White Paper

Design a Reliable and Highly Available Fibre Channel SAN

Author:

Fausto Vaninetti, EMEAR Consultant Systems Engineer, Datacenter Technologies

Contributors:

Paresh Gupta, Technical Marketing Engineer

Ed Mazurek, Technical Leader, Services

Pete Rotella, Technical Leader, Engineering

Shi-Jie Wen, Distinguished Engineer, Supply Chain

Contents

What You Will Learn

This document explains how to design highly available Fibre Channel networks. Such a design requires switches with an appropriate hardware design architecture, a solid software implementation, a careful selection of fabric topology, and adherence to implementation best practices. The document explains the business need for high availability and examines the basic concepts of reliability engineering. It also discusses the main differences between switches and directors. The intrinsically superior architecture of the Cisco® MDS 9700 Series of 32-Gbps Fibre Channel–capable Multilayer Data Switches is described and contrasted with implementations of older architectures to demonstrate the importance of mission-critical directors. The MDS 9700 Series directors are the first in the industry to provide the arbitrated multifabric nonblocking architecture with true N+1 redundancy that is required for best-in-class availability. Custom application-specific integrated circuits (ASICs), meticulous component selection, smart software design, a comprehensive management suite, powerful yet simple-to-use diagnostics and monitoring tools, and enhanced serviceability features all contribute to the success of the MDS 9700 Series. This document also presents recommended Fibre Channel fabric topologies and best practices for interconnecting networking devices to achieve a highly available implementation. An appendix is also offered to the reader for a deeper explanation of reliability engineering concepts.

