



## 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 shift to a worldwide economy, facilitated by the Internet, has shifted normal operations from 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
Teleticket	0.07 million	—	—

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 the core of 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, SAN architecture enables enterprise-wide high-availability configurations that will grow with the enterprise and protect your investment in data storage. Several factors are involved in designing a highly available SAN:

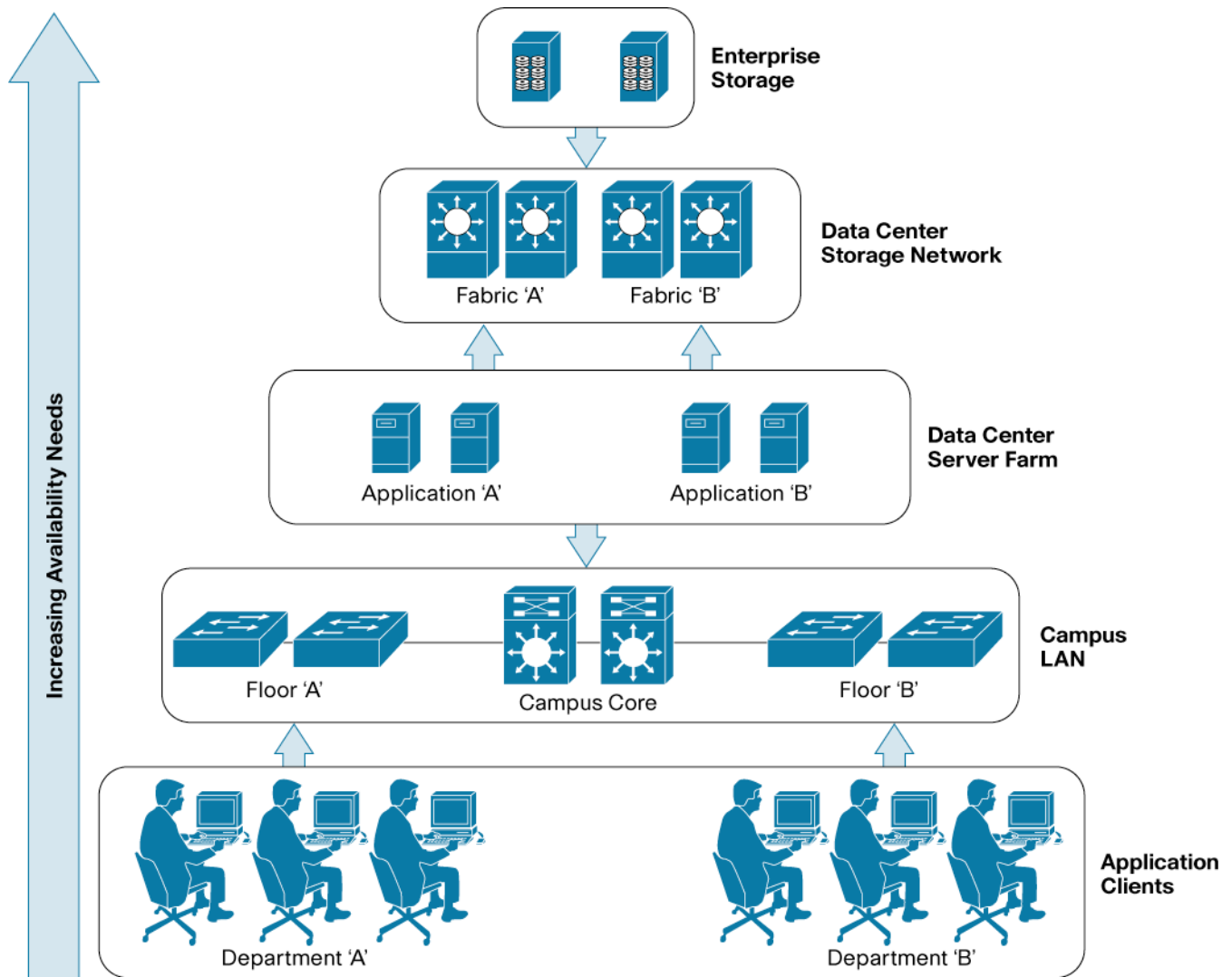
- Possible negative effects from human error and ways to prevent these effects
- Environmental problems such as power disruption, air conditioning failure, or plumbing problems
- Software or hardware failures of infrastructure devices such as switches or directors
- Planned downtime due to software upgrades and hardware maintenance
- Threats posed by hackers

Some of these events, such as a hardware failure and a power disruption, can be alleviated through a solid implemented design. Other factors, such as human error, are not as easy to alleviate through design.

