limited inside each MST region.

The Instance 0 is used to communicate to outside the MST region and being present on each port of each bridge, only 1 root bridge for the whole IST 0 will exist inside the whole Layer 2 network. Aggr1 on DC1 have been chosen as the root bridge for the Instance 0 (IST Root) and Aggr2 on DC1 to be the secondary root bridge for the Instance 0.

Consequently, you have the following STP state on each switch:

- MST Interface State for Data Center 1 (MST Region 1—Instance 0):

```
Aggregation 1
Aggr1 - Interface 3/46 connected to Aggr2 is Designated Forwarding
Aggr1 - Interface 3/47 connected to the core (MAN) is Designated Forwarding and
boundary using RSTP
Aggr1 - Interface 3/48 connected to Access switch is Designated Forwarding

Interface        Role Sts Cost      Prio.Nbr Type
Fa3/46           Desg FWD 200000    128.302  P2p
Fa3/47           Boun FWD 200000    128.303  P2p Bound(RSTP)
Fa3/48           Desg FWD 200000    128.304  P2p

Aggregation 2
Aggr2 - Interface 3/46 connected to Aggr1 (root) is Root Forwarding
Aggr2 - Interface 3/47 connected to the core (MAN) is Designated Forwarding and
boundary using RSTP
Aggr2 - Interface 3/48 connected to Access switch is Designated Forwarding

Interface        Role Sts Cost      Prio.Nbr Type
Fa3/46           Root FWD 200000    128.302  P2p
Fa3/47           Boun FWD 200000    128.303  P2p Bound(RSTP)
Fa3/48           Desg FWD 200000    128.304  P2p

Access 1
Acc1 - Interface 2 connected to Aggr1 is Root Forwarding
Acc1 - Interface 4 connected to Aggr2 is Alternate
There is no Boundary interface on this switch.

Interface        Role Sts Cost      Prio.Nbr Type
Gi1/0/2          Root FWD 200000    128.2    P2p
Gi1/0/4          Altn BLK 200000    128.4    P2p
```

- MST Interface State for Data Center 2 (MST Region 2—Instance 0):

```
Aggregation 1
Aggr1 - Interface 1/46 connected to Aggr1 is Designated Forwarding
Aggr1 - Interface 1/47 connected to the core (MAN) is Root Forwarding and boundary
using RSTP
Aggr1 - Interface 1/48 connected to Access switch is Designated Forwarding

Interface        Role Sts Cost      Prio.Nbr Type
Fa1/46           Desg FWD 200000    128.46   P2p
Fa1/47           Root FWD 100000    128.47   P2p Bound(RSTP)
Fa1/48           Desg FWD 200000    128.48   P2p


Aggregation 2
Aggr2 - Interface 3/46 connected to Aggr2 is Root Forwarding (active path to the root)
Aggr2 - Interface 3/47 connected to the core (MAN) is Alternate blocking and boundary
using RSTP
Aggr2 - Interface 3/48 connected to Access switch is Designated Forwarding

Interface        Role Sts Cost      Prio.Nbr Type
```

```
Fa1/46           Root FWD 200000    128.46   P2p
Fa1/47           Altn BLK 200000    128.47   P2p Bound(RSTP)
Fa1/48           Desg FWD 200000    128.48   P2p

Access 1
Acc1 - Interface 2 connected to Aggr1 is Alternate
Acc1 - Interface 4 connected to Aggr2 is Root Forwarding
There is no Boundary interface on this switch.

Interface        Role Sts Cost      Prio.Nbr Type
Gi1/0/2          Altn BLK 200000    128.2    P2p
Gi1/0/4          Root FWD 200000    128.4    P2p
```

For VLAN mapping, VLAN 601 and 602 are mapped to Instance 1, all other VLANs are mapped to Instance 2.

In these tests, VLAN 601 is used for the cluster VIP and VLAN 602 is used for the health check of the cluster's members and these are extended up to the remote data center.

In this configuration, the boundaries of the MST regions are located at the inbound interface used for the loopback cable (n/47). On those boundary ports, IST0 is used to communicate between the two MST regions.

Note The loopback cables shown in Figure 5-5 are used to allow the VLAN dedicated for the pseudowire to be switched to outside of the data center (for example, the VLAN 601 used for VIP).

Configuration

This section provides examples for the MST configuration.

Aggregation Switch Left (Primary Root Bridge for MST Region DC1)

```
spanning-tree mode mst
no spanning-tree optimize bpdu transmission
spanning-tree extend system-id
spanning-tree mst configuration
 name DC1
 revision 10
 instance 1 vlan 601-602
 instance 2 vlan 1-600, 603-4094
!
spanning-tree mst 1-2 priority 24576
```

Aggregation Switch Right (Secondary Root Bridge for MST Region DC1)

```
spanning-tree mode mst
no spanning-tree optimize bpdu transmission
spanning-tree extend system-id
spanning-tree mst configuration
 name DC1
 revision 10
 instance 1 vlan 601-602
 instance 2 vlan 1-600, 603-4094
!
spanning-tree mst 1-2 priority 28672
```

Access Switch (MST Region DC1)

```
spanning-tree mode mst
no spanning-tree optimize bpdu transmission
spanning-tree extend system-id
spanning-tree mst configuration
 name DC1
 revision 10
 instance 1 vlan 601-602
 instance 2 vlan 1-600, 603-4094
```

Aggregation Switch Left (Primary Root Bridge for MST Region DC2)

```
spanning-tree mode mst
no spanning-tree optimize bpdu transmission
spanning-tree extend system-id
spanning-tree mst configuration
 name DC2
 revision 20
 instance 1 vlan 601-602
 instance 2 vlan 1-600, 603-4094
!
spanning-tree mst 1-2 priority 24576
```

Aggregation Switch Right (Secondary Root Bridge for MST Region DC2)

```
spanning-tree mode mst
no spanning-tree optimize bpdu transmission
spanning-tree extend system-id
spanning-tree mst configuration
 name DC2
 revision 20
 instance 1 vlan 601-602
 instance 2 vlan 1-600, 603-4094
!
spanning-tree mst 1-2 priority 28672
```

Access Switch (MST Region DC2)

```
spanning-tree mode mst
no spanning-tree optimize bpdu transmission
spanning-tree extend system-id
!
spanning-tree mst configuration
 name DC2
 revision 20
 instance 1 vlan 601-602
 instance 2 vlan 1-600, 603-4094
```

EoMPLS Port-based Xconnect Redundancy with EtherChannels

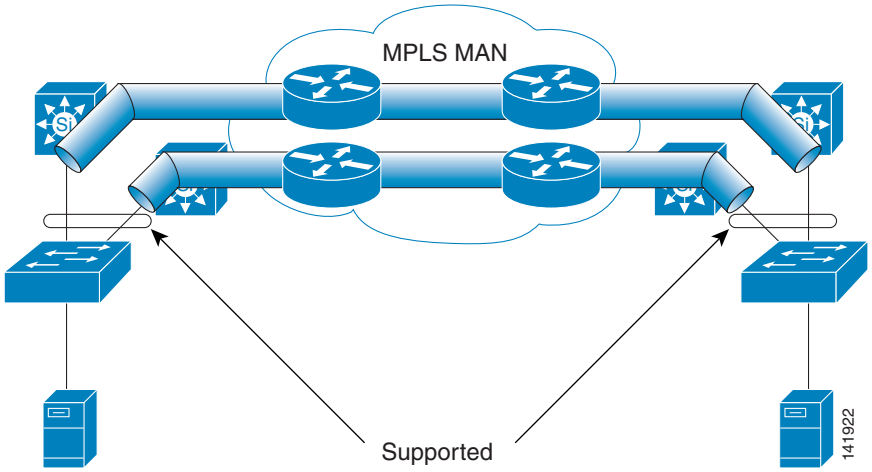
EtherChannel provides incremental trunk speeds from one to up to eight physical links, between Fast Ethernet, Gigabit Ethernet, and 10 Gigabit Ethernet. EtherChannel combines multiple Fast Ethernet up to 800 Mbps, Gigabit Ethernet up to 8 Gbps, and 10 Gigabit Ethernet up to 80 Gbps.

EtherChanneling provides an alternative to MST to keep redundant Layer 2 paths free from loops.

One option you can use to deploy an EtherChannel design while preventing a single point of failure is to encapsulate each physical link into a tunnel.

Figure 5-23 shows EtherChanneling between ports of remote switches. The switch at each data center is connected to each local aggregation switch. EtherChanneling end-to-end is possible because of the EoMPLS pseudowire that provides Layer 2 connectivity between the remote switches.

Figure 5-23 Split EtherChannel Over EoMPLS



As long as the Layer 2 VPN tunnel is maintained from end to end, the flow path can take multiple routes inside the WAN MAN, using Layer 3 or MPLS fast convergence algorithms. This keeps any convergences on the WAN MAN fully transparent for the edge devices (or access switches).

Figure 5-23 shows the logical equivalent of two access switches connected together through an EtherChannel built with two direct physical links.

Figure 5-24 Logical EtherChannel Over EoMPLS

Remote Failure Detection

In a native Layer 2 EtherChannel, if a segment within the channel fails, the traffic previously carried over the failed link switches to the remaining links within the EtherChannel.

However, this requires that the affected interface of the EtherChannel detects a physical link down. When using a Layer 2 VPN, the Logical Layer 2 link is built on top of existing transport layer (a pseudowire). This means that if one of the remote Layer 2 links (remote data center) fails for any reason, the local EtherChannel is not able to detect the remote failure as a direct physical link failure, therefore, the EtherChannel protocol used to communicate (LACP or PAgP) times out after the certain time. With LACP (802.1ad), the timeout happens after 60 seconds, which means during 1 minute or so, the local site continues to send the traffic over this link, which is no longer terminated on the remote site.

You can use UDLD to make the detection faster.

Unidirectional Link Detection (UDLD)

UDLD has been designed and implemented by Cisco to detect unidirectional links and improve a Layer 1-Layer 2 Loop detection, usually controlled by a Layer 2 algorithm, such STP or RSTP.

UDLD is a Layer 2 protocol that works in conjunction with Layer 1 mechanisms to determine the physical status of a link. At Layer 1, auto-negotiation takes care of physical signaling and fault detection. UDLD detects the identities of neighbors and shuts down misconnected ports. When enabling both auto-negotiation and UDLD, Layer 1 and Layer 2 detections work together to prevent physical and logical unidirectional connections and the malfunctioning of other protocols.

UDLD works by exchanging protocol packets between the neighboring devices. In order for UDLD to work, both devices on the link must support UDLD and have it enabled on their respective ports.

Each switch port configured for UDLD sends UDLD protocol packets containing the port's own device or port ID, and the neighbor's device or port IDs, as seen by UDLD on that port. Neighboring ports should see their own device or port ID (echo) in the packets received from the other side. If the port does not see its own device or port ID in the incoming UDLD packets for a specific duration of time, the link is considered unidirectional. This heartbeat, based on a echo-algorithm, allows detection of several issues, such as damaged wiring, fiber mistakes, or, in this design, a remote link failure after a specific timeout.

To take advantage of EtherChannel, it is important to enable UDLD to improve the detection of a remote failure before the timeout. Port-based Xconnect of the Layer 2 VPN tunnel must be created at the ingress port directly linked to the access layer, as shown in Figure 5-23. This architecture is described in Figure 5-4. Until recently, the default message interval value was 7 seconds; therefore, the timeout before detecting a remote link failure was 21 seconds. This value is not fast enough for a cluster, which requires a maximum of 10 seconds to detect a remote link failure and keep the cluster recovery algorithm working as expected.

The minimum message interval has recently been reduced to 1. Deploying UDLD with this low interval can cause problems. Some tests through long distances (>20kms) have been conducted with the message interval value set to 1 and these have shown a very unstable behavior of UDLD. The minimum message interval value setting for UDLD, to keep it stable over long distances, is >3-4 seconds. Therefore, the minimum time to detect a remote failure becomes between 12 and 15 seconds, a value still insufficient for HA clusters. UDLD requires more analysis to understand why it becomes unstable when using 1 sec for the message interval.

UDLD Modes

UDLD can operate in two modes:

- Normal mode
- Aggressive mode

In normal mode, if the link state of the port is determined to be bi-directional and the UDLD information times out, no action is taken by UDLD. The port state for UDLD is marked as undetermined. The port behaves according to its STP state.

In aggressive mode, if the link state of the port is determined to be bi-directional and the UDLD information times out while the link on the port is still up, UDLD tries to re-establish the state of the port. If unsuccessful, the port is put into an errdisable state. This requires a manual action to re-enable the port using the following command:

```
switch#udld reset
```

It is up to the network manager to decide if any disruption on the tunnel should be transparent and dynamically re-initiate, or if it should force a manual reset.

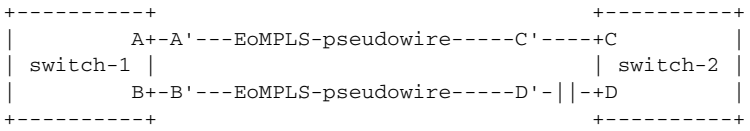
UDLD Configuration

From a global setting, define the same message interval on both access switches:

```
switch#udld message time 4

On each uplink interface enable UDLD

interface GigabitEthernet1/0/2
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan 1,601,602
 switchport mode trunk
 load-interval 30
 udld port aggressive
 no mdix auto
```

EoMPLS Port-based Xconnect Redundancy with Spanning Tree

Spanning tree is an alternative to the use of EtherChannels, in that it can keep the topology free from loops. By using 802.1w, the convergence time for any failure is around ~4s, which is better than what you can achieve with EtherChannels combined with UDLD. The spanning tree BPDUs perform the role of UDLD frames by verifying the availability of the path between the sites (one BPDUs is sent every two seconds). The topology with spanning tree follows. The main drawback is that one link is blocking, while with EtherChannels, all links are used.



For example, consider two simple failures. When C' fails, there is almost no traffic loss. The reason is that switch-2 switches D into Forwarding immediately, while on switch-1 traffic gets flooded on both port A and port B.

