White Paper

High-Availability Storage Networks with Cisco MDS 9500 Series Multilayer Directors

In today's enterprise environments, high availability is no longer optional. Data availability is more important than ever as data growth rates continue to accelerate. As enterprises and applications grow, the ability to increase the size of the associated data center infrastructure is critical. The advent of a worldwide economy, facilitated by the Internet, has shifted normal operations from a workday-only to a 24-hour model. In this "always on" world, more stringent requirements have been placed on high availability. To keep an enterprise running, data, a company's most crucial asset, must be available at all times. Not only can loss of data have catastrophic effects, but also the inability to access that data can be extremely costly.

Although 99 percent uptime can seem like a significant achievement, such a "highly available" environment would be down for over 83 hours per year. This amount of downtime could have a significant effect on a business of any size. In designing a highly available solution, the cost of downtime must be considered. A 99 percent uptime environment could cost a large financial brokerage firm over US$540 million in lost revenue and productivity per year. Table 1 illustrates the cost of downtime in several industries. Increasing uptime to 99.999 percent reduces this loss to US$540,000 per year.

Table 1. Cost of Downtime

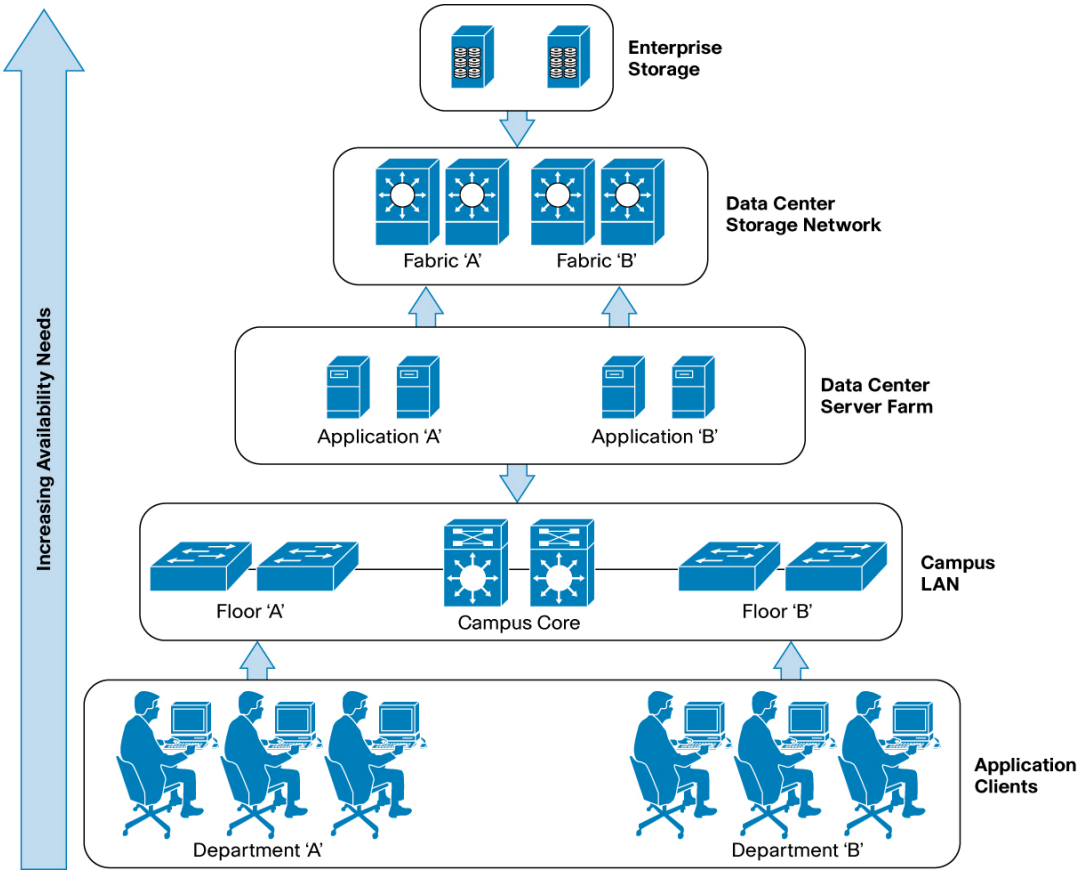
Type of Business	Cost per Hour (US$)	Availability (Percentage)	Minimum Downtime Hours per Year
Financial Brokerage	6.5 million	99.999	5
Credit Card Authorization	2.6 million	99.99	50
Home Shopping	0.1 million	99.9	500
Catalog Sales	0.09 million	99	5000
Airline Reservations	0.09 million	90	50,000

Source: Fibre Channel Industry Association, "Business Continuity When Disaster Strikes," http://www.fibrechannel.com/technology/index.master.html, Horison, Inc.