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 problems arise concerning storage availability, the effect is sure to be felt throughout the entire organization.

The highest possible uptime must be achieved to limit or eliminate any possible negative business effect. (See Figure 1)

**Figure 1.** Enterprise High-Availability Priorities



## DESIGNING HIGH-AVAILABILITY SOLUTIONS

An end-to-end approach is required for designing a highly available storage environment. It is not sufficient to simply consider components of the storage solution. Each of the following components must also be considered.

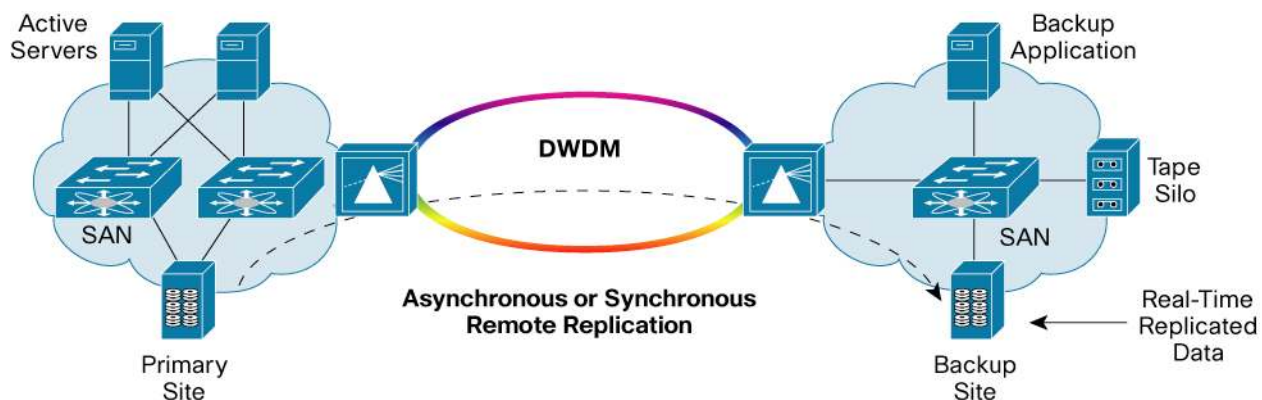
### The Storage Subsystem

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, a scenario that can be costly if written directly to a disk in terms of overall latency. When an application server issues a write command, the storage subsystem will write the data in the cache, a much lower latency, and inform the application server that the write was 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 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. Although replication is commonly used over longer distances in an asynchronous form, this is outside the scope of this document. Synchronous storage replication 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. (See Figure 2)