On switch-1, before the link going down on the remote switch, you have the following entries in the Layer 2 forwarding table:

```
20    0000.0c07.ac01    DYNAMIC     Fa1/0/15
20    0005.5f0b.2800    DYNAMIC     Gi1/0/1 <<<<<<
20    0011.bb0f.b301    DYNAMIC     Gi1/0/1 <<<<<<
20    0030.4880.4d1f    DYNAMIC     Fa1/0/5
20    0030.4880.4d23    DYNAMIC     Gi1/0/1 <<<<<<
20    00d0.020e.7400    DYNAMIC     Fa1/0/15
```

After the failure of port C, you have the following entries:

```
20    0000.0c07.ac01    DYNAMIC     Fa1/0/15
20    0005.5f0b.2800    DYNAMIC     Gi2/0/2
20    0011.bb0f.b31b    DYNAMIC     Gi2/0/2
20    0030.4880.4d1f    DYNAMIC     Fa1/0/5
20    0030.4880.4d23    DYNAMIC     Gi2/0/2
20    00d0.020e.7400    DYNAMIC     Fa1/0/15
```

In the case of port A' failing, the downtime is higher; there is packet drop for ~4s. The reason is that it takes longer for switch-2 to put the alternate port into forwarding mode.

```
Interface        Role Sts Cost      Prio.Nbr Type
---------------- ---- --- --------- --------------
```

```
Gi1/0/1          Root FWD 20000     128.1    P2p
Fa1/0/3          Desg FWD 200000    128.5    Edge P2p
Fa1/0/15         Desg FWD 200000    128.17   P2p
Gi1/0/3          Altn BLK 20000     128.27   P2p

Switch-2#
Nov  1 16:17:02: %SPANTREE-5-TOPOTRAP: Topology Change Trap for vlan 21
Nov  1 16:17:03: %SPANTREE-2-LOOPGUARD_BLOCK: Loop guard blocking port
GigabitEthernet1/0/1 on VLAN0020.
Nov  1 16:17:03: %SPANTREE-5-TOPOTRAP: Topology Change Trap for vlan 20

Switch-2#show spanning-tree vlan 20

Interface        Role Sts Cost      Prio.Nbr Type
---------------- ---- --- --------- --------
Gi1/0/1          Root BKN*20000     128.1    P2p *LOOP_Inc
Fa1/0/3          Desg FWD 200000    128.5    Edge P2p
Fa1/0/15         Desg FWD 200000    128.17   P2p
Gi1/0/3          Root FWD 20000     128.27   P2p
```

Layer 2 loops with this design are a concern, but they are less disruptive than in a regular LAN design. The maximum bandwidth that they can use from the MPLS core is limited by the link bandwidth connecting the CE switch to the aggregation switch. As long as the EoMPLS link carries only the LAN extension traffic, and the FC traffic uses the MPLS network or another transport, a Layer 2 loop is going to cause a high bandwidth utilization on the local and remote LAN. However, it is not going to make the aggregation switches unmanageable, and it is not going to cause the storage arrays to be in a split state. Design the network to avoid loops (spanning tree is used for this purpose) because bugs or other mistakes (such as configuring the teamed NICs of a server for forwarding) can potentially introduce loops.

Metro Ethernet Services

Metro Ethernet Service Framework

This chapter describes the typical Metro Ethernet Services available from service providers (SPs). For the most part, these services are derived from and map to the following Metro Ethernet Forum (MEF) specifications:

- MEF 6, Ethernet Services Definitions—Phase 1, June 2004
- MEF 10, Ethernet Services Attributes—Phase 1, November 2004

Note The MEF technical specifications can be found at the MEF website at the following URL: http://www.metroethernetforum.org/.

These MEF technical specifications describe the attributes and associated parameters that define specific Ethernet services. They also provide a framework for characterizing Ethernet services, which can be used by SPs in their deployments, or by design and sales engineers in responding to SP request for proposals (RFPs).

Following the MEF approach, the services that comprise the Metro Ethernet (ME) solution can be classified into the following two general categories:

- Point-to-point (PtP)—A single point-to-point Ethernet circuit provisioned between two User Network Interfaces (UNIs).
- Multipoint-to-multipoint (MPtMP)—A single multipoint-to-multipoint Ethernet circuit provisioned between two or more UNIs. When there are only two UNIs in the circuit, more UNIs can be added to the same Ethernet virtual connection if required, which distinguishes this from the point-to-point type.

In the MEF terminology, this maps to the following Ethernet service types:

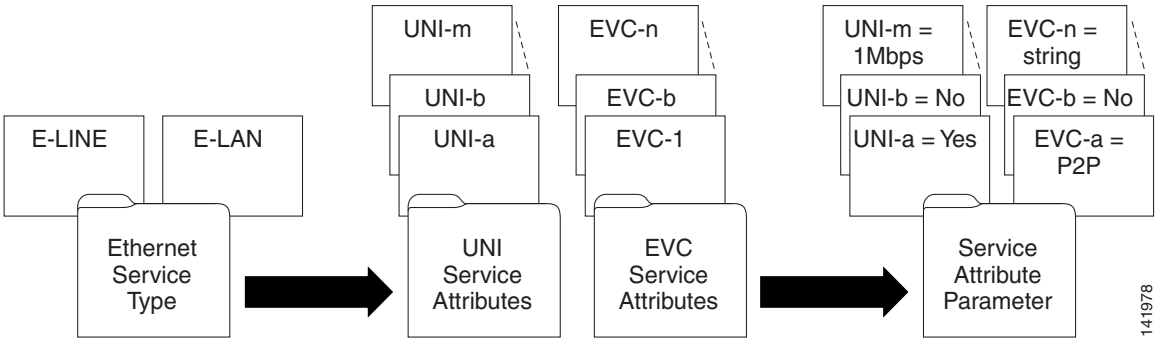
- Ethernet Line Service Type (E-Line)—Point-to-point Ethernet service
- Ethernet LAN Service Type (E-LAN)—Multipoint-to-multipoint Ethernet service

Within these two service types, Metro Ethernet services can be created by assigning values to a set of attributes grouped according to the following:

- User Network Interface (UNI)—Physical demarcation point between the responsibility of the SP and the responsibility of the subscriber.
- Ethernet virtual connection (EVC)—Association of two or more UNIs that limits the exchange of service frames to UNIs within the EVC.

Figure 6-1 illustrates the service definition framework described above.

Figure 6-1 Metro Ethernet Framework



MEF Services

MEF 6 defines two examples of E-Line services:

- Ethernet private line (EPL)—Uses a point-to-point EVC between two UNIs to provide a high degree of transparency such that service frames, headers, and most Layer 2 protocols are identical at both the source and destination UNI. It does not allow for service multiplexing; that is, a dedicated UNI (physical interface) is used for the EPL.
- Ethernet virtual private line (EVPL)—Uses a point-to-point EVC between two UNIs, but does not provide full transparency as with the EPL; that is, all Layer 2 control protocols are discarded at the UNI. The EVPL also allows for service multiplexing, which means that more than one EVC can be supported at the UNI, which cannot happen for the EPL.

As of publication of this document, the MEF has not yet defined multipoint services. However, a multipoint service type (E-LAN) does exist.

See Table 6-1 for a mapping of the above services with Cisco terminology.

Metro Ethernet Services

Before discussing ME services, note the following two definitions.

- Metro Ethernet service—A Metro Ethernet service is the combination of the UNI, EVC, and all associated attributes and parameters that together can be used by a SP to create an offering to their customers. These attributes and parameters describe specific properties of the UNI and EVC, as well as define the associated QoS, resiliency, security, and management features. This combination of attributes and parameters allows the SP to offer a service level agreement (SLA) to their customers. This section focuses on those attributes and parameters that describe the UNI and EVC.
- Metro Ethernet service frame—An Ethernet frame transmitted across the UNI toward the SP or an Ethernet frame transmitted across the UNI toward the subscriber.

ME services consist of various types of UNIs that are used in combination with EVCs, which can be built over Layer 1, Layer 2, or Layer 3 networks. This section provides a brief summary of these services, which are subsequently described in more detail:

- Ethernet relay service (ERS)—Point-to-point VLAN-based E-Line service that is used primarily for establishing a point-to-point connection between customer routers.

- Ethernet wire service (EWS)—Point-to-point port-based E-Line service that is used primarily to connect geographically remote LANs over an SP network.
- Ethernet multipoint service (EMS)—Multipoint-to-multipoint port-based E-LAN service that is used for transparent LAN applications.
- Ethernet relay multipoint service (ERMS)—Multipoint-to-multipoint VLAN-based E-LAN service that is used primarily for establishing a multipoint-to-multipoint connection between customer routers.
- Ethernet private line (EPL)—Port-based point-to-point E-Line service that maps Layer 2 traffic directly on to a TDM circuit.
- ERS access to MPLS VPN—Mapping of an Ethernet connection directly onto an MPLS VPN that provides Layer 2 access using an ERS UNI, but is a Layer 3 service as it traverses the MPLS VPN.
- ERS access to ATM service interworking (SIW)—Point-to-point VLAN-based E-Line service that is used for Ethernet to ATM interworking applications.

The ME services map to the MEF services (and service types in case of undefined services) described in Table 6-1

Table 6-1 MEF to Cisco Metro Ethernet Services Mapping

ME Service	MEF Equivalent Service/Service Type
EWS	EPL
ERS	EVPL
EPL	EPL
EMS	E-LAN service type
ERMS	E-LAN service type

These Metro Ethernet services are then defined by assigning specific attributes for the UNIs and EVCs. They are characterized by associating parameters to the attributes. The following sections describe the attributes for the EVC and UNI.

EVC Service Attributes

An EVC allows Ethernet service frames to be exchanged between UNIs that are connected via the same EVC. Some frames are subscriber data service frames while others are Ethernet control service frames. The following attributes describe the EVC:

- EVC type—The EVC can either be point-to-point or multipoint-to-multipoint.
- UNI list—This is the list of UNIs associated with an EVC.
- Service frame transparency—All fields of each egress service frame must be identical to the same fields of the corresponding ingress service frame, except as follows:
  - The egress service frame may have an IEEE 802.1Q tag, while the corresponding ingress service frame does not. In this case, the egress service frame must have a recalculated FCS.
  - The egress service frame may not have an IEEE 802.1Q tag, while the corresponding ingress service frame does have a tag. In this case, the egress service frame must have a recalculated FCS.

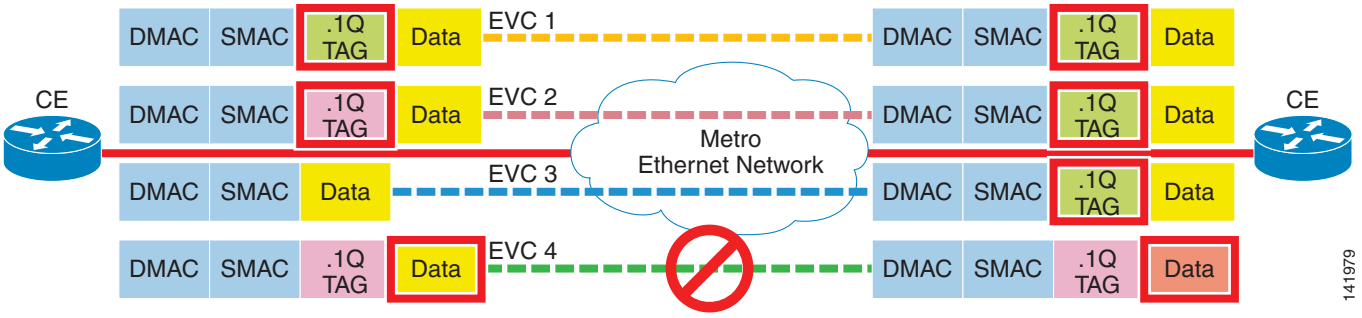
– If both the egress service frame and corresponding ingress service frame have an IEEE 802.1Q tag, the content of the tag in the egress service frame may be different from the content of the tag in the corresponding ingress service frame. If the contents of the ingress and egress tags are different, the egress service frame must have a recalculated FCS.

Note The service frame transparency EVC attribute is used in the solution to track the ability of the platform to maintain customer "DSCP transparency".

Figure 6-2 shows three possible cases (EVC 1 through EVC 3) of EVCs with service transparency, as well as a case (EVC 4) where service frame transparency is not achieved:

- For EVC 1, the entire ingress and egress frames are identical.
- For EVC 2, ingress and egress frames are identical with the exception of the 802.1Q tag.
- For EVC 3, ingress and egress frames are identical with the exception of the presence of an 802.1Q tag in the egress frame.
- For EVC 4, ingress and egress frames are not identical in the payload section of the frames. Examples of changes of the payload include changes in the IP header (for example, ToS field). EVC 4 is *not* service frame-transparent.

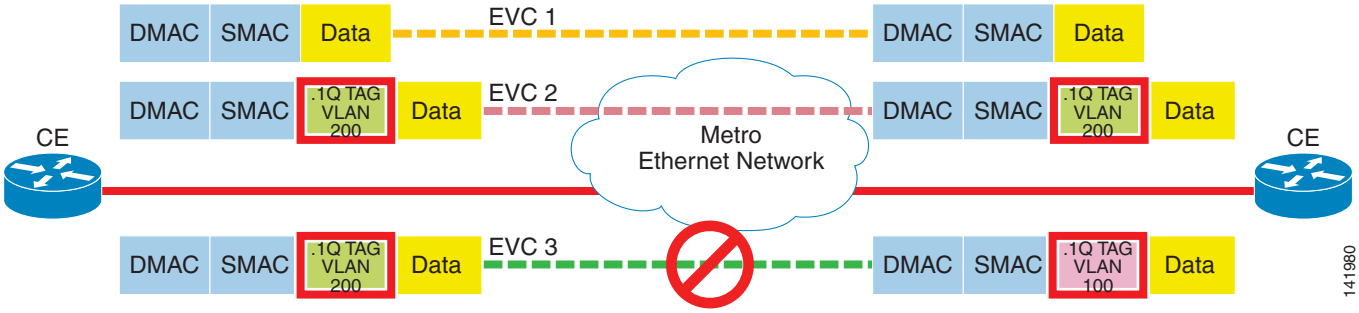
Figure 6-2 Service Frame Transparency EVC Attribute



- CE-VLAN ID preservation—Defines whether the CE-VLAN ID is preserved (unmodified) across the EVC. CE-VLAN ID preservation also implies that there is no constraint on the subscriber choice of VLAN ID or the number of VLAN IDs. Figure 6-3 shows two EVCs with support for this attribute (EVC 1 and EVC 2), and one without it (EVC 3).

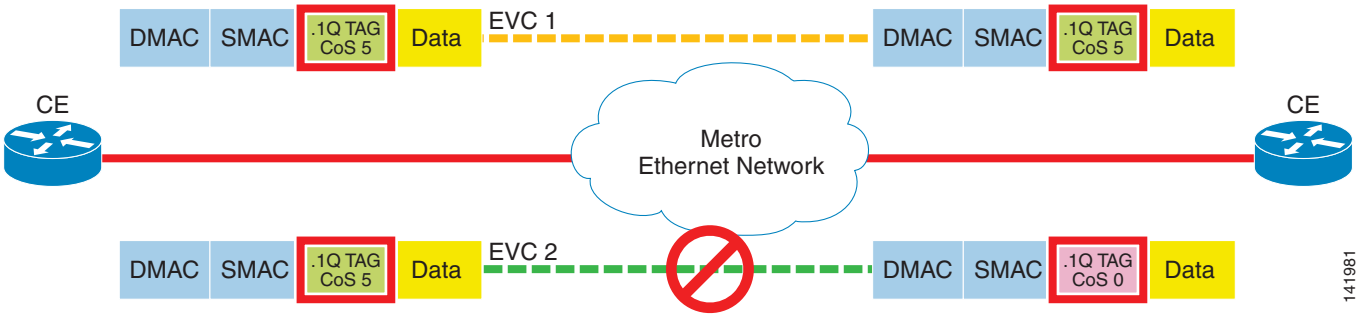
Note The CE-VLAN ID preservation EVC attribute used to be tracked as "VLAN transparency" in previous versions of the solution.