Achieving 99.999 percent uptime is not always easy. A highly available storage infrastructure is required for achieving data availability. It includes several components, including Redundant Array of Independent Disks (RAID) technology, multiple copies of data across a clustered system, clustering over distance, Storage Area Networks (SANs), and reliable tape backups. Among these component pieces, SAN architecture, the focus of this paper, enables enterprisewide high-availability configurations that will grow with the organization and protect your investment in data storage.

Storage uptime plays a primary role in the entire organization. Each employee relies on access to storage, whether through an application server or directly from the employee's workstation through file servers, to make important business decisions. When storage availability problems arise, the effect is sure to be felt throughout the entire organization, as Figure 1 shows.

Figure 1. Enterprise High-Availability Priorities



Designing High-Availability Solutions

To design a highly available storage environment, it's critical to take an end-to-end approach. You must consider every system carefully, including all aspects of storage subsystem, the storage network, and the application host.

The Storage Subsystem

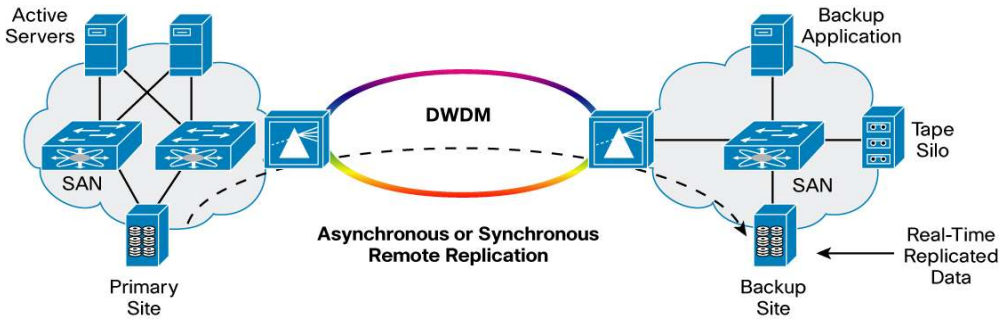
Three main aspects of the storage subsystem are critical to designing a highly available solution.

Data Protection

- **Redundant cache** - Virtually all storage subsystems employ some type of front-end cache. The cache increases the subsystem's response time by caching write operations, and the cost in overall latency can be high if data is written directly to a disk. When an application server issues a write command, the storage subsystem will write the data in the cache, which is a much lower-latency operation, and inform the application server that the write has been completed. The data is then written, or destaged, to physical disk at a later time. Many subsystems mirror the front-end cache for extra availability. If one cache fails, the data is not lost because of the mirrored copy.

- **RAID** - RAID technology is used in almost all subsystems to provide higher data availability in terms of protection and access speed. The RAID implementation can be just simple RAID 1 (mirroring data to two or more disks) or RAID 5, using advanced striping with data parity calculations. RAID 1 and RAID 5 techniques provide different levels of protection and performance, but both provide additional availability in the case of a disk failure.
- **Data replication** - Storage replication is commonly deployed to protect against an entire subsystem failure, with a backup site connected possibly through Dense Wavelength Division Multiplexing (DWDM) to the primary site. Although replication is commonly used over longer distances in an asynchronous form, asynchronous replication is outside the scope of this document. Synchronous storage replication ((Figure 2) can be used to help ensure data availability within a local data center with minimal performance effect to the overriding application. This replication can be done by the storage subsystem or through an external host-based application. In either case, the end result is two independent storage subsystems, each with a real-time copy of the same data.