Figure 2. Synchronous Data Replication Model

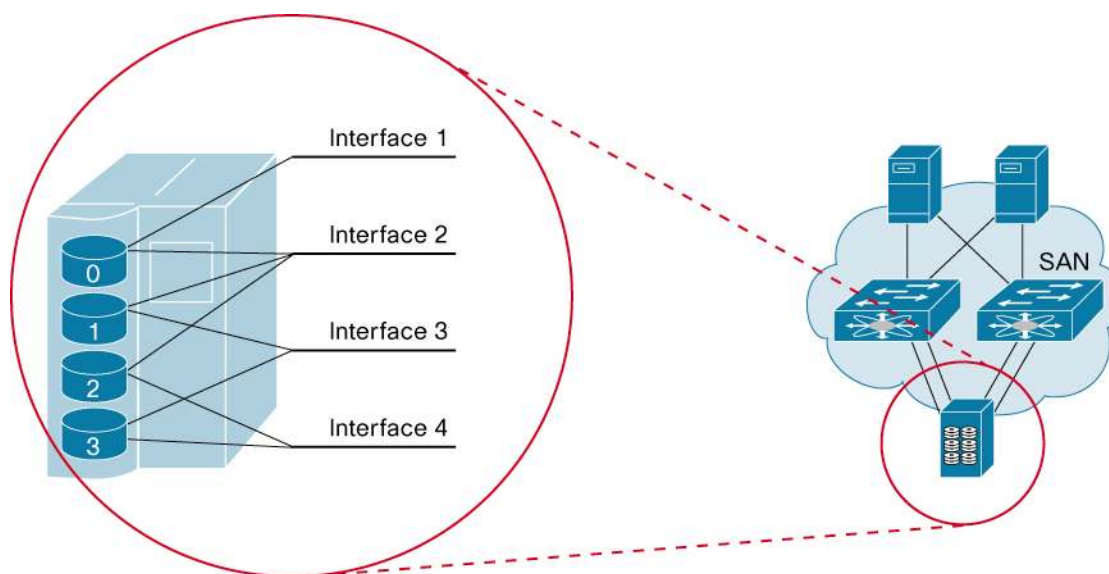


## Subsystem Connectivity

Connectivity to the storage is almost as important as the integrity of the storage 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 the overall storage high-availability solution.

**Redundant interfaces**—Connectivity must be redundant to achieve true high availability. A disk logical unit must be exported through multiple interfaces on the storage subsystem. This not only allows for multi-pathing at the host level but also provides the added redundancy of two physical connections from the disk subsystem itself. (See Figure 3)

**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 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 up to 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 amount per subsystem, are only utilized 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 of 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 component of the overall high-availability solution. 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 cause degradation in failure recovery time.

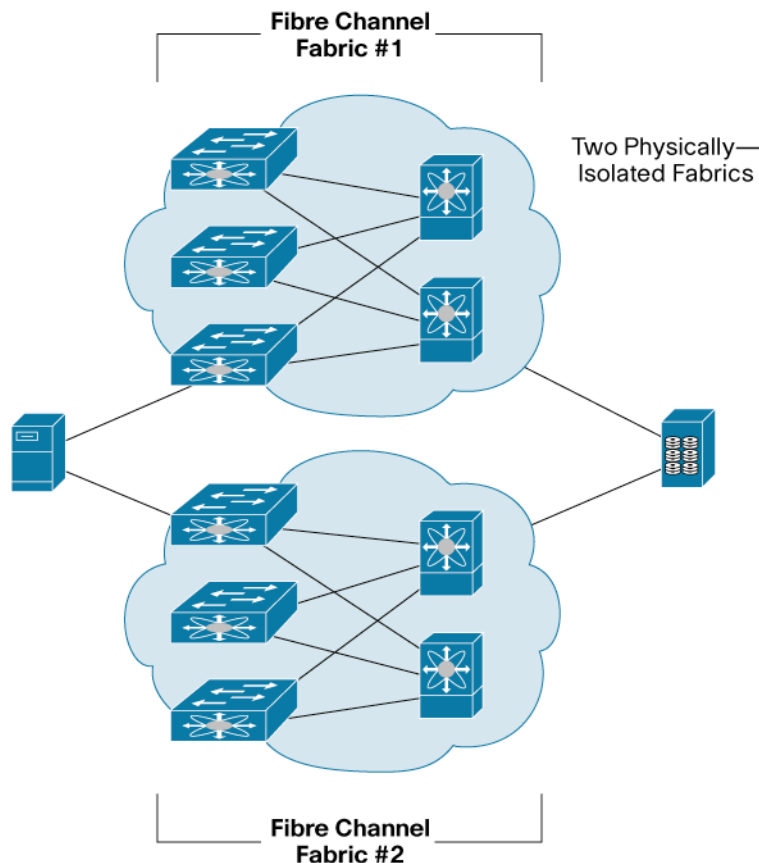
## Storage Network Hardware

**Switch 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. Not only do they support redundant power, but also every other major component is redundant. The control modules provide failover capability. Crossbars are also embedded in a redundant configuration. Software upgrades must be non-disruptive. Director-class hardware therefore 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. Therefore, designing with separate connected fabrics helps to isolate the scope of any such events. The Cisco Systems® Virtual SAN (VSAN) capability offers a way to replicate this environment, namely, the isolation of events, using the same physical infrastructure. VSANs are discussed in more detail later in this paper. (See Figure 4)