Figure 6-3 CE-VLAN ID Preservation EVC Attribute



- CE-VLAN CoS preservation—Defines whether the CE-VLAN CoS bits are preserved (unmodified) across the EVC. Figure 6-4 shows an EVC supporting the attribute (EVC 1) and an EVC without it (EVC 2).

Figure 6-4 CE-VLAN CoS Preservation EVC Attribute



- Unicast service frame delivery—A unicast service frame has a unicast destination MAC address. This EVC attribute specifies whether unicast service frames are discarded, delivered unconditionally, or delivered conditionally for each ordered UNI pair. If the services frames are delivered conditionally, the conditions must be specified.
- Multicast service frame delivery—A multicast service frame has a multicast destination MAC address. This EVC attribute specifies whether multicast service frames are discarded, delivered unconditionally, or delivered conditionally for each ordered UNI pair. If the services frames are delivered conditionally, the conditions must be specified.
- Broadcast frame delivery—A broadcast service frame has a broadcast MAC address. This EVC attribute specifies whether broadcast service frames are discarded, delivered unconditionally, or delivered conditionally for each ordered UNI pair. If the services frames are delivered conditionally, the conditions must be specified.
- Layer 2 control protocol processing—Can be applied at the EVC, and describes how to treat incoming Layer 2 control protocols. The allowable options are discard, tunnel, or peer.
- Class of service (CoS) identifier—Derived from one of the following:
  - The EVC to which the service frame is mapped
  - The combination of the EVC to which the service frame is mapped and a set of one or more CE-VLAN CoS values
  - The combination of the EVC to which the service frame is mapped and a set of one or more CE-VLAN DSCP values

- EVC performance—Specified for all service frames on an EVC with a particular CoS instance, which is identified by a CoS identifier (see previous attribute) associated with each service frame. The following parameters define the EVC performance:
  - CoS identifier
  - Frame delay
  - Frame delay variation
  - Frame loss

Table 6-2 summarizes the EVC attributes as defined generically in MEF 10, Ethernet Services Attributes, Phase 1 standard.

Table 6-2 Summary of MEF EVC Service Attributes

Attribute	Type of Parameter Value
EVC type	Point-to-point or multipoint-to-multipoint
UNI list	A list of UNI identifiers
Service frame transparency	Yes or no
CE-VLAN ID preservation	Yes or no
CE-VLAN CoS preservation	Yes or no
Unicast service frame delivery	Discard, deliver unconditionally, or deliver conditionally. If deliver conditionally is used, then the conditions *must* be specified.
Multicast service frame delivery	Discard, deliver unconditionally, or deliver conditionally. If deliver conditionally is used, then the conditions *must* be specified.
Broadcast service frame delivery	Discard, deliver unconditionally, or deliver conditionally. If deliver conditionally is used, then the conditions *must* be specified.
Class of service identifier	<EVC>, <EVC, DSCP>, or <EVC, COS>
EVC performance	Frame delay Frame delay variation Frame loss
Layer 2 Control Protocols Processing[1]	
Bridge block of protocols: 0x0180.C200.0000 through 0x0180.C200.000F	Discard or tunnel
GARP block of protocols: 0x0180.C200.0020 through 0x0180.C200.002F	Discard or tunnel
All bridges protocol 0x0180.C200.0010	Discard or tunnel

1. Note that SPs may define additional addresses as Layer 2 control in addition to those listed here.

ME EVC Service Attributes

Table 6-3 summarizes the EVC service attributes for each of the ME services. Note that not all of the MEF attributes are listed in this table (attributes used for record-keeping/inventory purposes have been omitted). Also, because the L2 control protocol processing for the ME services happens at the UNI, those attributes are not included for the EVC.

Table 6-3 ME EVC Service Attributes

	ME Services				
EVC Service Attribute	**ERS**	**ERMS**	**EWS**	**EMS**	**EPL**
EVC type	PtP	MPtMP	PtP	MPtMP	PtP
Service frame transparency[1]	Yes[2]	Yes	Yes	Yes	Yes
CE-VLAN ID preservation	Yes[3] or No	Yes[4] or No	Yes	Yes	Yes
CE-VLAN CoS preservation	No[5]	No[6]	Yes	Yes	Yes
Unicast[7] frame delivery	Deliver un-conditionally	Deliver un-conditionally	Deliver un-conditionally	Deliver un-conditionally	Deliver un-conditionally
Multicast frame delivery	Deliver conditionally per threshold	Deliver conditionally per threshold	Deliver conditionally per threshold	Deliver conditionally per threshold	Deliver un-conditionally
Broadcast frame delivery	Deliver conditionally per threshold	Deliver conditionally per threshold	Deliver conditionally per threshold	Deliver conditionally per threshold	Deliver un-conditionally
Class of service identifier	EVC <EVC, DSCP> <EVC, CoS>[8]	EVC <EVC, DSCP> <EVC, CoS>[9]	EVC <EVC, CoS>[10]	EVC <EVC, CoS>[11]	EVC
EVC performance	For each CoS instance, specify the frame delay, frame delay variation, and frame loss				

1. This is a *mandatory* attribute for all Layer 2 services of the solution.
2. In some cases, where an ERS is used as an Ethernet local loop for L3 services such as Ethernet Internet Access (EIA), SPs have expressed interest in changing customer DSCP values (typically to zero (0)).
3. The CE-VLAN ID preservation attribute can be achieved for ERS/ERMS services with the use of the 1:1 VLAN Mapping feature.
4. Same as above.
5. CE-VLAN CoS preservation for ERS/ERMS (that is, when at most only a single 802.1Q tag is present) is only possible if: a) SP employs a restricted and direct mapping from authorized CE-VLAN CoS values to SP-VLAN CoS. b) SP directly maps ingress CE-VLAN CoS to MPLS EXP in cases where the UNI resides at the MPLS PE device.
6. Same as above.
7. Assumes that the unicast traffic conforms to the service policy.
8. The <EVC, CoS> CoS identifier for ERS/ERMS is a new capability for the ME solution.
9. Same as above.

10. The <EVC, CoS> CoS identifier for EWS/EMS is a new capability for ME solution. This requires CE-VLAN CoS inspection to derive the SP-VLAN CoS value.
11. Same as above.

UNI Service Attributes

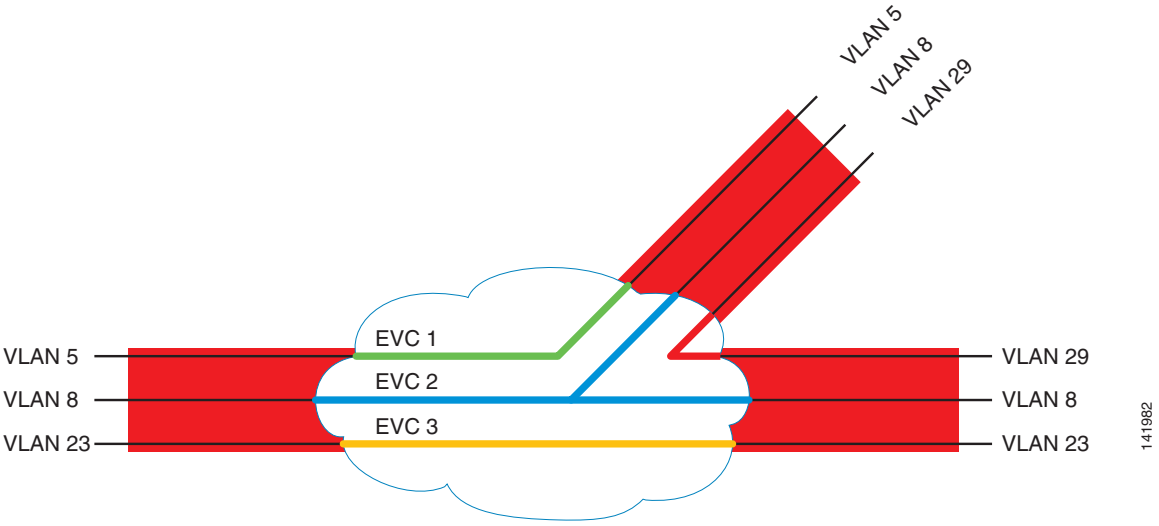
A UNI can have a number of characteristics that influence the way the Customer Edge (CE) device sees a service. The UNI service attributes are as follows:

- UNI identifier—Value that is assigned to the UNI by the SP that *may* have any string as a value and *must* be unique among all UNIs for the Metro Ethernet network (MEN).
- Physical medium—Specifies the physical interface as defined by the IEEE 802.3-2002 standard. Examples of physical media include 10BaseT, 100BaseT, 100BaseFX, and 1000BaseT.
- Speed—Specifies the standard Ethernet speeds of 10 Mbps, 100 Mbps, 1 Gbps, and 10 Gbps.
- Mode—Specifies whether the UNI supports full, half duplex, or auto speed negotiation.
- MAC layer— The UNI must support the IEEE 802.3-2002 frame formats.
- UNI EVC ID—Arbitrary string administered by the SP that is used to identify an EVC at the UNI.
- CE-VLAN ID/EVC map—For an UNI, there must be a recorded mapping of each CE-VLAN ID to at most one EVC called the CE-VLAN ID/EVC map.

Note In some scenarios, it may be necessary for the subscriber and the SP to agree upon the CE-VLAN ID/EVC map at the UNI. One way to implement this is to have the SP dictate the mapping. This is what is frequently done with the mapping between DLCIs and PVCs for Frame Relay.

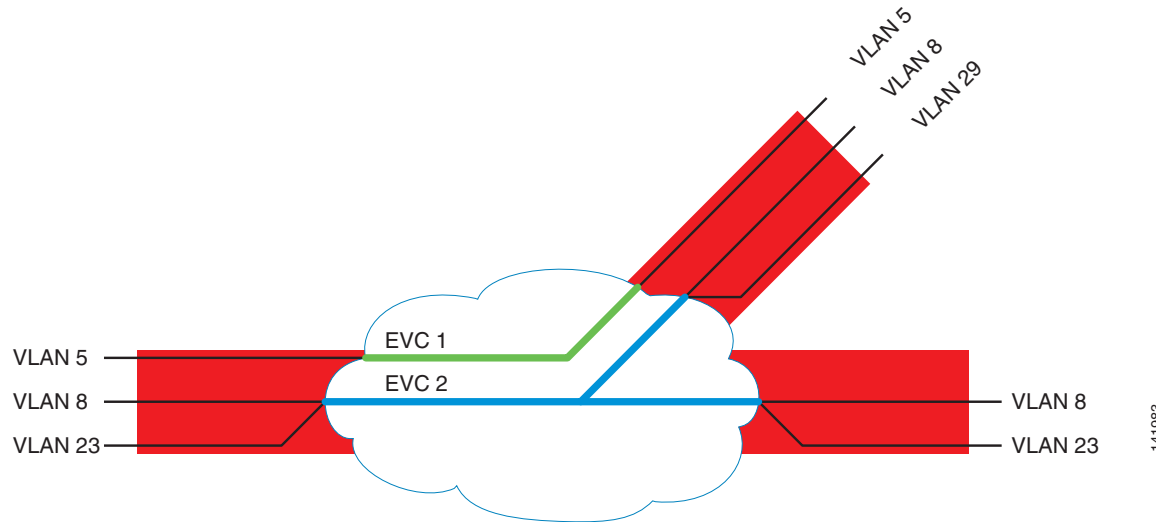
- Maximum number of EVCs—An integer greater than or equal to one (1)
- Service multiplexing—A UNI with the service multiplexing attribute must be able to support multiple EVCs (see Figure 6-5). Point-to-point and multipoint-to-multipoint EVCs may be multiplexed in any combination at the UNI. Following is the relationship of this attribute with others:
  - Service multiplexing *must* be “No” if all-to-one bundling is “Yes” (see Table 6-5 for more details).

Figure 6-5 Service Multiplexed UNIs that Support Multiple EVCs



- Bundling—When a UNI has the bundling attribute, it must be configurable so that two or more CE-VLAN IDs can map to an EVC at the UNI (see Figure 6-6). Following is the relationship of this attribute with others:
  - If an UNI supports bundling, the EVC *must* have the CE-VLAN ID preservation EVC attribute, and the list of CE-VLAN IDs mapped to the EVC *must* be the same at each UNI in the EVC.
  - Bundling *must* be "No" if all-to-one bundling is "Yes" (see Table 6-5 for more details).

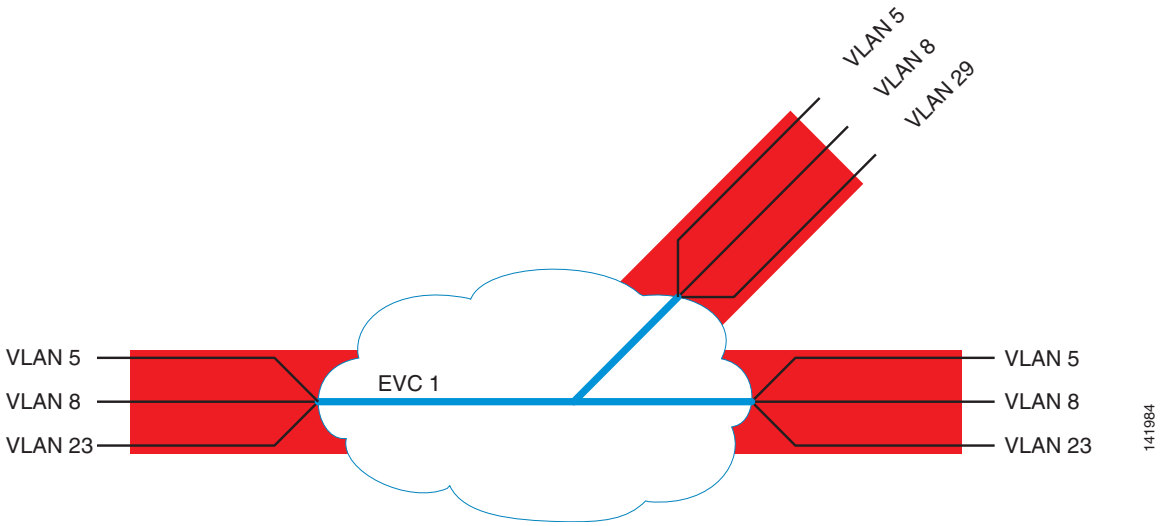
Figure 6-6 Bundling Attribute on a UNI



- All-to-one bundling—When a UNI has the all-to-one bundling attribute, all CE-VLANs *must* map to a single EVC at the UNI (see Figure 6-7). Following is the relationship of this attribute with others:
  - If an UNI supports all-to-one bundling, the EVC *must* have the CE-VLAN ID preservation service attribute, and the list of CE-VLAN IDs mapped to the EVC *must* be the same at each UNI in the EVC.