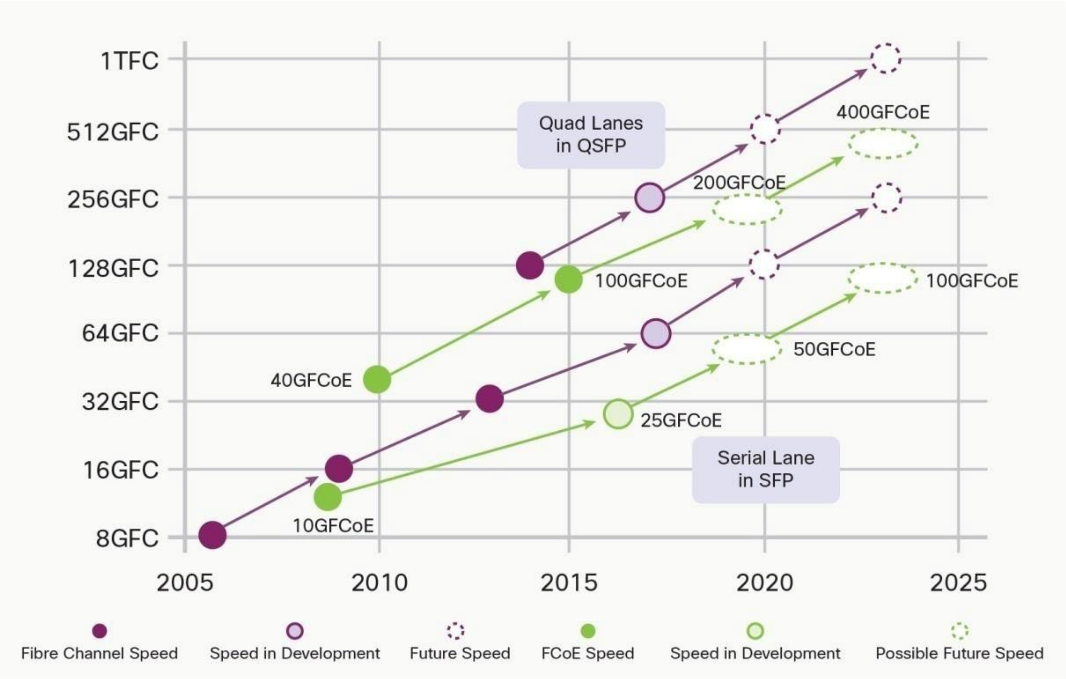
Business Need for High Availability

The main purpose of data centers is to provide the resources required to run applications and securely store data. As enterprises of every size move to digitize their businesses and adapt to new cloud-native workloads while preserving support for well-established applications, they need to rely on a storage network with no compromises. The advent of a worldwide economy, facilitated by the Internet and mobile devices, has shifted normal operations from the regular workday to a 24-hour model. In this always-on environment, more stringent requirements have been placed on high availability. With the modern expectation that businesses run 24 hours a day every day, IT services must always be available. Connecting to a favorite application and getting an error message is not acceptable. People now expect all IT services to be available at all times from any location and from any device, just as people expect electricity to always be available in their homes. Not surprisingly, business continuity is a top concern for chief information officers (CIOs), because IT downtime can have catastrophic consequences for a business. Both the loss of data and the inability to access that data can be extremely costly.

Most enterprise applications rely on relational databases and block storage to host their data. In this context, Fibre Channel networking devices are the preferred choice for connecting computational resources to data repositories in the form of disk arrays and tape libraries. Even today, with a variety of file- and object-based storage solutions on the market and the block-based alternative offered by Small Computer System Interface over IP (iSCSI) technology, almost all financial institutions and most Fortune 500 companies rely on Fibre Channel, sometime in combination with its derivative Fibre Channel over Ethernet (FCoE) protocol, as their trusted storage networking infrastructure, capable of functioning with almost no downtime. The cost of downtime varies by industry, but in all cases it can be high. An impact of US$300,000 per hour has been estimated for some organizations, but lost revenue, reduced employee productivity, customer dissatisfaction, and brand reputation dilution can increase this number tenfold. In today's enterprise environments, network high availability is no longer optional. Data must be available at all times.

Fibre Channel is a fast and reliable data transport technology and can scale to meet the requirements of small and midsize businesses as well as large enterprises. Today, installations range from small systems based on 12 port switches to very large deployments linking thousands of users, servers, and storage arrays with a switched Fibre Channel SAN. The success of this technology is demonstrated by the numbers. According to estimates presented in 2014 by the Storage Networking Industry Association (SNIA), more than 11 million Fibre Channel ports are sold every year and more than 11 exabytes of Fibre Channel storage are shipped each year, for an investment of more than US$11 billion in Fibre Channel enterprise-class storage systems. Moreover, Fibre Channel technology has a clear roadmap, as shown in Figure 1, with assured backward compatibility.

Figure 1. Fibre Channel Roadmap According to Fibre Channel Industry Association (FCIA)



Enterprise data centers have long invested in Fibre Channel storage and networks to help ensure that the requirements of their mission-critical applications are met with highly available solutions with low latency, high bandwidth, and deterministic performance. These features are especially important for organizations using densely virtualized servers and modern flash-based storage systems. The backward compatibility of any new generation of Fibre Channel speed, as required by INCITS T11 standards, and the flexibility of this technology in supporting multiple use cases and topologies, has further promoted widespread adoption. Fibre Channel can easily be integrated into existing (brownfield) data centers and uses operation best practices derived from Information Technology Infrastructure Library (ITIL) recommendations.

One reason for the success of Fibre Channel technology is the exceptional level of high availability it can deliver. With a nearly 20-year history of performance, stability, and high availability, the Fibre Channel protocol is a top consideration in organizations of all sizes when they have to design a storage network. With enterprises starting their journey toward big data solutions and real-time analytics, which are the new business enablers for the digital economy, fast and reliable access to structured data records and storage volumes is of the highest importance on both traditional systems and all-flash disk arrays.

The starting point in this journey is a switching product that offers highly reliable hardware and software. In addition, comprehensive and easy-to-use management tools have to support administrators in the daily activities of running and operating the network. Powerful diagnostics and monitoring tools are needed to assess the health of the solution and quickly track problems and move to the remediation phase. Serviceability features demonstrate their value in this scenario. Also, multiple networking devices need to be appropriately connected in accordance with best practices to build larger networks, sometime spanning multiple data centers in business-continuity and disaster-recovery scenarios.