Figure 2. Synchronous Data Replication Model

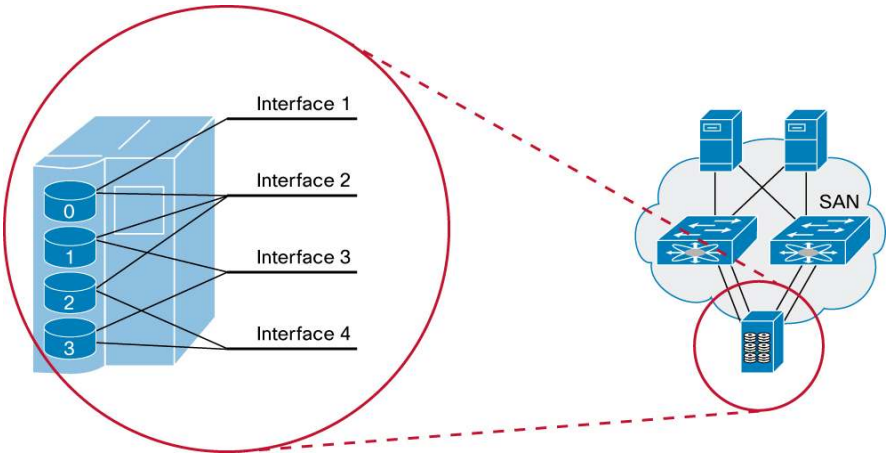


Subsystem Connectivity

Connectivity to the storage subsystem is almost as important as the integrity of the subsystem itself. If an application cannot access its storage, it will experience downtime. Therefore, the way in which storage is provisioned within a storage subsystem is very important to achieving high-availability overall.

Redundant interfaces are a critical piece here. Connectivity must be redundant to achieve true high availability. A disk logical unit must be exported through multiple interfaces on the storage subsystem (Figure 3). This not only allows for multipathing at the host level but also provides the added redundancy of two physical connections from the disk subsystem itself.

Figure 3. Redundant Disk Subsystem Interfaces for High Availability



Subsystem Hardware Redundancy

- **Power redundancy** - Power is critical in storage subsystems. Dual power supplies are standard equipment in most storage subsystems. Additionally, most subsystems with a front-end cache have some level of battery backup for the cache. Some subsystems use smaller batteries to keep power to the cache only for several days. Larger batteries are also used in some subsystems to keep the entire system running long enough to destage the data from the cache to the physical disk.
- **Hot disk sparing** - Most storage subsystems provide spare physical disks. These spare disks, which can vary in number amount per subsystem, are utilized only if a disk shows signs of failure or suddenly fails. The subsystem monitors each physical hard disk for potential signs of failure. If the subsystem notices failure signs, data from the failing disk can be copied to the hot spare. Also, with RAID typically used in storage subsystems, if a disk in a RAID group suddenly fails, a hot spare disk can be used to rebuild the lost data. In either case, the subsystem is able to recover, and access to the data is not disrupted.

The Storage Network

The network or fabric that provides the connectivity between hosts and storage is also an important aspect of achieving high availability. Best design practices are employed to help ensure that there are no single points of failure within the design. Such design practices also help ensure that the right level of redundancy is used, because excessive redundancy can potentially degrade failure recovery performance, causing recovery to take longer.

Storage Network Hardware

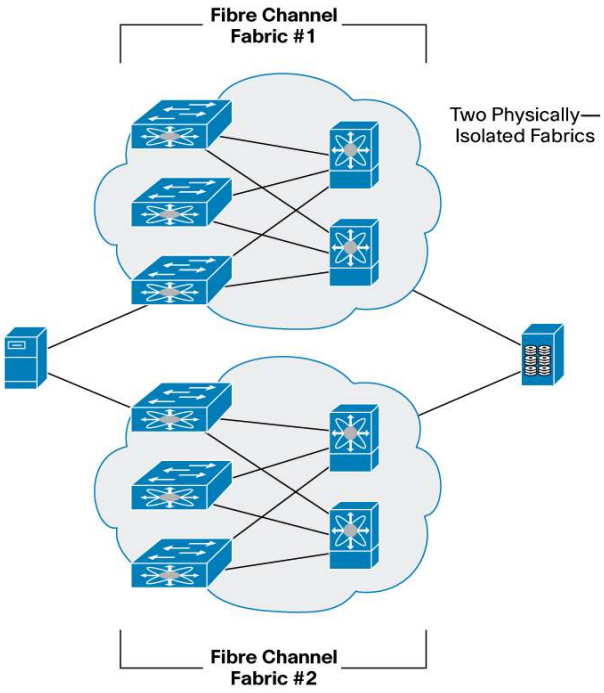
As with all other hardware components making up a storage solution, the hardware in a Fibre Channel switch must be redundant. In the switch class of products, hardware redundancy is typically limited to dual power supplies. This solves power disruption problems but does not address other switch component failures.

Director-class Fibre Channel switches bring a new level of availability to the storage network. In addition to supporting redundant power, director-class switch provides redundancy in other major switch components. The control modules provide failover capability. Crossbars are also embedded in an active-active configuration, so that loss of one crossbar does not bring the system down. Software upgrades must be nondisruptive, and with director-class switches, they are. In all these ways, director-class hardware helps contribute to a true 99.999 percent uptime within the system.