– All-to-one bundling *must* be “No” if bundling or service multiplexing is “Yes” (see Table 6-5 for more details).

Figure 6-7 All-to-One Bundling UNI Attribute



- Bandwidth profile attributes—A bandwidth profile is a characterization of the lengths and arrival times for ingress service frames at the UNI. When a bandwidth profile is applied to a given sequence of ingress frames, each service frame in the sequence is declared to be compliant or not compliant with the bandwidth profile. It also includes a specification of the disposition of ingress frames that do not comply with the profile. In this regard, only discard and marking the frame for priority discard are currently defined. The MEF has defined the following three bandwidth profile service attributes:
  - Ingress bandwidth profile per UNI—A single bandwidth profile must be applied to all ingress service frames at the UNI.
  - Ingress bandwidth profile per EVC—A single bandwidth profile must be applied to all ingress service frames for an instance of an EVC at the UNI.
  - Ingress bandwidth profile per CoS identifier—A single bandwidth profile must be applied to all ingress frames with a specific CoS identifier.

Each bandwidth profile consists of the following parameters:

- Committed information rate (CIR)
- Committed burst size (CBS)
- Peak information rate (PIR)
- Peak burst size (PBS)

Note Multiple bandwidth profile applications may exist simultaneously at a UNI. However, a UNI must be configured such that only a single bandwidth profile applies to any given service frame.

- Layer 2 control processing—Can be applied at the UNI or per EVC, and describes how to treat incoming CE Layer 2 control protocols. The allowable options are peer (process), discard, or pass them to the EVC (tunnel).

Table 6-4 provides a summary of the MEF UNI attributes as defined generically in MEF 10, Ethernet Services Attributes, Phase 1 standard.

.)

Table 6-4 Table Summary of MEF UNI Attributes

Attribute	Type of Parameter Value
UNI identifier	Any string
Physical medium	A standard Ethernet PHY[1]
Speed	10 Mbps, 100 Mbps, 1 Gbps, or 10 Gbps
Mode	Full Duplex or Auto negotiation
MAC Layer	802.3–2002
UNI EVC ID	An arbitrary string for the EVC supporting the service instance
CE-VLAN ID/EVC map	Mapping table of CE-VLAN IDs to EVCs at the UNI
Maximum number of EVCs	An integer greater than or equal to 1
Service multiplexing	Yes or no
Bundling	Yes or no
All-to-one bundling	Yes or no
Ingress bandwidth profile per ingress UNI	No or <CIR, CBS, PIR, PBS>
Ingress bandwidth profile per EVC	No or <CIR, CBS, PIR, PBS>
Ingress bandwidth profile per Class of Service identifier	No or <CIR, CBS, PIR, PBS>
Layer 2 Control Protocols Processing[2]	
Bridge block of protocols: 0x0180.C200.0000 through 0x0180.C200.000F	Discard, peer, or pass to EVC
GARP block or protocols: 0x0180.C200.0020 through 0x0180.C200.002F	Discard, peer, or pass to EVC
All bridges protocol 0x0180.C200.0010	Discard, peer, or pass to EVC

1. Per IEEE P 802.3-2002.
2. Note that SPs may define additional addresses as Layer 2 control in addition to those listed here.

Relationship between Service Multiplexing, Bundling, and All-to-One Bundling

Table 6-5 shows the valid combinations for three of the most relevant UNI attributes highlighted in the previous section. Some are mutually exclusive and therefore only some combinations are allowed. For example, if a UNI exhibits the all-to-one bundling attribute, service multiplexing and bundling *must* not be present.

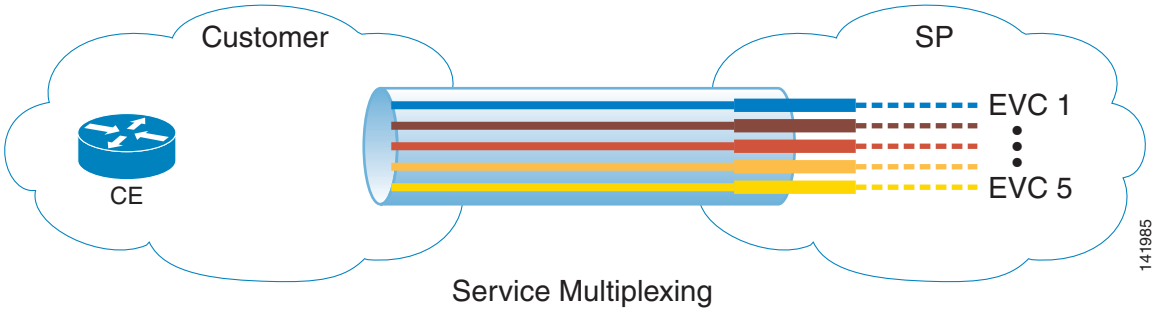
Table 6-5 UNI Attribute Valid Combinations

UNI Attribute	Valid Combinations		
	Option 1	Option 2	Option 3
Service multiplexing	Yes	Yes	No
Bundling	No	Yes	No
All-to-one bundling	No	No	Yes

Figure 6-8 through Figure 6-10 support the content of the previous table and show three UNIs with the allowed attribute combination. Observe that in these examples, UNIs are receiving service frames from five (5) CE-VLAN IDs.

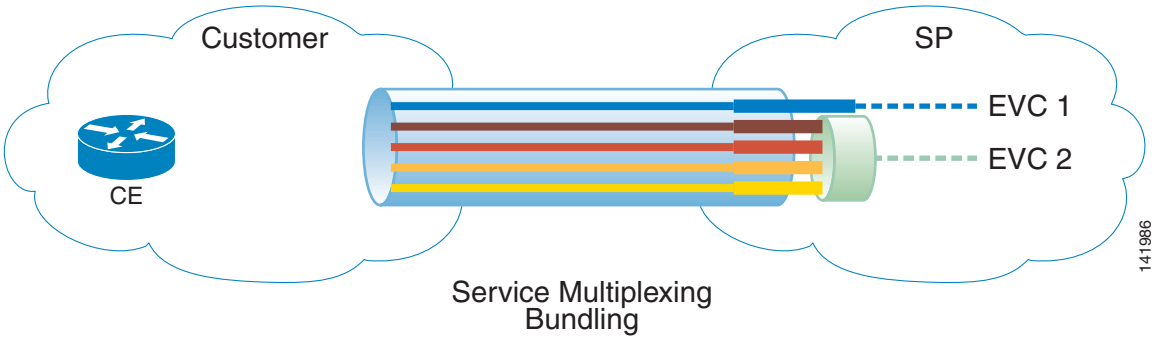
In the first scenario, each CE-VLAN ID is mapped to one EVC for a total of five (5) EVCs at the UNI (also known as one-to-one mapping). This UNI only has the service multiplexing attribute.

Figure 6-8 Option 1—UNI with Service Multiplexing Attribute



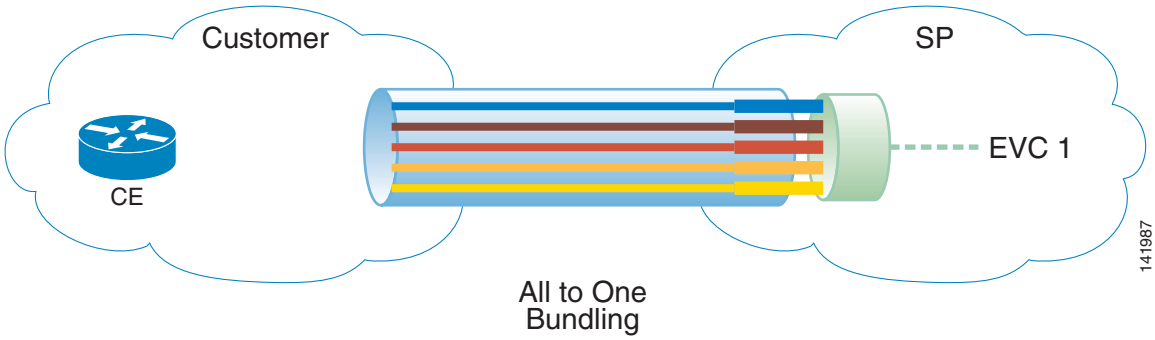
In the following example, (UNI with bundling and service multiplexing attributes), the first CE-VLAN ID is mapped to one EVC and the remaining four (4) to a second EVC. As seen, this UNI contains only two (2) EVCs.

Figure 6-9 Option 2—UNI with Bundling and Service Multiplexing Attributes



Lastly, the last UNI highlights the case where all CE-VLAN IDs are mapped to just one EVC. In this case, the UNI has the all-to-one bundling attribute.

Figure 6-10 Option 3—UNI with All-to-One Bundling Attribute



ME UNI Service Attributes

Table 6-6 summarizes the UNI service attributes for each of the ME services. Note that not all of the MEF attributes are listed in this table (attributes used for record-keeping/inventory purposes have been omitted). Also, the table expands the Layer 2 control processing section from the one included in MEF 10.

Table 6-6 ME UNI Service Attributes

	ME Services				
UNI Service Attribute	**ERS**	**ERMS**	**EWS**	**EMS**	**EPL**
Speed	10/100/1000 Mbps				
MAC layer	IEEE 802.3-2002				
Service multiplexing	Yes	Yes	Yes[1] or No	Yes[2] or No	No
Bundling	No[3]	No[4]	Yes[5] or No	Yes[6] or No	No
All-to-one bundling	No	No	No or Yes	No or Yes	Yes
Maximum number of EVCs	>=1	>=1	>=1[7]	>=1[8]	== 1
Ingress bandwidth profile per UNI	No or <CIR, CBS, EIR, EBS>[9]				No or: CIR > 0, CBS > largest frame size PIR == 0, PBS == 0
Ingress bandwidth profile per EVC	No or <CIR, CBS, EIR, EBS>[10]				n/a
Ingress and egress bandwidth profile per CoS identifier	No or <CIR, CBS, EIR, EBS>[11]				n/a
Layer 2 Control Protocol Processing					
802.3x handling	Discard	Discard	Discard	Discard	Discard
LACP handling	Discard	Discard	Pass to EVC	Discard	Pass to EVC

Table 6-6 ME UNI Service Attributes (continued)

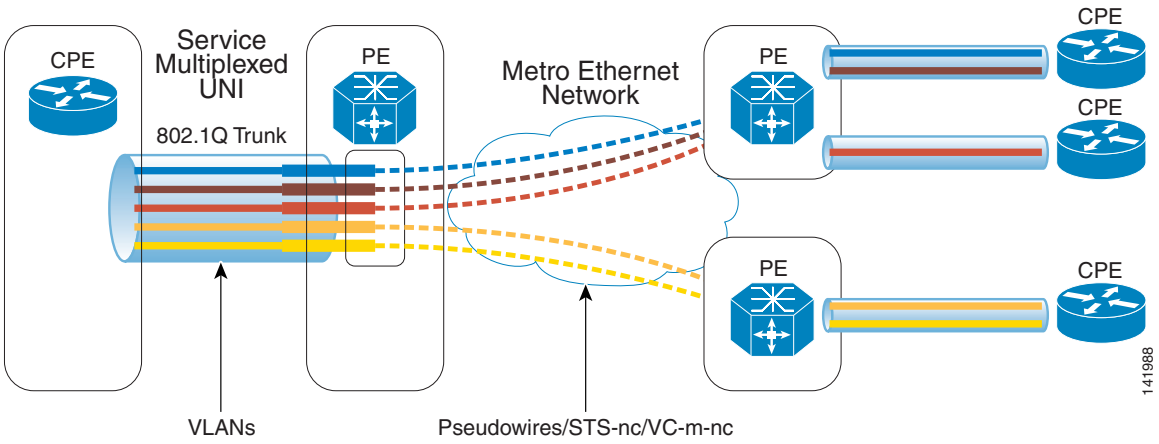
802.1x handling	Discard	Discard	Discard	Discard	Pass to EVC
GARP handling	Discard	Discard	Discard	Discard	Pass to EVC
STP handling	Discard	Discard	Pass to EVC	Pass to EVC	Pass to EVC
Protocol that multicasts to all bridges in a bridged LAN	Discard	Discard	Discard	Discard	Pass to EVC
Layer 2 Protocols not listed in MEF Framework					
CDP handling	Discard	Discard	Pass to EVC	Pass to EVC	Pass to EVC
VTP handling	Discard	Discard	Pass to EVC	Pass to EVC	Pass to EVC
PAgP handling	Discard	Discard	Pass to EVC	Discard	Pass to EVC
UDLD handling	Discard	Discard	Pass to EVC	Discard	Pass to EVC

1. Service multiplexing on a UNI with an EWS/EMS service is *optional* and is achieved when the bundling UNI attribute is present. In this case, the all-to-one bundling attribute *must* be No.
2. Same as above.
3. ERS/ERMS services are defined with a one-to-one relationship for the CE-VLAN ID/EVC map attribute (that is, one EVC maps to no more than one CE-VLAN ID). Therefore, the UNI bundling attribute *may* exist at the UNI but is not associated with the corresponding EVC for the mentioned services.
4. Same as above.
5. Bundling *may* be present on a UNI with an EWS/EMS service. If present, the all-to-one bundling attribute *must* be No.
6. Same as above.
7. With the presence of the bundling attribute at the UNI, it is possible to have more than one service in a UNI that holds an EWS/EMS service.
8. Same as above.
9. Ingress BW profile per UNI is mostly applied on cases where the UNI holds a single EVC. The ability to support CIR/PIR depends mostly on the U-PE capabilities. Certain services (for example, multipoint) might be offered with CIR == 0.
10. Ability to support CIR/PIR depends mostly on the U-PE capabilities. Certain services (or example, multipoint) might be offered with CIR == 0 .
11. Ability to support CIR/PIR and <EVC, CoS> or <EVC, DSCP> CoS IDs depends mostly on the U-PE capabilities. Certain services (or example, multipoint)) might be offered with CIR == 0.

Ethernet Relay Service

ERS is an Ethernet point-to-point VLAN-based service targeted to Layer 3 CEs (routers). Among its applications, ERS represents a high-speed alternative to existing Frame Relay and ATM offerings. For example, the VLANs in Figure 6-11 can be considered equivalent to DLCIs in FR circuits that carry the traffic from a corporate office to the regional office sites.

Figure 6-11 ERS Deployment Scenario—Multiple ERS Multiplexed Over a Single UNI Interface



Note With ERS, the SP assigns unique VLAN IDs to their customers as they would do for Frame Relay DLCIs. Using the new VLAN 1:1 translation feature, the SP may accommodate customer requests for specific VLAN ID values. See "VLAN Translation Analysis" (EDCS-447318) for details.