This document explores the architectural differences among switches, directors, and mission-critical directors. The Cisco MDS 9700 Series of 32-Gbps Fibre Channel–capable storage networking products is described, with details about the series' industry-leading arbitrated multifabric nonblocking architecture with true N+1 redundancy. This document also presents some considerations related to fabric-level design best practices so that network downtime can be reduced or, ideally, avoided. The appendix explains the basic concepts of reliability and availability in the context of Fibre Channel SANs and provides guidance for calculating system availability and downtime. Hardware reliability and software reliability are described together with relevant terminology, such as failure rate, failures in time (FITs), mean time between failure (MTBF), mean time between repair (MTBR), thousands of lines of code (KLOC), in-service software upgrade (ISSU), and fault domain.

Reliability, Availability, and Serviceability

Reliability, availability, and serviceability (RAS) is a set of related attributes that must be considered when designing, manufacturing, purchasing, or using a network device. Reliability refers to the ability to consistently perform according to specifications. Availability is the ratio of the time that a system is functional to the total time it is expected to function. This ratio is often expressed as a percentage (for example, 99.99 percent). Serviceability refers to the ease with which a component, device, or system can be maintained and repaired. Early detection of potential problems therefore is critical so that the time to restore the service can be reduced.

The most important aspect of nearly every network is availability. Performance, scalability, management, agility, and other features are relevant only if the network is online. Availability contributes to system cost, possibly more than bandwidth, but compromises should not be accepted. Too often IT managers are judged by their short-term performance. Consequently, they may choose a system based only on its initial cost and without regard to its real long-term cost, inclusive of the unavoidable damage from the lack of availability.

Availability may seem a simple concept, but it is not. It must be carefully defined, and it is not easy to measure and benchmark. Most IT professionals today accept that availability is not simply the proportion of time that a system is online. Availability also has to be measured from the user's perspective. If a network fault denies access to 15 percent of the users, the system is technically available, but not for those affected users. If a real-time application is expected to provide response times of less than a second but is currently responding in 10 seconds, users may consider it unusable and therefore down. The concept of partial operational availability, in contrast to full operational availability, is sometimes helpful in describing these ambiguous situations. When assessing aggregate network bandwidth for any individual network device, a reduction of more than 15 percent is typically assumed to be equivalent to an out-of-service condition.

Switches, Directors, and Mission-Critical Directors

Fibre Channel networking devices are generally categorized into three classes according to the system architecture and embedded redundancy of critical components:

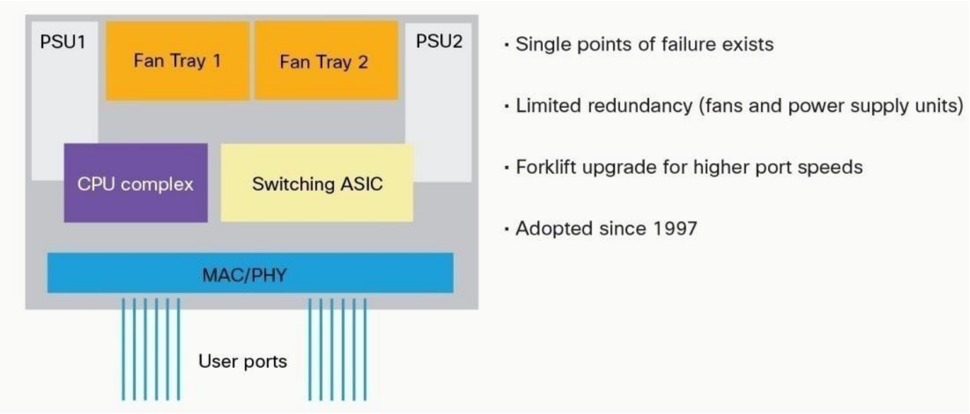
- Fabric switches
- Directors (also called director-class switches)
- Mission-critical directors

Moving from one class to the next, customers get higher port count, greater scalability for the entire fabric, multiprotocol support, investment protection and greater availability. Both Cisco and other vendors sell Fibre Channel fabric switches. However, for modular devices, Cisco has recently decided to focus on mission-critical directors, whereas other vendors are still selling directors only. Each product category has different main attributes.