**Figure 4.** Designing SANs with Isolated Fabrics



- **Interswitch 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. This “multi-pathing” 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 multi-pathing 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 that 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 multi-pathing. An example of this would be Dynamic Multi-pathing (DMP) from VERITAS. This type of solution is not specific to a subsystem vendor.
- **HBA drivers**—Some HBA vendors are now providing multi-pathing features in the HBA driver on the host. This solution is also not specific to a particular storage vendor, although it limits the multi-pathing to a specific HBA vendor and possibly a particular HBA model.
- **OS**—Several OSs now support multi-pathing features native in the OS. This allows the multi-pathing 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

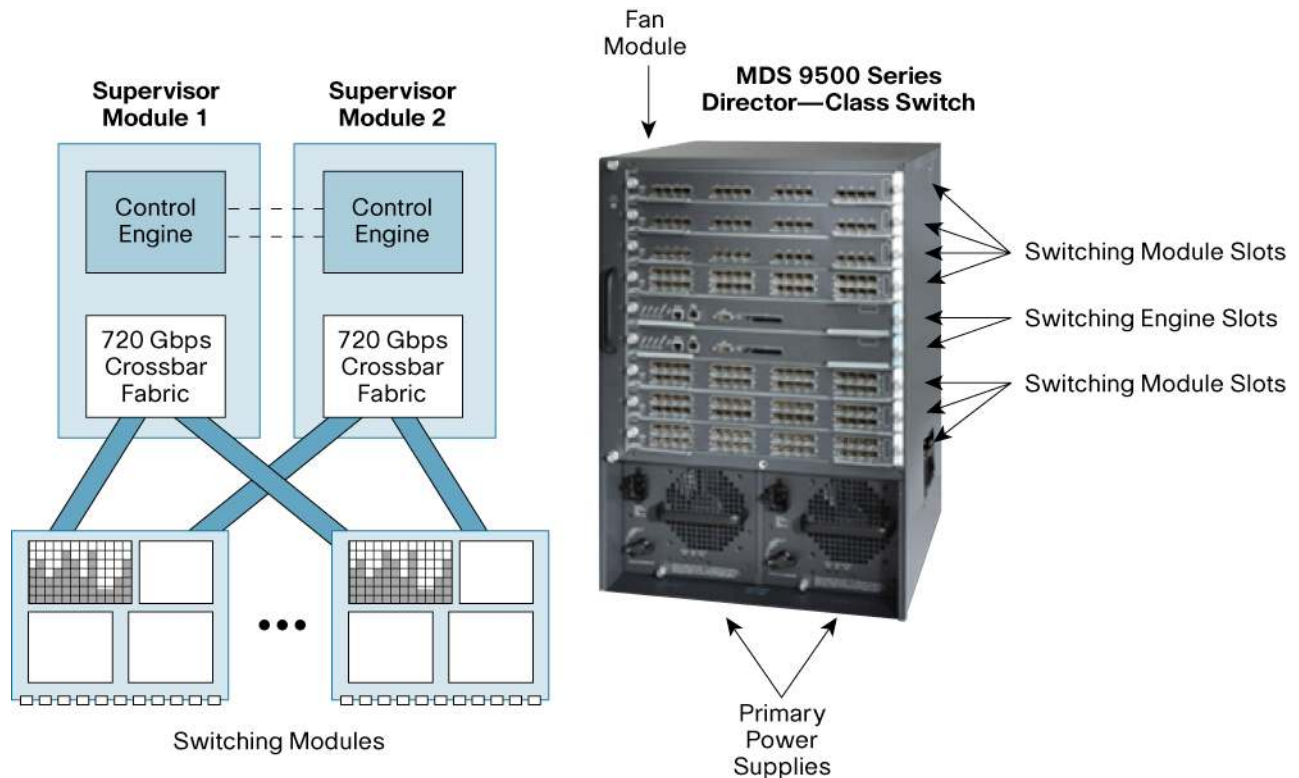
Cisco MDS 9500 Series Multilayer Directors support the ability to have two supervisor modules in the chassis for redundancy. Each supervisor module consists of a control engine and a crossbar fabric. The control engine is the central processor responsible for the management of the overall system. In 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 800 Gbps and serves 90 Gbps of bandwidth to each slot. Since each switching module of the Cisco MDS 9500 Series does not consume more than 90 Gbps of bandwidth to the crossbar, the system will operate at full performance even with one supervisor module. Therefore, in a fully populated Cisco MDS 9500 Series Multilayer Director, the system will not experience any disruption or any loss of performance with the removal or failure of one supervisor module. (See Figure 5)