Storage Network Design

- **Fabric redundancy** - Another area that requires attention in a Fibre Channel SAN is the fabric itself. Each device connected to the same physical infrastructure is in the same Fibre Channel fabric. This opens up the SAN to fabric-level events that could disrupt all devices on the network. Changes such as adding switches or changing zoning configurations could ripple through the entire connected fabric. However, designing with separate connected fabrics helps to isolate the scope of any such events. Cisco® Virtual SAN (VSAN) technology offers a way to isolate fabrics - and therefore isolate events.-using the same physical infrastructure. VSANs are discussed in more detail later in this paper. Figure 4 illustrates VSAN technology.

Figure 4. Designing SANs with Isolated Fabrics



- **Inter-Switch Links (ISLs)** - The connectivity between switches is important as the SAN grows. Relying on a single physical link between switches reduces overall redundancy in the design. Redundant ISLs provide failover capacity if a link fails.

The Application Host

Host bus adapters (HBAs) are the interface between an application server and the SAN. Similar to a network interface card, an HBA is inserted into a bus slot in the server. Although most servers do not generate enough I/O to stress a single Fibre Channel link, dual HBAs are still a requirement in a high-availability environment. Two or more HBAs provide multiple paths to storage. This not only facilitates failover if one HBA fails, but also provides load balancing across the HBAs.

HBA “multipathing” can be achieved in several ways. The following options are available to provide high availability for HBAs:

- **Subsystem software** - Most major storage subsystem providers have developed multipathing software to provide for load balancing and failover of certified HBAs. An example of this would be PowerPath from EMC. Such software is usually designed specifically for a specific vendor’s subsystem or offers an enhanced mode of operation when operating with the vendor’s own subsystem.
- **Volume management software** - Some host-based volume management applications support multipathing. An example is Veritas Dynamic Multipathing (DMP) from Symantec. This type of solution is not specific to a subsystem vendor.
- **HBA drivers** - Some HBA vendors are now providing multipathing features in the HBA driver on the host. This solution is also not specific to a particular storage vendor, although it limits the multipathing to a specific HBA vendor and possibly a particular HBA model.
- **Operating systems** - Several operating systems now support multipathing features native in the OS. This allows the multipathing features to be decoupled not only from the storage subsystem but also from the HBA.

Enhancing Storage Network Availability

Cisco MDS 9500 Series Multilayer Directors provide a number of hardware and software features that enable advanced availability within the Fibre Channel network.

Hardware Features

The following section outlines the hardware aspects of high availability within Cisco MDS 9500 Series Multilayer Directors.

Supervisor Modules

Cisco MDS 9500 Series Multilayer Directors support two supervisor modules in the chassis for redundancy. Each supervisor module consists of a control engine and a crossbar fabric. The control engine is the central processor responsible for the management of the overall system. In addition, the control engine participates in all of the networking control protocols, including all Fibre Channel services. In a redundant system, two control engines operate in an active-standby mode, with one control engine always active. The control engine that is in standby mode is actually in a stateful-standby mode such that it keeps sync with all major management and control protocols that the active control engine maintains. Although the standby control engine is not actively managing the switch, it continually receives information from the active control engine. This allows the state of the switch to be maintained between the two control engines. If the active control engine fails, the secondary control engine will transparently resume its function.

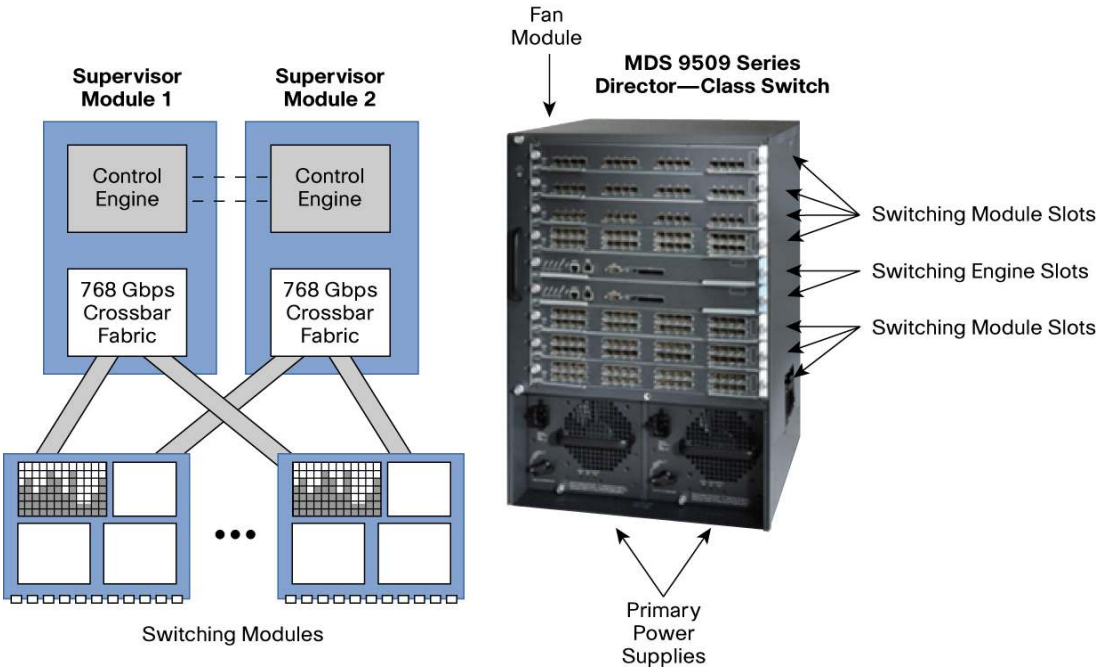
The supervisor module is a hot-swappable module. In a dual supervisor module system, this allows the module to be removed and replaced without causing disruption to the rest of the system.