Fabric Switches

Fabric switches have a single control plane and a non-redundant data plane. These switches are governed by a single CPU complex, and its failure makes the device unmanageable and compromises its function. These switches do not provide redundancy in the data plane either. In many cases, the entire data plane is built around a single ASIC, whose function is critical to the entire device. In some cases, an additional ASIC is used to expand the port count or add specialized capabilities. The only redundant components of fabric switches are the power supply units (PSUs) and fans, and today all vendors offer such components, at least as options. This architecture is shown in Figure 2.

Figure 2. Internal Architecture of a Fibre Channel Fabric Switch



- Single points of failure exists
- Limited redundancy (fans and power supply units)
- Forklift upgrade for higher port speeds
- Adopted since 1997

In most cases, fabric switches come in a predefined, fixed configuration, even if port utilization can start low and be expanded over time by the purchase of additional port licenses. Recent switches have a port range from 12 to 96 ports. Beyond that, a modular device is needed. In some cases, additional connectivity options and capabilities have been added to Fibre Channel switches, and in that case the switches are referred to as multiservice switches. For example, some switches offer FCoE ports for unified storage connectivity; Fibre Channel over IP (FCIP) ports for SAN extension applications; and compression, acceleration, and encryption capabilities. These additional features are enabled on specific ASICs, so the data plane becomes more complex than with a single ASIC. In all cases, a failure of any component of the switch cannot be repaired onsite, and the entire switch needs to be replaced. Consequently, a failure in a switch implies some downtime. This downtime has a negative effect on switch availability, and product data sheets tend to avoid reporting this value.

Directors

Directors are more resilient and scalable than fixed SAN switches. Directors are modular platforms whose port count can be scaled through the addition of connectivity line cards. Typically, a line card offers 48 ports, but in the past fewer ports were used. Depending on the chassis size and number of hosted line cards, a director today can include from 48 to 384 or even 528 ports.

One important benefit of a modular device is the capability to expand your connectivity options by adding specific line cards, choosing from cards that offer a variety of interfaces and different generations of Fibre Channel ASICs. For example, one line card might provide 48 ports of 8 Gbps Fibre Channel, another line card might provide 48 ports at 16 Gbps Fibre Channel, a third line card might provide multiple ports supporting 10-Gbps FCoE, and a fourth line card might enable 1- and 10-Gbps FCIP ports. Also, a director doesn’t just allow different types of protocols and line cards. It also allows all ports to communicate with each other through the backplane.

Directors offer greater availability than fabric switches, normally reported to be about 99.999 percent (also referred to as five-nines availability). The improved availability is not really associated with a higher quality of individual components. Instead it derives from the overall architectural design and duplication of components, so that no single point of failure exists. Clearly, for complete duplication, directors need to be equipped with at least two line cards. Just like switches, they come with redundant power supplies and fans. In most cases, more than one power supply is needed to meet the system power requirements. Because of this, the power supply redundancy schema used is the N+N model with load sharing. This approach is often referred to as grid redundancy because every set of independent power supplies is connected to a different electrical grid: one for each electricity provider. Moreover, directors have a dual control plane. This is represented by a pair of supervisor units working in active-standby mode and constantly synchronized.

Just like with line cards, the supervisor unit is a hot-swappable module. If a failure occurs, the supervisor unit can be removed and replaced without disrupting the rest of the system.

Additionally, directors provide functional duplication of the data plane. All line cards are cross-connected to the others through two dedicated crossbar fabric units (or equivalent modules for shared-memory designs), often mounted on the back of the chassis. Upon failure of one crossbar fabric unit, the director remains functional, but bandwidth drops by half. In this scenario, the director clearly offers only partial availability, and IT managers are increasingly treating such a scenario as a complete system failure.

The overall director architecture with duplicated crossbar fabric modules was developed in 1995 for Ethernet devices and ported to Fibre Channel devices in 2002. Despite being based on a consolidated and field-proven architecture, this approach is now yielding to a more modern approach.