The ERS UNI is typically an 802.1Q trunk (or access port) that allows the SP to multiplex services on a single interface. This gives the SP the capability to direct customer traffic to different destination routers using the appropriate VLAN IDs.

When ERS is implemented over an MPLS core, there is a one-to-one mapping between 802.1Q VLAN IDs and EoMPLS pseudowires.

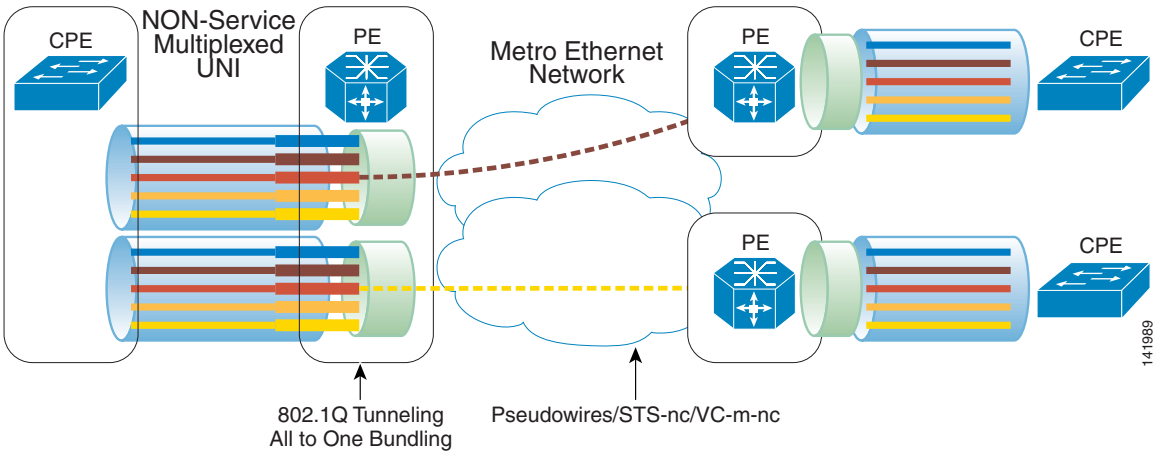
As mentioned earlier, the intended termination device of this service is a Layer 3 CE. Thus, ERS purposely does not tunnel Layer 2 control PDUs (for example, STP BPDUs, VTP) typically exchanged by Layer 2 CEs (bridges). With the selection of a Layer 3 CE, the SP reduces the number of MAC addresses that need to be learned by the network (that is, only two MACs per VLAN for a point-to-point service).

Note SP VLAN IDs can be different on each side of the EVC, thereby permitting a more scalable and flexible use of the VLAN ID space.

Ethernet Wire Service

EWS is an Ethernet point-to-point port-based service targeted to Layer 2 CEs (bridges). Among its main applications, this service can be used to connect geographically remote LANs over an SP network.

Figure 6-12 EWS Deployment Scenario—Two Services Provisioned Over the SP Network



When implemented on Cisco Catalyst switches, the EWS UNI is an 802.1Q tunnel (or QinQ) interface, which allows the SP to tunnel any number of CE-VLANs through the SP network. This is accomplished by encapsulating customer traffic inside a pre-defined SP VLAN, thus allowing overlapping customer VLAN IDs. In MEF terms, a UNI of these characteristics is described as supporting the all-to-one bundling UNI attribute.

Note that an 802.1Q tunnel increases the supported number of customer VLANs. Therefore, it is possible to support 4094 customers per Metro access domain, where each UNI could potentially receive up to 4094 VLANs per customer.

Because the service is intended for Layer 2 CEs, VLAN transparency and Layer 2 PDU transparency are key characteristics provided by it. One of the ways to achieve VLAN transparency is with the QnQ behavior described in the previous paragraph. Secondly, Layer 2 PDU (for example, STP, VTP) transparency can be achieved with the use of features such as Layer 2 Protocol Tunneling (L2PT), which effectively makes the remote LANs appear as if they were on the same segment (from a control plane perspective). An example of an EWS is shown in Figure 6-12.

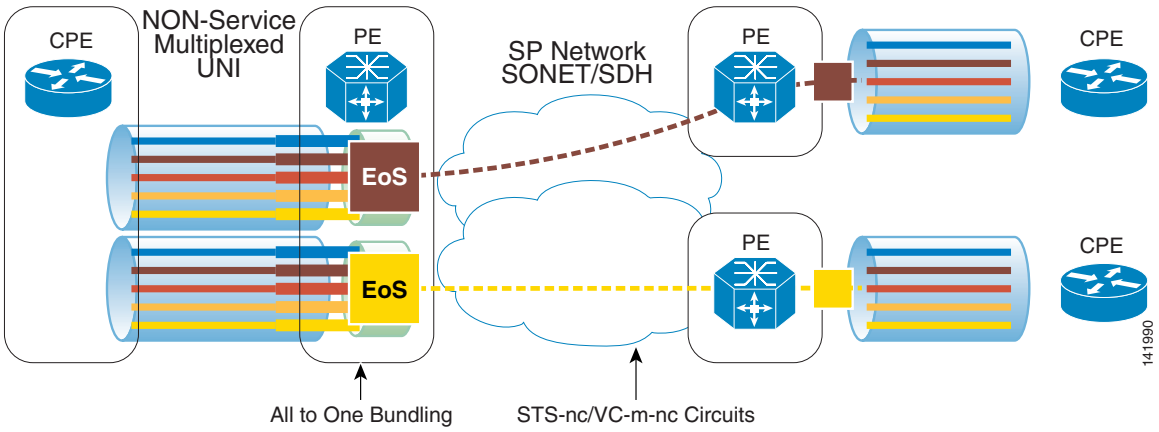
Note SP VLAN IDs can be different on each side of the EVC, thereby permitting a more scalable and flexible use of the VLAN ID space.

Ethernet Private Line

EPL is an Ethernet point-to-point, port-based service that maps customer Layer 2 traffic directly onto a TDM circuit. It is considered by US-based SP transport/transmission groups as the alternative to offer a "private" service. With an EPL, the customer Ethernet stream is encapsulated directly into the SONET or SDH frame and mapped exclusively onto an STS or VC circuit. From a service attribute perspective, EPL is a VLAN and L2PDU transparent service that supports all-to-one bundling, but not service multiplexing.

Figure 6-13 illustrates a sample EPL service offering.

Figure 6-13 EPL Deployment Scenario



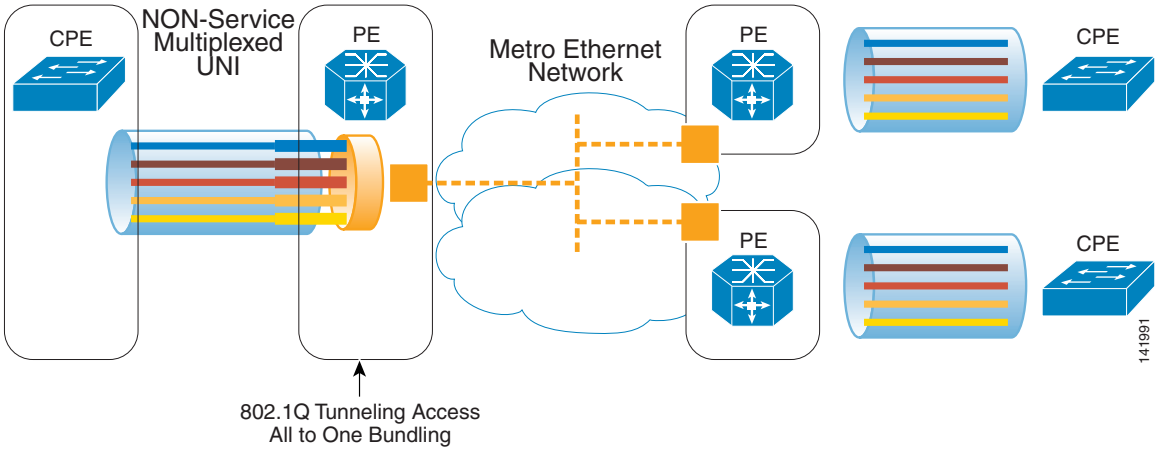
Ethernet Multipoint Service

EMS is an Ethernet multipoint-to-multipoint port-based service targeted to Layer 2 CEs (bridges). It is used primarily for transparent LAN service applications.

For the most part, EMS and EWS share the same service characteristics. Their main difference is that EMS is a multipoint-to-multipoint service. See Ethernet Wire Service, page 6-15 for a basic description of this service.

When implemented over MPLS, the SP VLAN is mapped to a virtual private LAN service (VPLS) forwarding instance (VFI). Figure 6-14 illustrates a sample EMS.

Figure 6-14 EMS Deployment Scenario for Three Customer Sites



ME EMS Enhancement

EMS enjoys the same service enhancements as EWS.

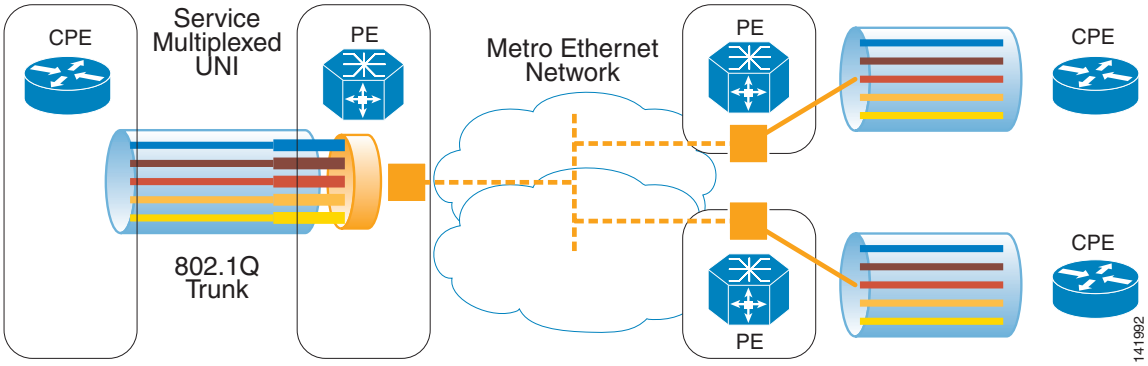
Ethernet Relay Multipoint Service

ERMS is an Ethernet multipoint-to-multipoint VLAN-based service targeted to Layer 3 CEs (routers). It is intended for scenarios where customers desire a multipoint-to-multipoint connection among WAN routers.

For the most part, ERMS and ERS share the same service characteristics. Their main difference is that ERMS is a multipoint-to-multipoint service. See Ethernet Relay Service, page 6-14 for a basic description of this service.

When implemented over MPLS, the SP VLAN is mapped to a virtual private LAN service (VPLS) VFI. Figure 6-15 illustrates a sample ERMS.