**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 fully redundant, 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 1.2 Tbps and serves 90 Gbps of bandwidth to each slot. Since each switching module of the Cisco MDS 9500 Series does not consume more than 90 Gbps of bandwidth to the crossbar, the system will operate at full performance even with one crossbar module. Therefore, in a fully populated Cisco MDS 9513 Multilayer Director, the system will not experience any disruption or any loss of performance with the removal or failure of one crossbar module.

### 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 non-redundant component, the tray is designed in a 1+1 redundancy configuration. Each fan on the fan tray is monitored individually. If a fan failure occurs, the system will notify the end user of the fan failure. However, the system can sustain a multiple-fan failure with no negative effect. Within a normal operating environment, up to four fans can fail before the system is affected. The entire fan tray is hot swappable, and the system can run for up to 30 minutes without the fan tray installed. This allows for time to replace a fan tray while the system is functional.

## 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.

### Non-Disruptive Software Upgrade

Planned downtime is a large percentage of equipment downtime per year. A common reason for planned downtime is to upgrade software in the networking devices. These upgrades might be 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 back 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

Many 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 are the following:

- **High availability**—A common practice is to build multiple parallel fabrics and “multihomed” 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 dedicate 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 each of these reasons is a valid reason for building out separate fabrics, doing so can become quite wasteful. The prospect of additional separate fabrics means more hardware, more money spent, and typically underutilized hardware.