Cisco MDS 9506 and MDS 9509 Multilayer Director Crossbar Fabrics

The crossbar fabric is the switching engine of the system. The crossbar provides a high-speed matrix of switching paths between all ports within the system. A crossbar fabric is embedded within each supervisor module. Therefore, a redundant system with two supervisor modules will also contain two crossbar fabrics. The two crossbar fabrics operate in a load-shared active-active mode. However, each crossbar fabric has a total switching capacity of 768 Gbps and serves 96 Gbps of bandwidth to each slot. Cisco MDS 9500 Multilayer Director will continue to switch traffic between all front-panel ports even if one of the supervisor modules fails or is removed (Figure 5).

There is an additional crossbar stage within each 8-Gbps Advanced Fibre Channel switching module that enables local switching between the ports on the module. Arbitration for this crossbar stage is enabled by the same, central redundant control logic on the supervisor modules that also does arbitration for the crossbar fabric on the supervisor modules. With the local switching enabled, the max switching bandwidth on Cisco MDS 9509 Multilayer Director is 2.7 Tbps, and on MDS 9506 it is 1.5 Tbps.

Figure 5. Cisco MDS 9500 Series Switching System



Cisco MDS 9513 Multilayer Director Crossbar Fabrics

The crossbar fabric is the switching engine of the system. The crossbar provides a high-speed matrix of switching paths between all ports within the system. Although a crossbar fabric is embedded within each supervisor module, the Cisco MDS 9513 Multilayer Director uses two dedicated crossbar modules located on the back of the chassis. The two crossbar modules operate in a load-shared active-active mode. However, each crossbar module has a total switching capacity of 2.9 Tbps and serves 256 Gbps of bandwidth to each slot. Cisco MDS 9513 Multilayer Director will continue to switch traffic between all front-panel ports even if one of the crossbar modules fails or is removed. With the local switching enabled using the crossbar stage on the 8-Gbps Advanced Fibre Channel switching modules, the system switching bandwidth on Cisco MDS 9513 is 8.4 Tbps.

Power Supplies

Cisco MDS 9500 Series Multilayer Directors support dual redundant power supplies. The power supplies run in an active-active configuration but operate independently of each other. If a power supply fails, a single power supply is sufficient to power the entire system. Each power supply is hot swappable. Individual power supplies are designed to power the whole system, allowing for replacement of a failed supply.

System Fans

Cisco MDS 9500 Series Multilayer Directors use a single fan tray to cool the entire system. Although this appears to be a nonredundant component, the tray in fact contains **multiple**, redundant fans with dual-fan controllers. All fans are variable speed fans, and the failure of one or more individual fans causes all other fans to speed up to compensate. In the unlikely event of multiple, simultaneous fan failures, the fan tray still provides sufficient cooling to keep the switch operational. The entire fan tray is hot swappable, and the system can run for up to five minutes without the fan tray installed. This allows for time to replace a fan tray while the system is running.

Software Features

Whereas traditional Fibre Channel switches rely solely on hardware redundancy for high availability, the Cisco MDS 9500 Series provides a robust set of software features to enhance the hardware-based redundancy in the typical storage network.

Nondisruptive Software Upgrades

Planned downtime is a large percentage of equipment downtime per year. A common reason for planned downtime is to upgrade software in networking devices - for instance, to fix software bugs or add new features. Regardless of the reason, even planned downtime can have a negative effect on business. A critical ability of any director-class Fibre Channel switch is the ability to load and activate new software on the switch without disrupting traffic across the SAN.

Cisco MDS 9500 Series Multilayer Directors support the ability to upgrade the supervisor module and the switching module software on the fly without disrupting traffic flowing through the switch. This allows maximum flexibility in upgrading the software while providing a path to revert to known stable software.