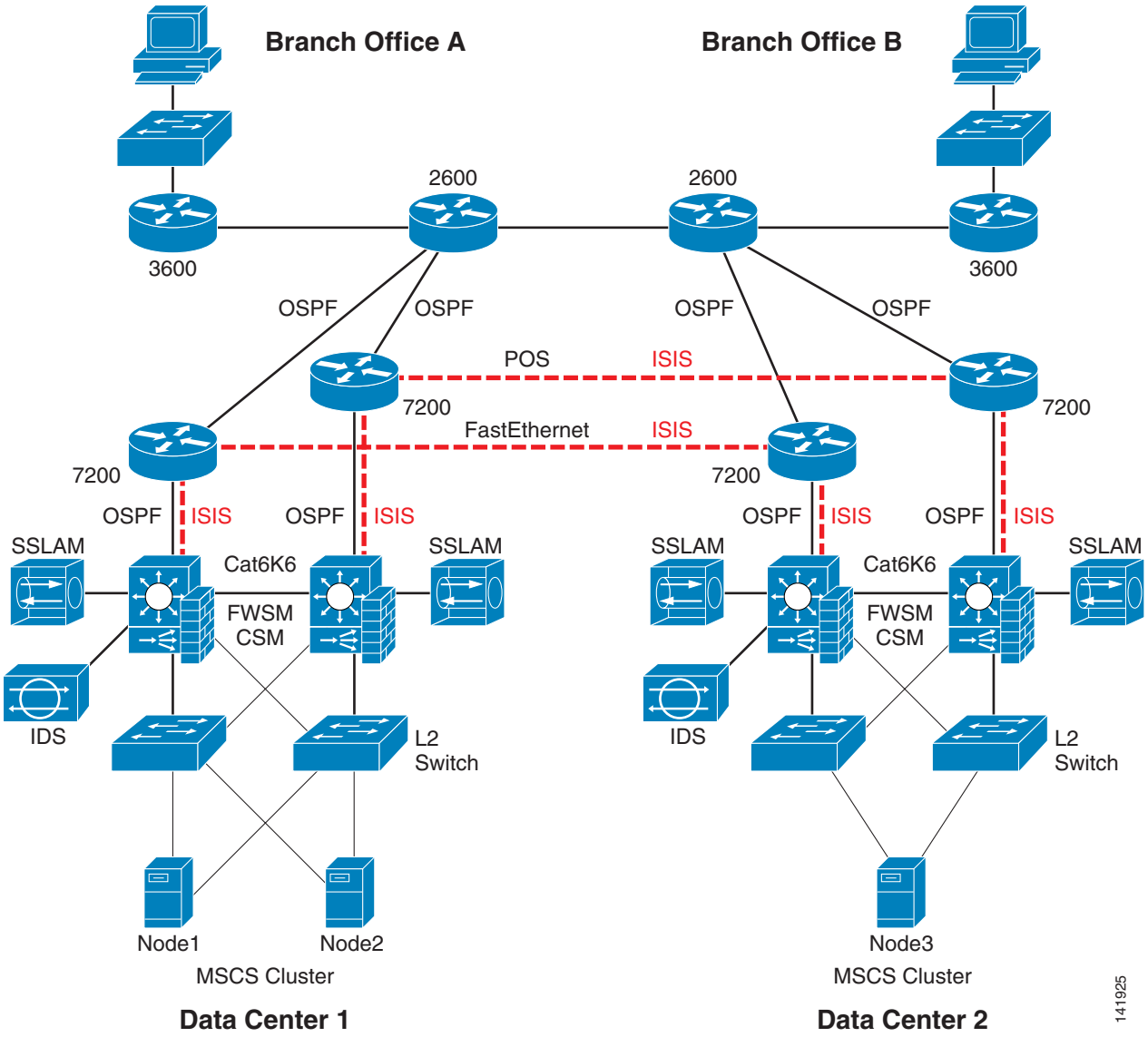
Figure 6-15 ERMS Deployment Scenario



Configurations for Layer 2 Extension with EoMPLS

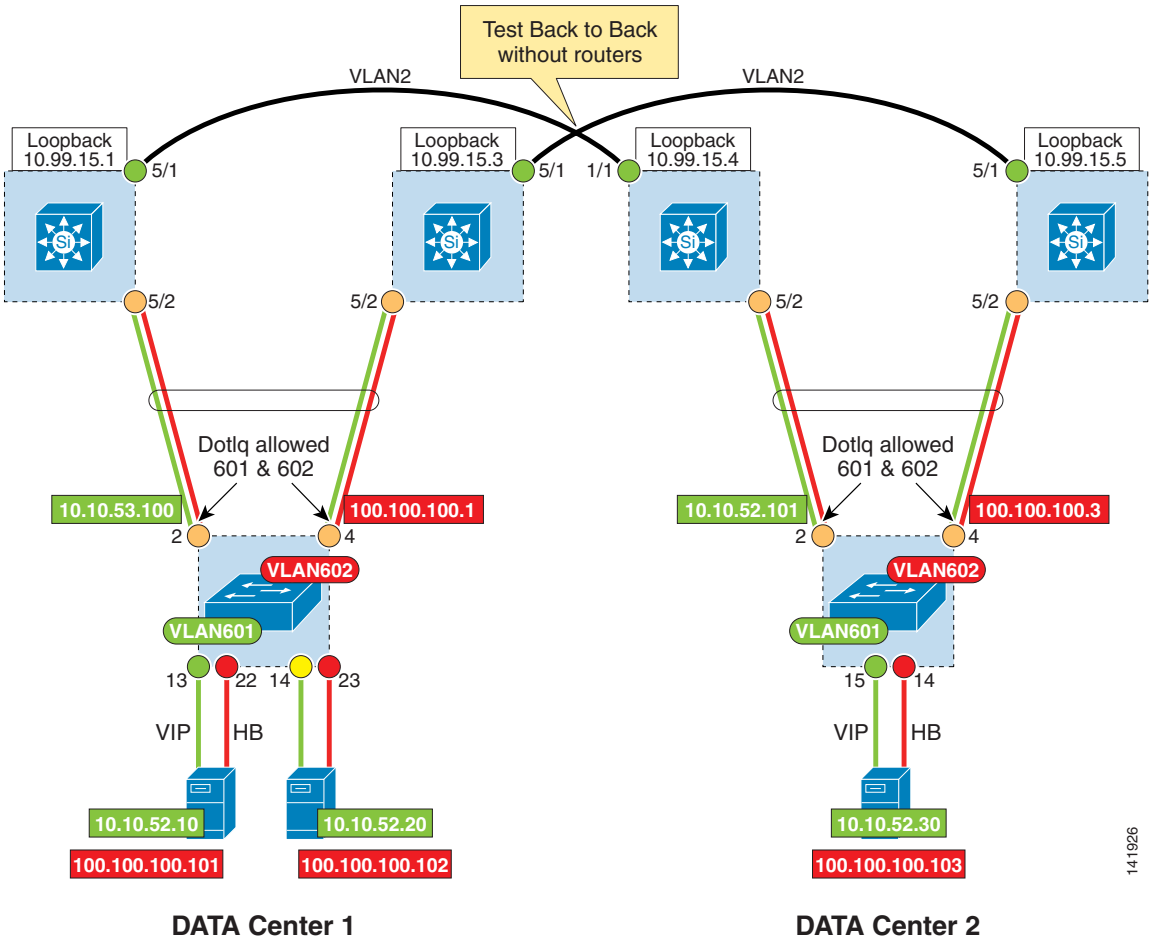
The tested architecture is based on two fully redundant data centers (DC1 and DC2) and two remote branch offices, all connected to a WAN access network built with 2 x 2600 routers. An MPLS network is built between the two data centers with the WAN access router (7200), as shown in Figure A-1.

Figure A-1 Corporate Data Centers and Branch Offices



For this test setup, a Microsoft MSCS cluster is built with three nodes, two nodes located on DC1 and a third node located on DC2. The cluster requires two distinct Layer 2 networks between each node. Each node is configured with two NICs, one used for the cluster heartbeat and one dedicated for the VIP (user access). Each interface is connected to a dedicated VLAN: VIP belongs to VLAN 601 and HB (heartbeat) belongs to VLAN 602.

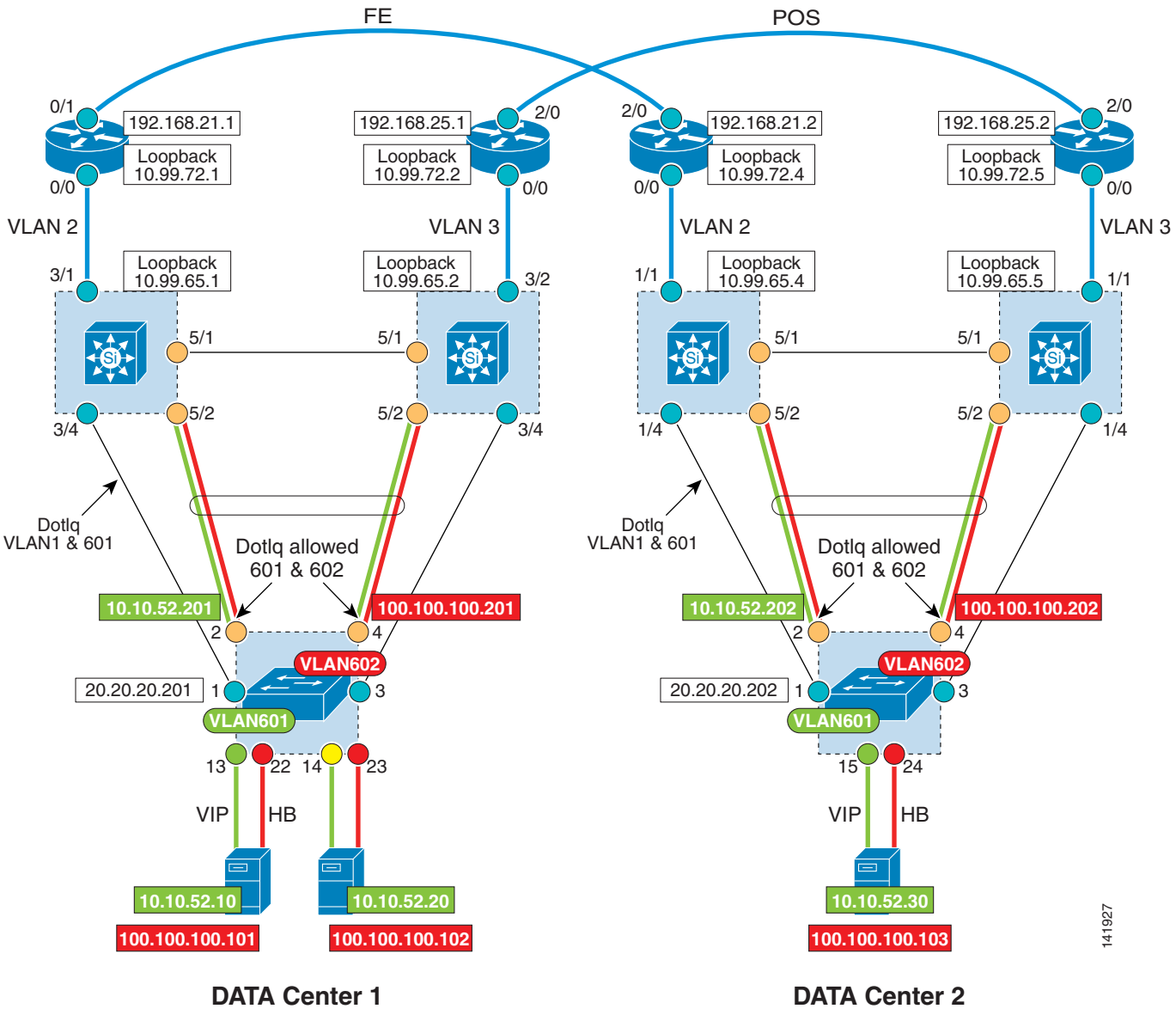
Figure A-2 Layer 2 Back-to-Back Testing



With Option 3, the Layer 2 VPN tunnel built with the Port-based Xconnect is initiated directly at the interface that connects to the access switch. With this design, the EoMPLS tunnel (pseudowire) is transparent to the access switch (see Figure A-3). Therefore, the EtherChannel feature might be useful to deploy, as an alternative to spanning tree.

However, deploying EoMPLS in native IOS (with no additional OSM or SIP card) does not allow switching the internal VLAN coming from the access switch to outside the Layer 2 tunnel. Therefore, an additional parallel physical link must be provisioned from the access switch to the aggregation switch to allow required VLANs to be routed outside the data center, such as the VLAN 601 used for the VIP of the cluster, or, in these tests, the VLAN 1 used for management purposes.

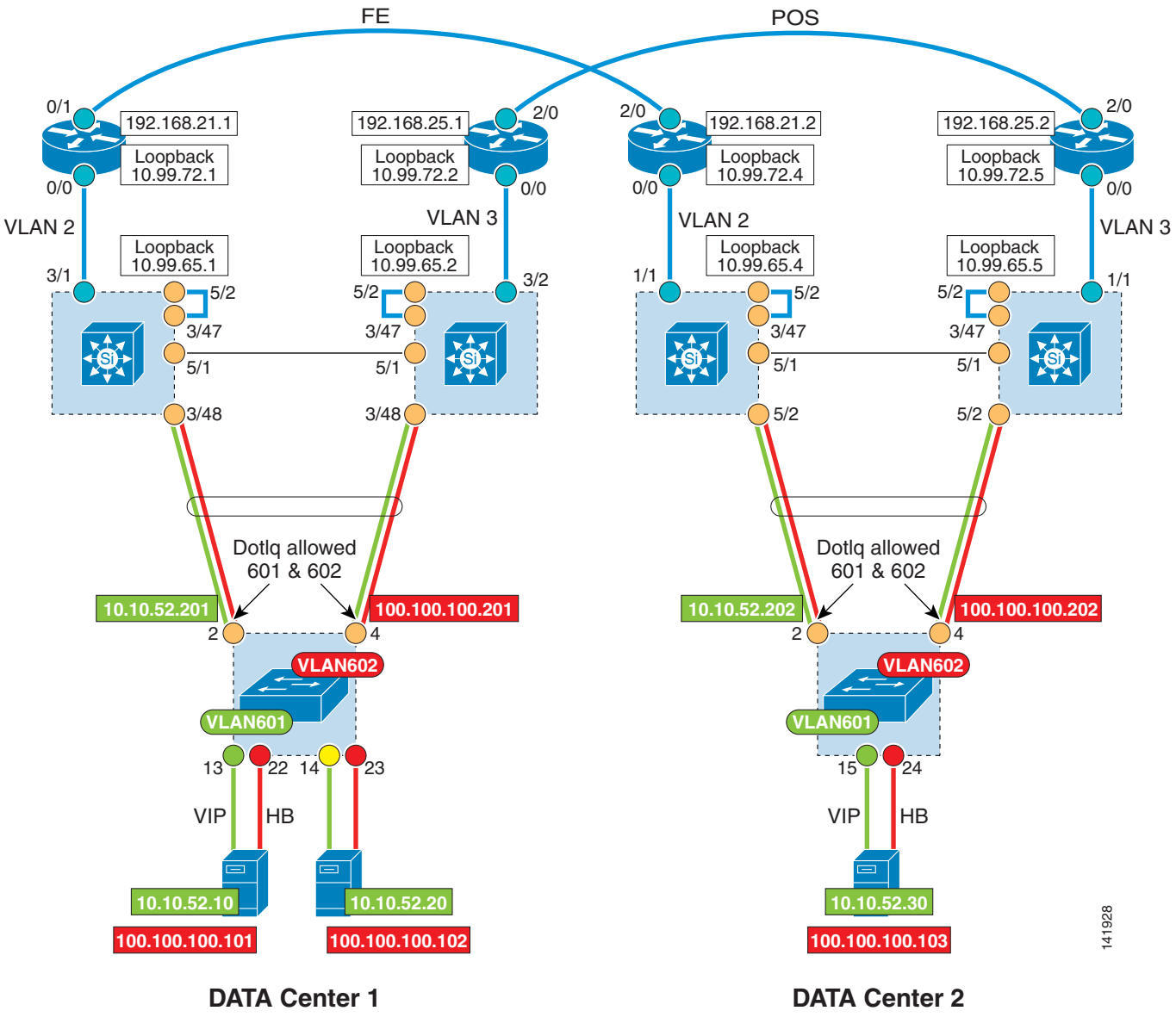
Figure A-3 Port-Based Xconnect with with Multiple Links from the Access Switches



As described in the previous configuration, Option 3 requires two parallel links between the aggregation and the access layers to be able to route the VLAN outside the pseudowires.

An alternative design to using additional links from the access switches consists in using a loopback cable, also referred to as Option 4 in this guide (see Figure A-4.)

Figure A-4 Port-Based Xconnect Using Loopback Cable



To use the loopback cable, a full trunk is created between the access switch and the aggregation switch on both interfaces. As you recall, with Option 3, only the interface on the access switch is using the mode trunk, the interface on the aggregation switch is configured in access mode for the pseudowire.

On this test, the interfaces n/48 are configured using the mode trunk allowing VLAN 601 (VIP), VLAN 602 (heartbeat) and VLAN 1 (management). Therefore, any of these VLANs can be routed by the MSFC as any traditional SVI.

For the pseudowire, an additional interface (n/47) is created in trunk mode and allowed the VLAN 601 and 602 for the Layer 2 tunnel. A loopback cable was added to interconnect interface n/47 to an interface used for the Port-based Xconnect interface (pseudowire) G5/2.