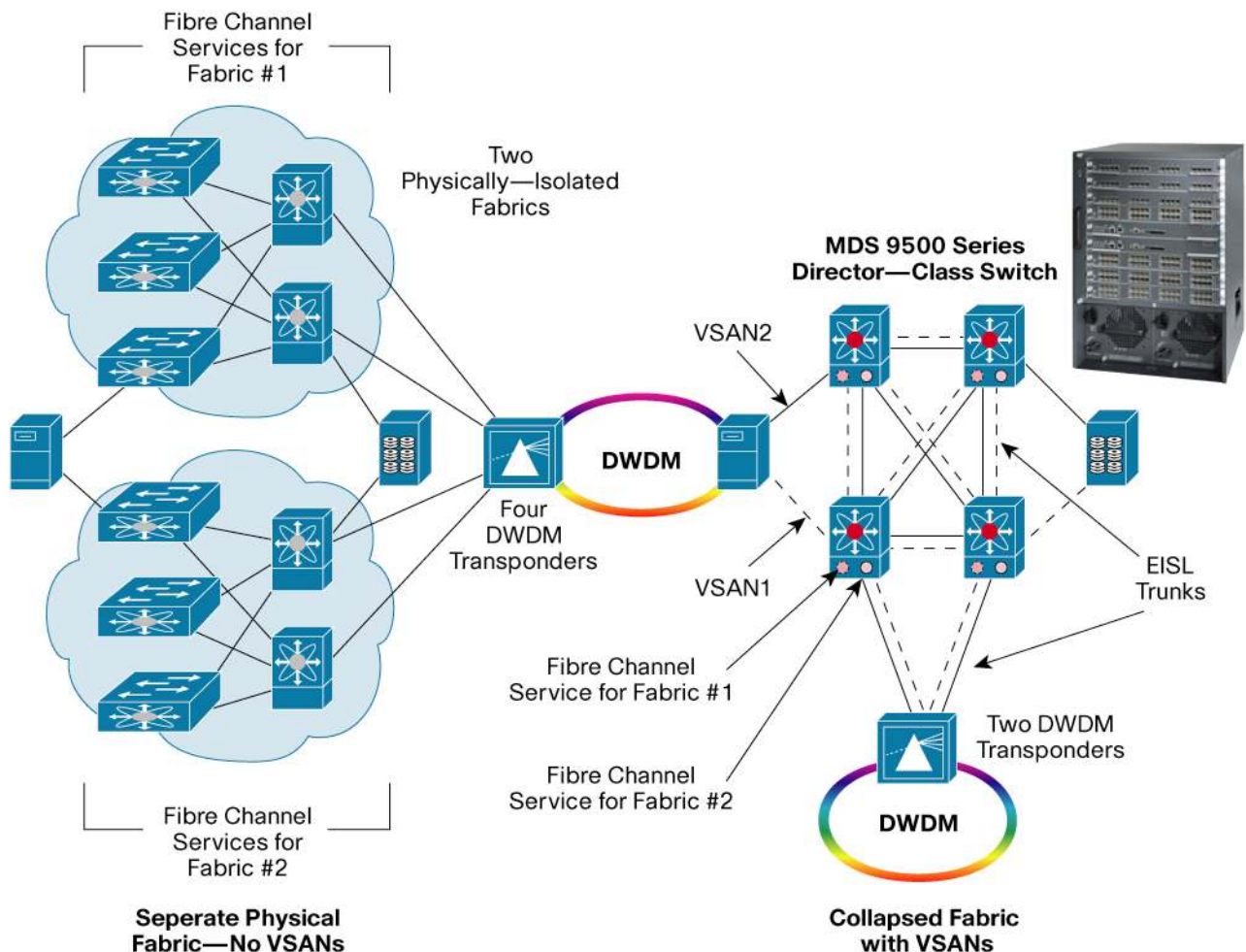
To help achieve the same isolated environments while eliminating the added expense of building physically separate fabrics, Cisco has introduced the VSAN within the Cisco MDS 9000 family. A VSAN provides the ability to create separate virtual fabrics on top of the same physical infrastructure. Each separate virtual fabric is isolated from the others using a hardware-based frame-tagging mechanism on ISLs. An EISL is an enhanced ISL that includes added tagging information for each frame and is supported on links interconnecting any Cisco MDS 9000 family switch product. Membership in a VSAN is based on 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 of products supports 1024 VSANs per physical infrastructure. Each VSAN can be selectively added to or pruned from an EISL to control the VSAN's reach. In addition, special traffic counters are provided to track statistics per VSAN.

Probably the most highly desired characteristic is the high-availability profile of VSANs. Not only do VSANs provide strict hardware isolation, but also a full replicated set of Fibre Channel services is created for each new VSAN. Therefore, 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 replica 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 across transports such as 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. VSANs bring scalability to a new level by using a common redundant physical infrastructure to build flexible isolated fabrics to achieve high-availability goals. (See Figure 6)