Internal Process Restart

A unique feature of the Cisco MDS 9500 Series is the ability to restart a failed software process. The supervisor module continually monitors all software processes. If a process fails, the supervisor can restart the process without disrupting the flow of traffic in the switch. This feature allows for increased reliability because failover of a supervisor is not required if a process can be restarted. If a process cannot be restarted or continues to fail, the primary supervisor module can then fail-over to the standby supervisor module.

Virtual Storage Area Networks

SAN designers build separate storage networks for a variety of reasons. In this case, a separate storage network refers to a completely physically isolated switch or group of switches used to connect hosts to storage. Some of the more popular reasons for building separate fabrics include the following:

- **High availability** - A common practice is to build multiple parallel fabrics and “multihome” hosts and disks into the parallel, physically isolated fabrics. Generally the primary reason for this isolation is to help ensure that fabric services such as the name service are isolated within each fabric. If a fabric service fails, it will not affect the other parallel fabrics. Therefore, the parallel fabrics provide isolated paths from hosts to disks.

- **Application and backup fabrics** - Many customers build at least two physically separate fabrics for their storage network environment. The primary idea is to dedicate one fabric to the application hosts and the second fabric to the backup environment. Using this method, backup traffic is physically isolated from main application traffic.
- **Departmental fabrics** - Many customers choose to build out separate storage network environments for departmental applications. In this case, a separate, smaller fabric is built for each department's applications.
- **Homogeneous OS fabrics** - Some customers follow a practice of building separate fabrics for different hosts with different operating systems. Because of the nature of some operating systems and their method for discovering and using storage, many customers isolate environments on separate fabrics. An example would be a Sun Solaris fabric and a Windows NT/2000 fabric.

Although these are all valid reasons for building out separate fabrics, doing so can become quite wasteful. Additional, separate fabrics mean more hardware and more money spent. What's more, the hardware is typically underutilized.

To help achieve the same isolated environments while eliminating the added expense of building physically separate fabrics, Cisco has introduced the Virtual SAN (VSAN) as a feature of the Cisco MDS 9000 Family of switches. A VSAN provides the ability to create separate virtual fabrics on top of the same physical infrastructure. Each separate virtual fabric is isolated from the others using a hardware-based frame-tagging mechanism on ISLs. Enhanced ISLs (EISLs) include added tagging information for each frame and are supported on links interconnecting any Cisco MDS 9000 Family switches. Membership in a VSAN is based on the physical port, and no physical port can belong to more than one VSAN. Therefore, whatever node is connected to a physical port becomes a member of that port's VSAN.

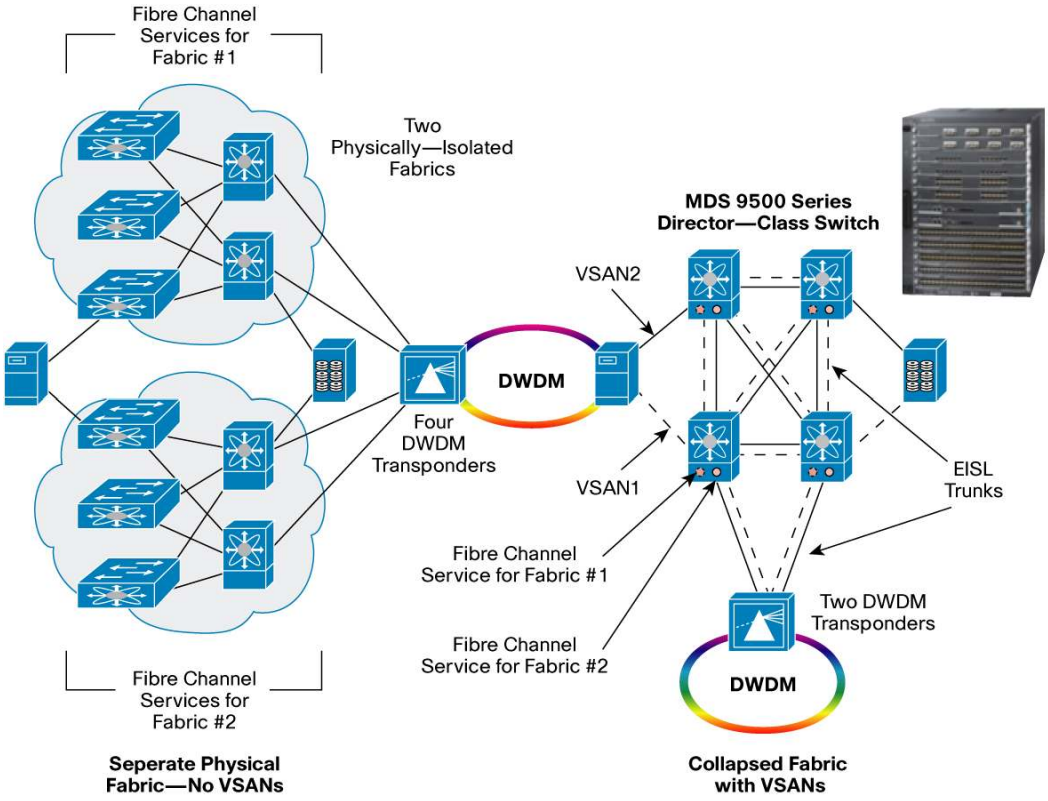
VSANs offer a great deal of flexibility to the user. For example, the Cisco MDS 9000 Family supports 1024 VSANs per physical infrastructure. Each VSAN can be selectively added to or deleted from an EISL to control the VSAN's reach. In addition, special traffic counters are provided to track statistics per VSAN.