Configurations

Enabling MPLS

```
!
mpls label protocol ldp
mpls ldp router-id Loopback99
mpls ldp neighbor 10.99.65.2 targeted ldp    ! to maintain label binding even on link
failure
mpls ldp neighbor 10.99.72.1 targeted ldp
no mpls ldp advertise-labels          ! to limit advertisement of label to EoMPLS
loopbacks
mpls advertise-tags for 1             !  …
access-list 1 permit 10.99.0.0 0.0.255.255   !  …
!
```

Port-based Xconnect

Interface G5/2 cannot be configured as a switchport. As it receives tagged frames (+4 bytes), the MTU must be increased to 1504 (4 Bytes dot1Q). The interface xconnect to the remote loopback uses encapsulation MPLS.

```
!
interface GigabitEthernet5/2
 mtu 1504
 no ip address
 load-interval 30
 media-type rj45
 xconnect 10.99.65.4 100 encapsulation mpls
!
```

Configuring the Loopback Interface

```
!
interface Loopback99
 ip address 10.99.65.1 255.255.255.255
 ip router isis
 isis circuit-type level-1
!
```

Configure VLAN 2 to interconnect both data centers. The same VLAN ID is used for both Catalyst 6000 Series switches. ISIS is configured with MPLS. The minimum MTU size for VLAN 2 must be set to 1522 (802.3 max frame size 1518 + 4 bytes for tagged frames). The uplinks from the access switch carry multiple VLANs to the aggregation switch where the uplink interface is configured in access mode. This means that any tagged frames from the access switch are seen as a raw larger frames.

```
!
interface Vlan2
 mtu 1522
 ip address 10.0.0.2 255.255.255.252
 ip router isis
 tag-switching ip
 tag-switching mtu 1526
 isis circuit-type level-1
```

```
 isis hello-multiplier 10 level-1
 isis hello-interval minimal level-1
!
```

Configure the interface fa6/1 that belongs to VLAN 2 and connected to the remote Catalyst 6000 Series switch. Note that the MTU is forced to 9216.

```
!
 interface FastEthernet6/1
 switchport
 switchport access vlan 2
 switchport mode access
 mtu 9216
 no ip address
 load-interval 30
 spanning-tree portfast
!
```

Configuring OSPF

```
!
router ospf 1
 log-adjacency-changes
 redistribute static subnets
 network 10.0.0.0 0.0.0.3 area 0
 network 10.0.2.0 0.0.0.255 area 0
 network 10.0.10.0 0.0.0.255 area 0
 network 10.0.20.0 0.0.0.255 area 0
 network 10.0.30.0 0.0.0.255 area 0
 network 10.0.40.0 0.0.0.255 area 0
 network 10.0.51.0 0.0.0.255 area 0
!
```

Configuring ISIS

```
!
router isis
 net 49.0001.0000.6500.1111.00
 is-type level-1
 metric-style wide
passive-interface Loopback99
 advertise passive-only
 spf-interval 20 100 20
 prc-interval 20 100 20
 lsg-gen-interval 1 1 20
 fast-flood 15
!
```

Aggregation Switch Right (Catalyst 6000 Series Switch-Sup720-B)—Data Center 1

Enabling MPLS

```
!
mpls label protocol ldp
mpls ldp router-id Loopback99
mpls ldp neighbor 10.99.65.1 targeted ldp
mpls ldp neighbor 10.99.72.2 targeted ldp
no mpls ldp advertise-labels
mpls advertise-tags for 1
access-list 1 permit 10.99.0.0 0.0.255.255
!
```

Port-based Xconnect

```
!
interface GigabitEthernet5/2
 no ip address
mtu 1504
 load-interval 30
 media-type rj45
 xconnect 10.99.65.5 100 encapsulation mpls!
!
```

Configuring the Loopback Interface

```
!
interface Loopback99
 ip address 10.99.65.2 255.255.255.255
 ip router isis
 isis circuit-type level-1
!
```

Configuring VLAN 2

```
!
interface Vlan2
mtu 1522
 ip address 10.10.0.2 255.255.255.252
 ip router isis
 tag-switching ip
 tag-switching mtu 1526
 isis circuit-type level-1
 isis hello-multiplier 10 level-1
 isis hello-interval minimal level-1
!
```

Configuring Interface fa5/1 (Connected to a Remote Catalyst 6000 Series Switch)

```
!
interface GigabitEthernet5/1
mtu 9216
 switchport
 switchport access vlan 2
```

```
 switchport mode access
 no ip address
 load-interval 30
 spanning-tree portfast
!
```

Configuring OSPF

```
!
router ospf 1
 log-adjacency-changes
 redistribute static subnets
 network 10.0.0.0 0.0.0.3 area 0
 network 10.0.2.0 0.0.0.255 area 0
 network 10.0.10.0 0.0.0.255 area 0
 network 10.0.20.0 0.0.0.255 area 0
 network 10.0.30.0 0.0.0.255 area 0
 network 10.0.40.0 0.0.0.255 area 0
!
```

Configuring ISIS

```
!
router isis
 net 49.0001.0000.6500.2222.00
 is-type level-1
 metric-style wide
passive-interface Loopback99
 advertise passive-only
 spf-interval 20 100 20
 prc-interval 20 100 20
 lsg-gen-interval 1 1 20
 fast-flood 15
!
```

Aggregation Switch Left (Catalyst 6000 Series Switch-Sup720-B)—Data Center 2

Enabling MPLS

```
!
mpls label protocol ldp
mpls ldp router-id Loopback99
mpls ldp neighbor 10.99.65.5 targeted ldp
mpls ldp neighbor 10.99.72.4 targeted ldp
no mpls ldp advertise-labels
mpls advertise-tags for 1
access-list 1 permit 10.99.0.0 0.0.255.255
!
```

Port-based Xconnect

Interface G5/2 cannot be configured as a switchport. As it receives tagged frames (+4 bytes), the MTU must be increased to 1504. The interface Xconnect to the remote loopback using encapsulation MPLS.

```
!
interface GigabitEthernet5/2
 description "to access switch Xconn"
 mtu 1504
 no ip address
 load-interval 30
 no mdix auto
 xconnect 10.99.65.1 100 encapsulation mpls
!
```

Configuring the Loopback Interface

```
!
interface Loopback99
 ip address 10.99.65.4 255.255.255.255
 ip router isis
 isis circuit-type level-1
!
```

Configure VLAN 2 to interconnect both data centers. The same VLAN ID is used for both Catalyst 6000 Series switches. ISIS is configured with MPLS.

```
!
interface Vlan2
 mtu 1522
 ip address 10.0.0.1 255.255.255.252
 ip router isis
 tag-switching ip
 tag-switching mtu 1526
 isis circuit-type level-1
 isis hello-multiplier 10 level-1
 isis hello-interval minimal level-1
!
```

Configure the interface fa1/1 that belongs to VLAN 2 and connected to the remote Catalyst 6000 Series switch. Note that the MTU is forced to 9216.

```
!
interface FastEthernet1/1
 description To-router-rack1
 switchport
 switchport access vlan 2
 switchport mode access
 mtu 9216
 no ip address
 load-interval 30
 spanning-tree portfast
 lan-name Router-L
!
```

Configuring OSPF

```
!
router ospf 1
 log-adjacency-changes
 redistribute static subnets
 network 10.10.0.0 0.0.0.3 area 0
network 10.10.2.0 0.0.0.255 area 0
 network 10.10.10.0 0.0.0.255 area 0
 network 10.10.20.0 0.0.0.255 area 0
 network 10.10.30.0 0.0.0.255 area 0
```

```
 network 10.10.50.0 0.0.0.255 area 0
!
```

Configuring ISIS

```
!
router isis
 net 49.0001.0000.6500.4444.00
 is-type level-1
 metric-style wide
passive-interface Loopback99
 advertise passive-only
 spf-interval 20 100 20
 prc-interval 20 100 20
 lsg-gen-interval 1 1 20
 fast-flood 15
!
```

Aggregation Switch Right (Catalyst 6000 Series Switch-Sup720-B)—Data Center 2

Enabling MPLS

```
!
mpls label protocol ldp
mpls ldp router-id Loopback99
mpls ldp neighbor 10.99.65.4 targeted ldp
mpls ldp neighbor 10.99.72.5 targeted ldp
no mpls ldp advertise-labels
mpls advertise-tags for 1
access-list 1 permit 10.99.0.0 0.0.255.255
!
```

Port-based Xconnect

```
!
interface GigabitEthernet5/2
description "to access switch Xconn"
mtu 1504
 no ip address
 load-interval 30
 media-type rj45
 xconnect 10.99.65.2 100 encapsulation mpls
 lan-name Cluster!
!
```

Configuring the Loopback Interface

```
!
interface Loopback99
 ip address 10.99.65.5 255.255.255.255
 ip router isis
 isis circuit-type level-1
!
```

Configuring VLAN 2

```
!
interface Vlan2
mtu 1522
 ip address 10.10.0.1 255.255.255.252
 ip router isis
 tag-switching ip
 tag-switching mtu 1526
 isis circuit-type level-1
 isis hello-multiplier 10 level-1
 isis hello-interval minimal level-1
!
```

Configuring Interface G5/1 (Connected to Remote Catalyst 6000 Series Switch)

```
!
interface GigabitEthernet5/1
mtu 9216
 switchport
 switchport access vlan 2
 switchport mode access
 no ip address
 load-interval 30
 spanning-tree portfast
 !
```

Configuring OSPF

```
!
router ospf 1
 log-adjacency-changes
 redistribute connected
 redistribute static subnets
 network 10.10.0.0 0.0.0.3 area 0
 network 10.10.2.0 0.0.0.255 area 0
 network 10.10.10.0 0.0.0.255 area 0
 network 10.10.20.0 0.0.0.255 area 0
 network 10.10.30.0 0.0.0.255 area 0
 network 10.10.50.0 0.0.0.255 area 0
!
```

Configuring ISIS

```
!
router isis
 net 49.0001.0000.6500.5555.00
 is-type level-1
 metric-style wide
passive-interface Loopback99
 advertise passive-only
 spf-interval 20 100 20
 prc-interval 20 100 20
 lsg-gen-interval 1 1 20
 fast-flood 15
!
```

MTU Considerations

The interface Xconnect receives tagged frames from all VLANs created at the access switches (601 and 602 in this example). Therefore, 4 bytes must be added to the Interface MTU (here Gig5/2).

The VLAN connecting the Catalyst 6000 Series switch to the edge router should support 1518 bytes (max Ethernet frame size) + 4 bytes, or 1522 bytes (VLAN 2 in this example).

Also, the physical Interface that connects to the remote router must use a bigger MTU. The minimum MTU size for the egress interface is 9216 (Int Gig5/1 in this example).

Spanning Tree Configuration

The design uses MST as Spanning Tree Protocol. Each data center is aggregated within a dedicated MST region. Therefore, between the two MST regions, RSTP is enabled to carry the BPDU for the Instance 0 (default). (See Figure A-5.)

This assumes the following:

- The root bridge for Instance 0 is located on the DC1 left Catalyst 6000 Series switch.
- The secondary root bridge for Instance 0 is located on the DC1 right Catalyst 6000 Series switch.

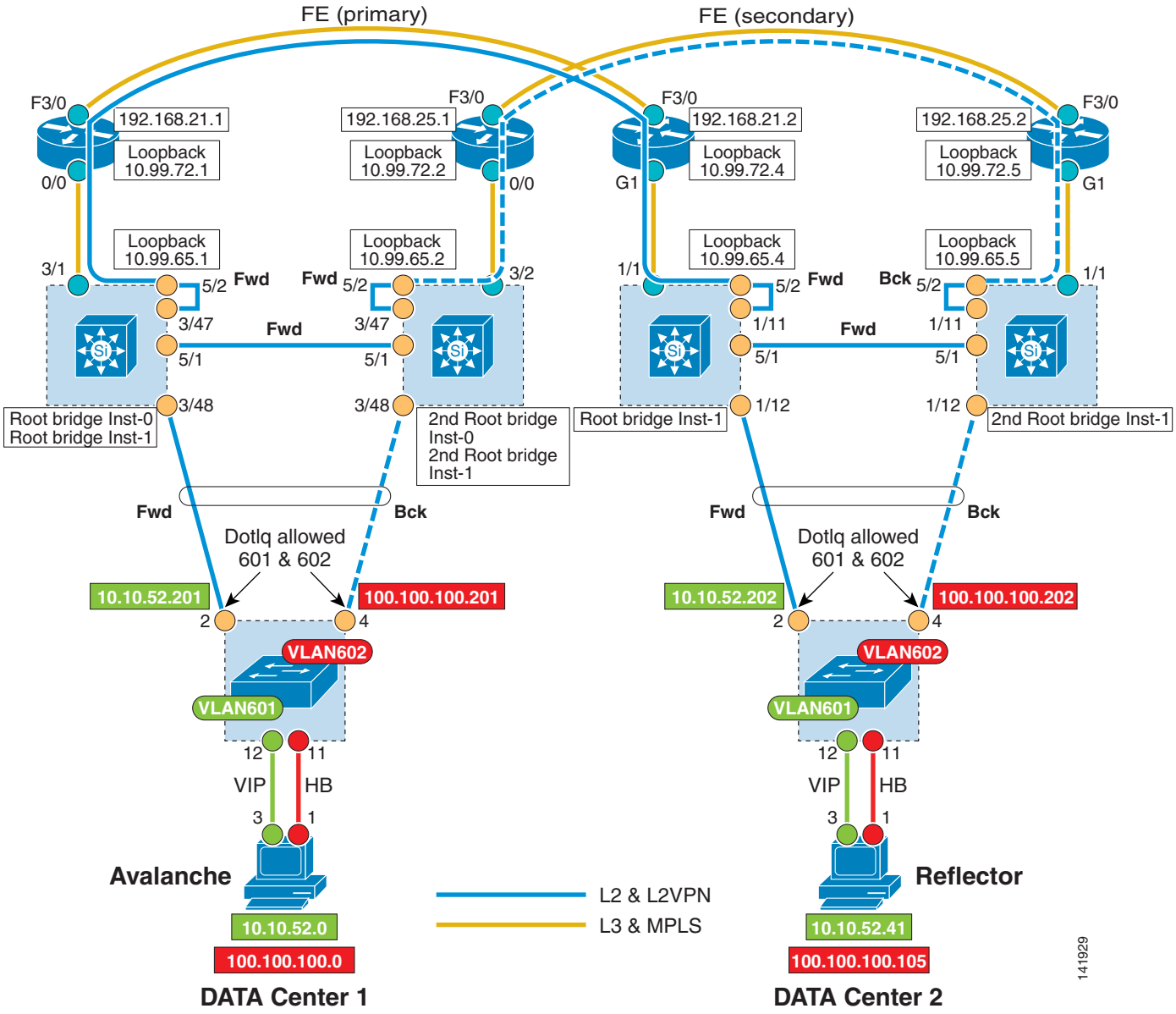
MST Region 1:

- The root bridge for Instance 1 is located on the Catalyst 6000 Series switch, DC1 Left.
  - VLAN 601 (VIP) and VLAN 602 (HB) are mapped to the Instance 1.
- The secondary root bridge for Instance 1 is located on the Catalyst 6000 Series switch DC1 Right.

MST Region 2:

- The root bridge for Instance 1 is located on the Catalyst 6000 Series switch DC2 Left.
  - VLAN 601 (VIP) and Vlan 602 (HB) are mapped to the Instance 1.
- The secondary root bridge for Instance 1 is located on the Catalyst 6000 Series switch DC2 Right.

Figure A-5 Global Test Layer 2 Site-to-Site



MST Configuration

By default, all switches use the same cost for the gigabit Ethernet interfaces (20000). The MST region appears as a logical bridge from outside. The two MST regions can be conceived as two logical bridges connected, as shown in Figure A-6.

Figure A-6 MST Configuration

Assuming the root bridge for Instance 0 is on site A (DC1 in this test), the spanning tree for the *logical* switch on site B (DC2) activates only one link to reach the remote switch (normal STP behavior to prevent any Layer 2 looping). This is computed based on the Port Cost. If the Port Costs are equal, then the lowest Port ID wins. Here, within the MST, there are two physical switches. Therefore, the Port ID cannot be taken into consideration to compute the forwarding path, and MST uses the lowest MAC address to enable the forwarding path to the root bridge located on the remote site.