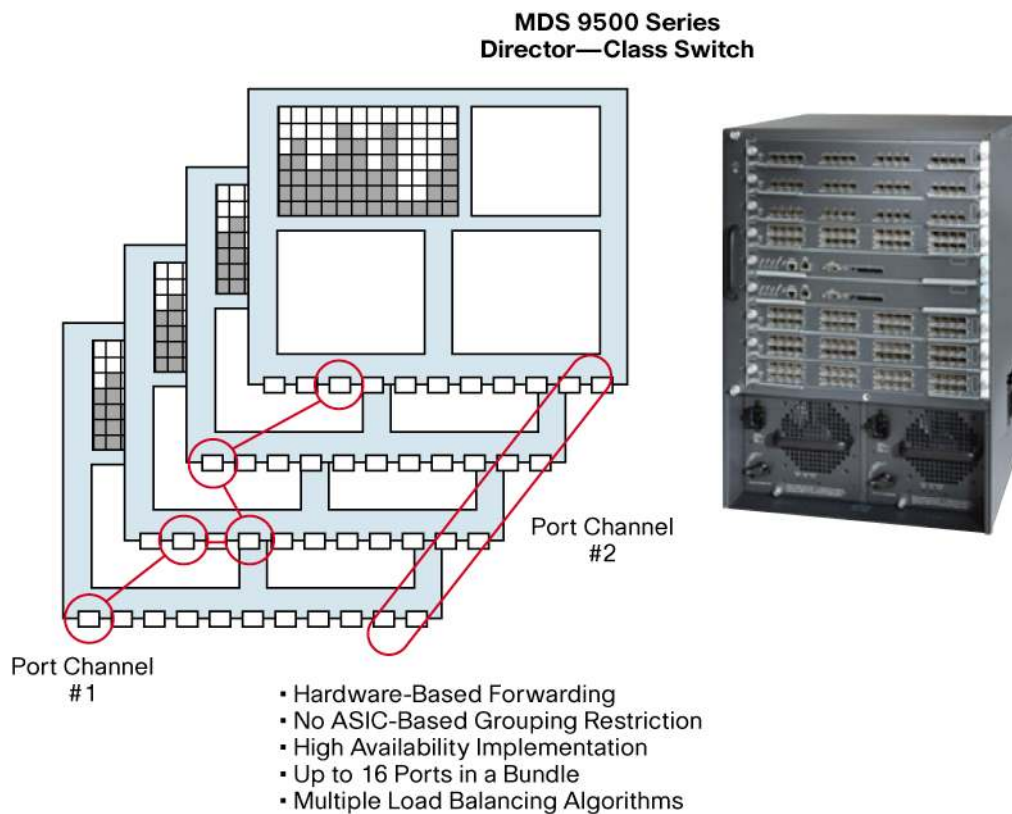
**Figure 6.** Using VSANs to Reduce SAN Complexity



## 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. 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. (See Figure 7)

**Figure 7.** Port Channeling in the Cisco MDS 9500 Series



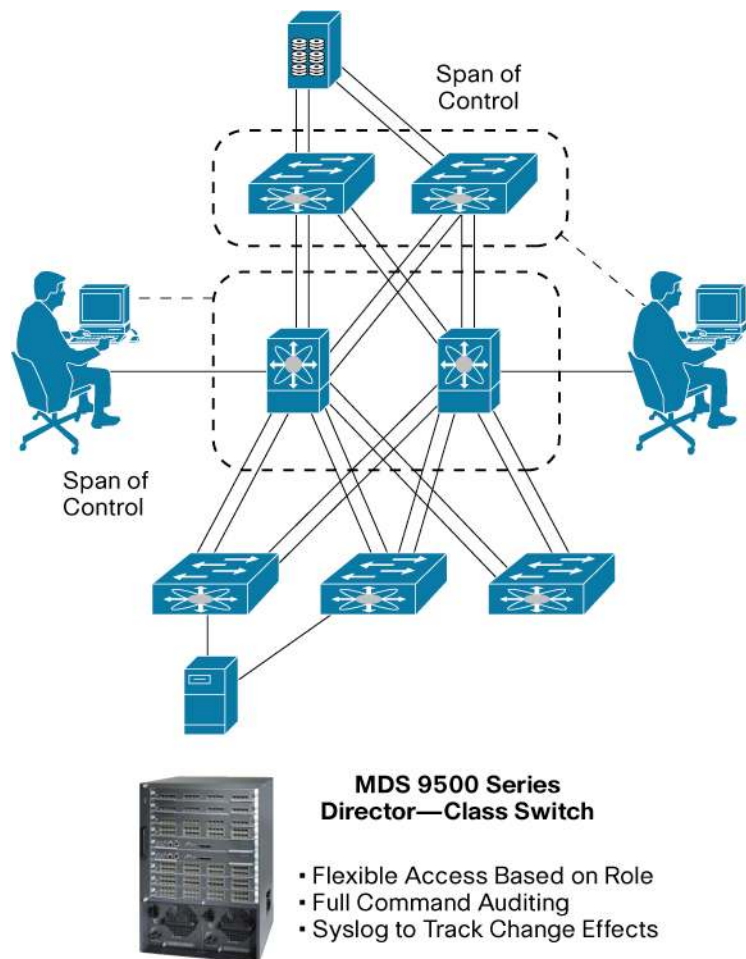
Cisco MDS 9500 Series Multilayer Directors support two different load-balancing algorithms across PortChannels. The first algorithm looks at the source and destination FCIDs of frames prior to entering the PortChannel. A hash is created in hardware from the source and destination FCID within the frame that serves as an index as to what 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 often not a consideration relating to high availability. However, one of the leading causes of downtime is human error. A user might mistakenly carry out a command without fully realizing the results of that command. The Cisco MDS 9000 Family of Multilayer Directors and Fabric Switches supports 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. 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. Therefore, one could 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. Therefore, a role could even be defined called “no debug fspf” that would allow 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. Two default roles, namely, network administrator (full access) and network operator (read-only access), are provided. Up to 64 custom roles can be defined by the user. Only a user within the network administrator role can create new roles. (See Figure 8)