Probably the most desirable characteristic of VSANs is their high-availability profile. In addition to providing strict hardware isolation, each new VSAN enables a fully replicated set of Fibre Channel services. Thus when a new VSAN is created, a completely separate set of services, including name server, zone server, domain controller, alias server, and login server, is created and enabled across those switches that are configured to carry the new VSAN. This replication of services provides the ability to build the isolated environments needed to address high-availability concerns over the same physical infrastructure. For example, an installation of an active zone set within VSAN 1 does not affect the fabric in any way within VSAN 2.

VSANs also provide a method to interconnect isolated fabrics in remote data centers over a common, long-haul infrastructure. Because the frame tagging is done in hardware and is included in every EISL frame, it can be transported using dense wavelength-division multiplexing (DWDM) or coarse wavelength-division multiplexing (CWDM). Therefore, traffic from several VSANs can be multiplexed across a single pair of fibers and transported a greater distance and yet still remain completely isolated.

As Figure 6 illustrates, VSANs bring scalability to a new level by using a common redundant physical infrastructure to build flexible isolated fabrics to achieve high-availability goals.

Figure 6. Using VSANs to Reduce SAN Complexity

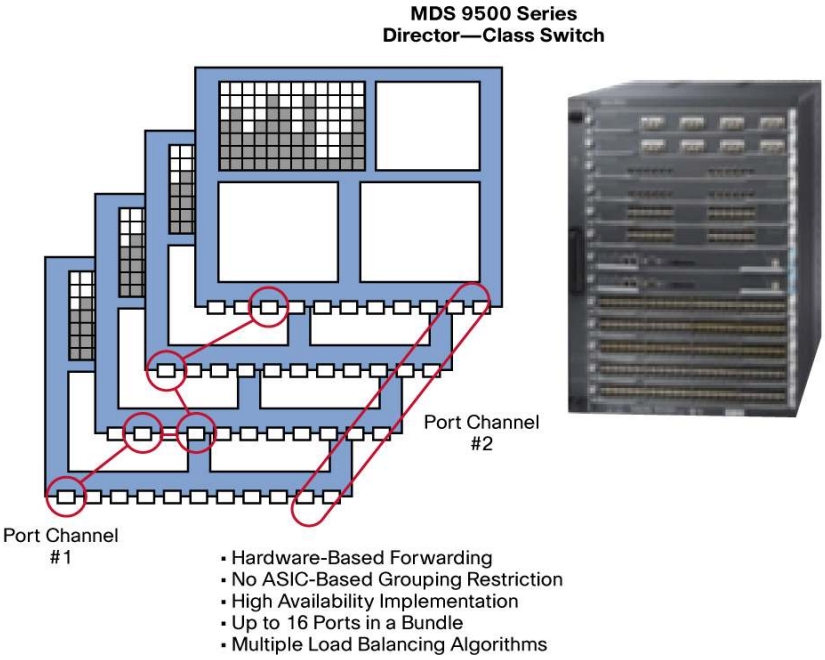


Fibre Channel PortChannels

As Fibre Channel fabrics grow larger, more switches are generally required to meet the port count requirements. ISLs facilitate switch-to-switch connectivity. As with all other connections in the SAN, these links must be redundant. With Cisco PortChannel technology, up to 16 independent physical links can be combined to create one logical ISL between two switches (Figure 7). This provides not only a completely resilient logical link but also up to 32 Gbps of bandwidth between two switches.