As previously stated, it was decided to use the left switches of DC1 (Catalyst 6000 Series switch-DC1-Left) to be the root bridge for IST-0. Unfortunately, the lowest Bridge ID of the Catalyst 6000 Series switch in DC2 is the Catalyst 6000 Series switch-DC2-Right. Therefore, the forwarding link to DC1 is on Catalyst 6000 Series switch-DC2-Right, the edge interface of the Catalyst 6000 Series switch-DC2-Left being backup for the remote DC1.

To position the STP path where desired for this test, the port cost of the edge interface G1/11 of the Catalyst 6000 Series switch-DC2-Right is increased by one, as follows:

```
Cat6k-DC2-right#sho span mst 0 Before any change

###### MST00        vlans mapped:   none
Bridge      address 0011.5de0.0c00  priority  32768 (32768 sysid 0)
Root        address 0005.dce7.1440  priority  32768 (32768 sysid 0)
            port    Gi1/11          path cost 20000
IST master  this switch
Operational hello time 2, forward delay 15, max age 20
Configured  hello time 2, forward delay 15, max age 20, max hops 20

Interface        Role Sts Cost      Prio.Nbr Type
---------------- ---- --- --------- -------- --------------------------------
Gi1/11           Root FWD 20000     128.11   P2p Bound(RSTP)
Gi1/12           Desg FWD 20000     128.12   P2p
Po10             Desg FWD 20000     128.1665 P2p
Po259            Desg FWD 5000      128.1667 Edge P2p

Cat6k-DC2-right#conf t
Enter configuration commands, one per line.  End with CNTL/Z.
Cat6k-DC2-right(config)#inter g1/11
Cat6k-DC2-right(config-if)#span cost 20001 default = 20000

Cat6k-DC2-right#sho span mst 0

###### MST00        vlans mapped:   none
Bridge      address 0011.5de0.0c00  priority  32768 (32768 sysid 0)
Root        address 0005.dce7.1440  priority  32768 (32768 sysid 0)
            port    Po10            path cost 20000
IST master  address 0012.449a.5000  priority  32768 (32768 sysid 0)
                                    path cost 20000     rem hops 19
Operational hello time 2, forward delay 15, max age 20
```

```
Configured  hello time 2, forward delay 15, max age 20, max hops 20

Interface        Role Sts Cost      Prio.Nbr Type
---------------- ---- --- --------- -------- --------------------------------
Gi1/11           Altn BLK 20001     128.11   P2p Bound(RSTP)
Gi1/12           Desg LRN 20000     128.12   P2p
Po10             Root FWD 20000     128.1665 P2p
Po259            Desg FWD 5000      128.1667 Edge P2p



Cat6k-DC1-left#sho span mst conf
Name      [DC1]
Revision  10
Instance  Vlans mapped
--------  ------------------------------------------------------------------------
0         none
1         601-602
2         1-600,603-4094
-------------------------------------------------------------------------------

Cat6k-DC1-left#sho span mst  0

###### MST00        vlans mapped:   none
Bridge      address 0005.dce7.1440  priority  32768 (32768 sysid 0)
Root        this switch for CST and IST
Configured  hello time 2, forward delay 15, max age 20, max hops 20

Interface        Role Sts Cost      Prio.Nbr Type
---------------- ---- --- --------- -------- --------------------------------
Gi3/47           Desg FWD 20000     128.303  P2p Bound(RSTP) to remote DC2
Gi3/48           Desg FWD 20000     128.304  P2p to Access Switch
Po10             Desg FWD 20000     128.1665 P2p Channel using interface G5/1
Po260            Desg FWD 5000      128.1667 Edge P2p

Cat6k-DC1-left#sho span mst  1

###### MST01        vlans mapped:   601-602
Bridge      address 0005.dce7.1440  priority  24577 (24576 sysid 1)
Root        this switch for MST01

Interface        Role Sts Cost      Prio.Nbr Type
---------------- ---- --- --------- -------- --------------------------------
Gi3/47           Boun FWD 20000     128.303  P2p Bound(RSTP)
Gi3/48           Desg FWD 20000     128.304  P2p
Po10             Desg FWD 20000     128.1665 P2p
Po260            Desg FWD 5000      128.1667 Edge P2p



Cat6k-DC1-right#sho span mst 0

###### MST00        vlans mapped:   none
Bridge      address 0007.0d0b.8400  priority  32768 (32768 sysid 0)
Root        address 0005.dce7.1440  priority  32768 (32768 sysid 0)
            port    Po10            path cost 0
IST master  address 0005.dce7.1440  priority  32768 (32768 sysid 0)
                                    path cost 20000     rem hops 19
Operational hello time 2, forward delay 15, max age 20
Configured  hello time 2, forward delay 15, max age 20, max hops 20

Interface        Role Sts Cost      Prio.Nbr Type
---------------- ---- --- --------- -------- --------------------------------
```

```
Gi3/47           Desg FWD 20000     128.303  P2p Bound(RSTP)
Gi3/48           Desg FWD 20000     128.304  P2p
Po10             Root FWD 20000     128.1665 P2p
Po260            Desg FWD 5000      128.1667 Edge P2p

Cat6k-DC1-right#sho span mst 1

###### MST01        vlans mapped:   601-602
Bridge      address 0007.0d0b.8400  priority  28673 (28672 sysid 1)
Root        address 0005.dce7.1440  priority  24577 (24576 sysid 1)
            port    Po10            cost      20000               rem hops 19

Interface        Role Sts Cost      Prio.Nbr Type
---------------- ---- --- --------- -------- --------------------------------
Gi3/47           Boun FWD 20000     128.303  P2p Bound(RSTP)
Gi3/48           Desg FWD 20000     128.304  P2p
Po10             Root FWD 20000     128.1665 P2p
Po260            Desg FWD 5000      128.1667 Edge P2p

3750-DC1 #sho span mst 0

##### MST0    vlans mapped:   none
Bridge        address 0013.1a65.4780  priority      32768 (32768 sysid 0)
Root          address 0005.dce7.1440  priority      32768 (32768 sysid 0)
              port    Gi1/0/2         path cost     0
Regional Root address 0005.dce7.1440  priority      32768 (32768 sysid 0)
                                      internal cost 20000     rem hops 19
Operational   hello time 2 , forward delay 15, max age 20, txholdcount 6
Configured    hello time 2 , forward delay 15, max age 20, max hops    20

Interface        Role Sts Cost      Prio.Nbr Type
---------------- ---- --- --------- -------- --------------------------------
Gi1/0/2          Root FWD 20000     128.2    P2p Pre-STD-Rx to Cat6-DC1-Left
Gi1/0/4          Altn BLK 20000     128.4    P2p Pre-STD-Rx to Cat6-DC1-Right
Gi1/0/11         Desg FWD 20000     128.11   Edge P2p to Avalanche Interf 1
Gi1/0/12         Desg FWD 20000     128.12   Edge P2p to Avalanche Interf 3

3750-DC1 #sho span mst 1

##### MST1    vlans mapped:   601-602
Bridge        address 0013.1a65.4780  priority      32769 (32768 sysid 1)
Root          address 0005.dce7.1440  priority      24577 (24576 sysid 1)
              port    Gi1/0/2         cost          20000     rem hops 19

Interface        Role Sts Cost      Prio.Nbr Type
---------------- ---- --- --------- -------- --------------------------------
Gi1/0/2          Root FWD 20000     128.2    P2p Pre-STD-Rx
Gi1/0/4          Altn BLK 20000     128.4    P2p Pre-STD-Rx
Gi1/0/11         Desg FWD 20000     128.11   Edge P2p
Gi1/0/12         Desg FWD 20000     128.12   Edge P2p

Cat6k-DC2-left#sho span mst 0

###### MST00        vlans mapped:   none
Bridge      address 0012.449a.5000  priority  32768 (32768 sysid 0)
Root        address 0005.dce7.1440  priority  32768 (32768 sysid 0)
            port    Gi1/11          path cost 20000
IST master  this switch
Operational hello time 2, forward delay 15, max age 20
Configured  hello time 2, forward delay 15, max age 20, max hops 20

Interface        Role Sts Cost      Prio.Nbr Type
---------------- ---- --- --------- -------- --------------------------------
Gi1/11           Root FWD 20000     128.11   P2p Bound(RSTP) to remote DC1
```

```
Gi1/12           Desg FWD 20000     128.12   P2p to Access Switch
Po10             Desg FWD 20000     128.1665 P2p Channel using interface G5/1
Po259            Desg FWD 5000      128.1667 Edge P2p

Cat6k-DC2-left#sho span mst 1

###### MST01        vlans mapped:   600-602
Bridge      address 0012.449a.5000  priority  24577 (24576 sysid 1)
Root        this switch for MST01

Interface        Role Sts Cost      Prio.Nbr Type
---------------- ---- --- --------- -------- --------------------------------
Gi1/11           Boun FWD 20000     128.11   P2p Bound(RSTP)
Gi1/12           Boun FWD 20000     128.12   P2p
Po10             Desg FWD 20000     128.1665 P2p
Po259            Desg FWD 5000      128.1667 Edge P2p

Cat6k-DC2-right#sho span mst 0

###### MST00        vlans mapped:   none
Bridge      address 0011.5de0.0c00  priority  32768 (32768 sysid 0)
Root        address 0005.dce7.1440  priority  32768 (32768 sysid 0)
            port    Po10            path cost 20000
IST master  address 0012.449a.5000  priority  32768 (32768 sysid 0)
                                    path cost 20000     rem hops 19
Operational hello time 2, forward delay 15, max age 20
Configured  hello time 2, forward delay 15, max age 20, max hops 20

Interface        Role Sts Cost      Prio.Nbr Type
---------------- ---- --- --------- -------- --------------------------------
Gi1/11           Altn BLK 20001     128.11   P2p Bound(RSTP)
Gi1/12           Desg LRN 20000     128.12   P2p
Po10             Root FWD 20000     128.1665 P2p
Po259            Desg FWD 5000      128.1667 Edge P2p

Cat6k-DC2-right#sho span mst 1

###### MST01        vlans mapped:   601-602
Bridge      address 0011.5de0.0c00  priority  28673 (28672 sysid 1)
Root        address 0012.449a.5000  priority  24577 (24576 sysid 1)
            port    Po10            cost      20000               rem hops 19

Interface        Role Sts Cost      Prio.Nbr Type
---------------- ---- --- --------- -------- --------------------------------
Gi1/11           Boun BLK 20001     128.11   P2p Bound(RSTP)
Gi1/12           Desg FWD 20000     128.12   P2p
Po10             Root FWD 20000     128.1665 P2p
Po259            Desg FWD 5000      128.1667 Edge P2p


3750-DC2 #sho span mst 0


##### MST0    vlans mapped:   none
Bridge        address 0013.1a4a.a080  priority      32768 (32768 sysid 0)
Root          address 0005.dce7.1440  priority      32768 (32768 sysid 0)
              port    Gi1/0/2         path cost     20000
Regional Root address 0012.449a.5000  priority      32768 (32768 sysid 0)
                                      internal cost 20000     rem hops 19
Operational   hello time 2 , forward delay 15, max age 20, txholdcount 6
Configured    hello time 2 , forward delay 15, max age 20, max hops    20

Interface        Role Sts Cost      Prio.Nbr Type
---------------- ---- --- --------- -------- --------------------------------
```

```
Gi1/0/2          Root FWD 20000     128.2    P2p Pre-STD-Rx To Cat6k-DC-Left
Gi1/0/4          Altn BLK 20000     128.4    P2p Pre-STD-Rx To Cat6k-DC-Right
Gi1/0/11         Desg FWD 20000     128.11   Edge P2p Reflector Intf. 1
Gi1/0/12         Desg FWD 20000     128.12   Edge P2p Reflector Intf. 3

3750-top-Rack-5#sho span mst 1

##### MST1    vlans mapped:   601-602
Bridge        address 0013.1a4a.a080  priority      32769 (32768 sysid 1)
Root          address 0012.449a.5000  priority      24577 (24576 sysid 1)
              port    Gi1/0/2         cost          20000     rem hops 19

Interface        Role Sts Cost      Prio.Nbr Type
---------------- ---- --- --------- -------- --------------------------------
Gi1/0/2          Root FWD 20000     128.2    P2p Pre-STD-Rx
Gi1/0/4          Altn BLK 20000     128.4    P2p Pre-STD-Rx
Gi1/0/11         Desg FWD 20000     128.11   Edge P2p
Gi1/0/12         Desg FWD 20000     128.12   Edge P2p
```

Failover Test Results

Data Center 1 (Catalyst 6000 Series Switch—DC1-Left)

Disconnect G3/47 (interface port Xconnect for the pseudowire)

Reconnect G3/47

Disconnect G3/48 (Forwarding interface to access switch)

Reconnect G3/48

Other interfaces have no impact.

Shutdown for maintenance of the root bridge (Catalyst 6000 Series switch-DC1-left)

Rebooting the original root bridge (Catalyst 6000 Series switch-DC1-Left) has no impact (zero packet lost) while it becomes Root back.

Data Center 1 (Catalyst 6000 Series Switch—DC1-Right)

Any actions to try to disrupt the traffic applied on the standby switch (Catalyst 6000 Series switch-DC1-Right) while the root bridge is up and running, have no impact on the traffic flow from data center 1 to data center 2 (zero packet loss).

Data Center 2 (Catalyst 6000 Series Switch—DC2-Left)

Disconnect G1/11 (interface port Xconnect for the pseudowire)

Reconnect G1/11

Disconnect G1/12 (Forwarding interface to access switch)

Reconnect G1/12

Shutdown of the Forwarding Bridge (Catalyst 6000 Series switch-DC1-Left) to remote DC

Rebooting the forwarding Bridge (Catalyst 6000 Series switch-DC2-Left) has no impact (zero packet loss) while it comes back to forwarding state.

Data Center 2 (Catalyst 6000 Series Switch—DC2-Right)

Any actions to try to disrupt the traffic applied on the standby switch (Catalyst 6000 Series switch—DC1-Right) while the root bridge is up and running have no impact on the traffic flow between the two data centers (zero packet loss).

Disconnect G3/47 (interface port Xconnect for the pseudowire)

Reconnect G3/47

Disconnect G3/48 (Forwarding interface to access switch)

Reconnect G3/48

Other interfaces have no impact.

Shutdown for maintenance of the root bridge (Catalyst 6000 Series switch-DC1-Left)

Rebooting the original root bridge (Catalyst 6000 Series switch-DC1-Left) has no impact (zero packets lost) while it becomes Root back.

Disconnect G1/11 (interface port Xconnect for the pseudowire)

Reconnect G1/11

Disconnect G1/12 (Forwarding interface to access switch)

Reconnect G1/12

Shutdown for maintenance of the Forwarding Bridge (Catalyst 6000 Series switch-DC1-Left) to remote DC

Rebooting the forwarding Bridge (Catalyst 6000 Series switch-DC2-Left) has no impact (zero packet loss) while it comes back to forwarding state.