**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.



### Corporate Headquarters

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

### European Headquarters

Cisco Systems International BV  
Haarlerbergpark  
Haarlerbergweg 13-19  
1101 CH Amsterdam  
The Netherlands  
www-europe.cisco.com  
Tel: 31 0 20 357 1000  
Fax: 31 0 20 357 1100

### Americas Headquarters

Cisco Systems, Inc.  
170 West Tasman Drive  
San Jose, CA 95134-1706  
USA  
www.cisco.com  
Tel: 408 526-7660  
Fax: 408 527-0883

### Asia Pacific Headquarters

Cisco Systems, Inc.  
168 Robinson Road  
#28-01 Capital Tower  
Singapore 068912  
www.cisco.com  
Tel: +65 6317 7777  
Fax: +65 6317 7799

Cisco Systems has more than 200 offices in the following countries and regions. Addresses, phone numbers, and fax numbers are listed on **the Cisco Website at [www.cisco.com/go/offices](http://www.cisco.com/go/offices).**

Argentina • Australia • Austria • Belgium • Brazil • Bulgaria • Canada • Chile • China PRC • Colombia • Costa Rica • Croatia • Cyprus  
Czech Republic • Denmark • Dubai, UAE • Finland • France • Germany • Greece • Hong Kong SAR • Hungary • India • Indonesia • Ireland • Israel  
Italy • Japan • Korea • Luxembourg • Malaysia • Mexico • The Netherlands • New Zealand • Norway • Peru • Philippines • Poland • Portugal  
Puerto Rico • Romania • Russia • Saudi Arabia • Scotland • Singapore • Slovakia • Slovenia • South Africa • Spain • Sweden • Switzerland • Taiwan  
Thailand • Turkey • Ukraine • United Kingdom • United States • Venezuela • Vietnam • Zimbabwe

Copyright © 2006 Cisco Systems, Inc. All rights reserved. CCSP, CCVP, the Cisco Square Bridge logo, Follow Me Browsing, and StackWise are trademarks of Cisco Systems, Inc.; Changing the Way We Work, Live, Play, and Learn, and iQuick Study are service marks of Cisco Systems, Inc.; and Access Registrar, Aironet, BPX, Catalyst, CCDA, CCDP, CCIE, CCIP, CCNA, CCNP, Cisco, the Cisco Certified Internetwork Expert logo, Cisco IOS, Cisco Press, Cisco Systems, Cisco Systems Capital, the Cisco Systems logo, Cisco Unity, Enterprise/Solver, EtherChannel, EtherFast, EtherSwitch, Fast Step, FormShare, GigaDrive, GigaStack, HomeLink, Internet Quotient, IOS, IP/TV, iQ Expertise, the iQ logo, iQ Net Readiness Scorecard, LightStream, Linksys, MeetingPlace, MGX, the Networkers logo, Networking Academy, Network Registrar, Packet, PIX, Post-Routing, Pre-Routing, ProConnect, RateMUX, ScriptShare, SlideCast, SMARTnet, The Fastest Way to Increase Your Internet Quotient, and TransPath are registered trademarks of Cisco Systems, Inc. and/or its affiliates in the United States and certain other countries.

All other trademarks mentioned in this document or Website are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (0601R)