An important advantage of Cisco PortChannel technology is the ability for the bundled physical links to be located on any port on any switching module in the switch. Because the physical links are spread across multiple switching modules, protection is provided not only from link failures, such as cable breaks and faulty optics, but also from a switching module failure.

Figure 7. Port Channeling in the Cisco MDS 9500 Series



Cisco MDS 9500 Series Multilayer Directors support two different load-balancing algorithms across PortChannels. The first algorithm looks at the source and destination Fibre Channel IDs (FCIDs) of frames prior to entering the PortChannel. A hash is created in hardware from the source and destination FCID within the frame that serves as an index to which physical link in the virtual link this traffic should take. Traffic from that source-destination FCID pair will always travel over the same link. Other combinations of source-destination FCIDs will make independent link decisions and might or might not travel across the same link. Traffic from the destination to the source does not necessarily travel across the same physical link, because the switch on the destination side makes an independent decision about link traffic.

The second algorithm in the Cisco MDS 9500 Series is load balancing based on the source-destination FCIDs as well as the Exchange IDs (OX_ID, RX_ID) of the operation. With every operation, a new Exchange ID is used, and a new physical link routing decision is made. This allows for maximum efficiency of the entire PortChannel, even between the same source and destination nodes. Using this algorithm, exchanges from the same source and destination can be distributed across the links of a PortChannel, while keeping all frames associated within any one particular exchange in order.

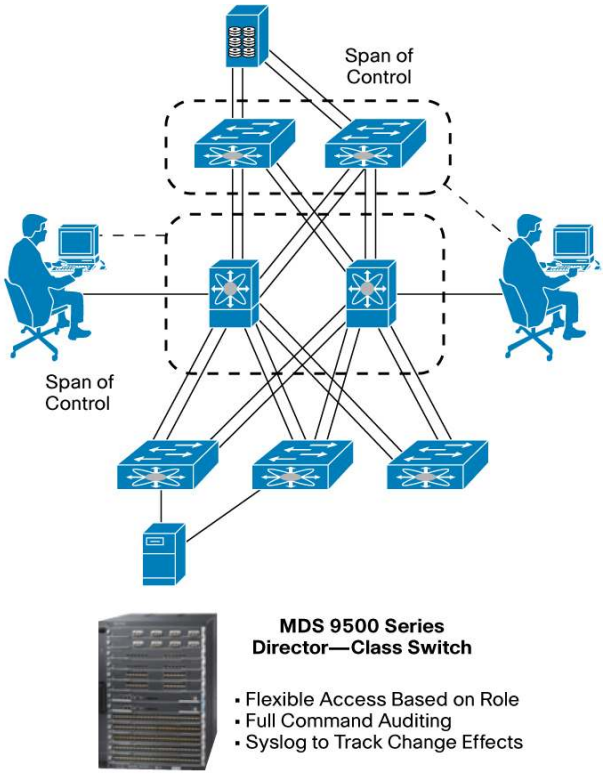
Role-Based Security

Security is not a consideration normally associated with high availability. However, one of the leading causes of downtime is human error. A user may mistakenly carry out a command without fully realizing the results of that command. The Cisco MDS 9000 Series Multilayer Directors and Fabric Switches support a role-based security methodology to help ensure that only authorized individuals have access to critical functions within the fabric.

Each user is assigned to a role, better known as a group ID, which is given a specific access level within the fabric (see Figure 8). This access level dictates the commands - or more specifically which nodes of the command-line interface (CLI) command parser tree - to which the particular role has access. For example, you can create a role called “no debug” that allows users assigned to the role to implement any command with the exception of any debug commands. The granularity of this permission system can be two levels deep within the parser tree. For instance, you can define a role called “no debug fspf” that allows a user to implement any system command, including debug commands, with the exception of FSPF debug commands.

Roles can be defined and assigned locally within a switch by using CLI commands. Role assignments can even be centralized in a RADIUS server for easier management. There are default roles - network administrator (full access) and network operator (read-only access). Up to 64 custom roles can be defined. Only a user within the network administrator role can create new roles.

Figure 8. Cisco MDS 9500 Series Role-Based Access



Summary

Downtime in a storage network can have a significant negative effect on the entire business infrastructure. This can cost millions of dollars in lost revenue on an annual basis. With use of a robust and highly resilient SAN, downtime can be significantly reduced or eliminated. Cisco MDS 9500 Series Multilayer Directors provide the hardware redundancy and reliability to achieve 99.999 percent hardware uptime. In addition to hardware redundancy, the Cisco MDS 9500 Series provides highly resilient software with an innovative high-availability feature set designed to eliminate downtime in the storage network.

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **www.cisco.com/go/offices.**

Printed in USA C11-330666-01 05/11