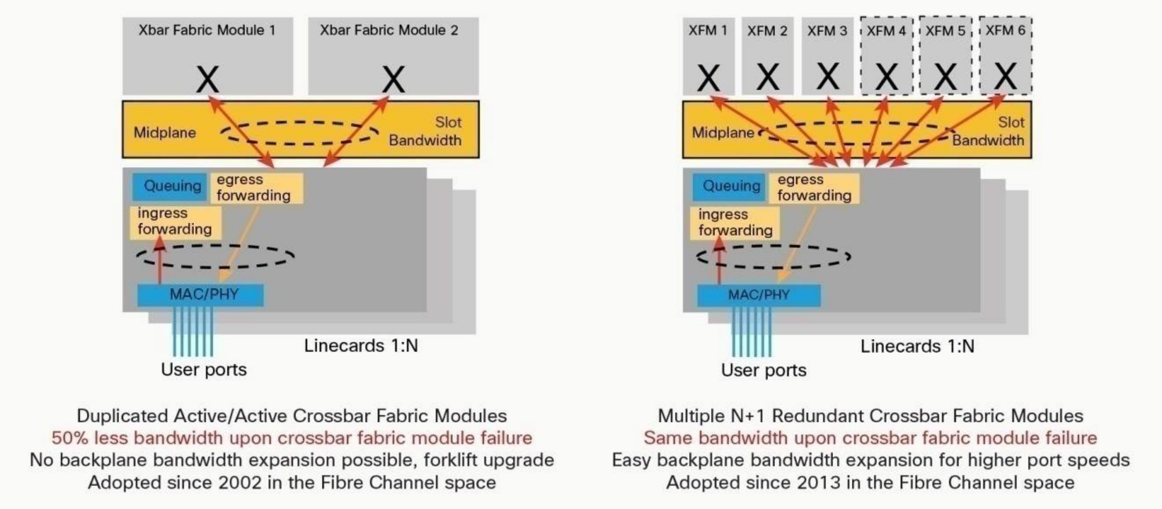
Mission-Critical Directors

Mission-critical directors are the latest enhancement to Fibre Channel devices. They provide all the capabilities of directors with the addition of some benefits derived from their enhanced architecture. This modern approach to modular network devices for data center was introduced in 2008, initially for Ethernet switches. In 2013 Cisco extended the concept to Fibre Channel devices. Today almost all vendors of modular data center switches, regardless of the protocol they serve, have adopted this superior architecture. Considerable investment in research and development has resulted in a truly redundant arbitrated crossbar multifabric architecture. The arbitration element makes traffic flow predictable and deterministic, avoiding congestion and head-of-line blocking even under high traffic load. True redundancy was achieved by building on the data-plane functional duplication of directors, bolstering it with elements to prevent performance degradation after a failure.

Customer appreciation of the crossbar multifabric approach is apparent. For their mission-critical applications, customers understand the value of high availability and expect a mission-critical networking platform.

A mission-critical director is similar to a director. It has N+N redundant power supplies and fans and a dual control plane. However, it improves uptime by offering N+1 redundancy for the data plane. This enhancement provides greater availability than the simple data-plane duplication found in directors. In fact, in a properly configured device, the failure of one crossbar fabric unit will not have any impact on the overall switching bandwidth available to line cards. Mission-critical directors thus provide full operational availability even in the event of a critical failure. In situations in which the N+1 redundancy of the data plane is not achieved, the crossbar multifabric architecture helps limit the bandwidth reduction resulting from a crossbar fabric module failure. Figure 3 shows the architectural difference between traditional directors and mission-critical directors.

Figure 3. Internal Architecture of Traditional Directors and Mission-Critical Directors


Xbar Fabric Module 1
Xbar Fabric Module 2
X
X
Midplane
Slot Bandwidth
Queuing
egress forwarding
ingress forwarding
MAC/PHY
Linecards 1:N
User ports
Duplicated Active/Active Crossbar Fabric Modules
50% less bandwidth upon crossbar fabric module failure
No backplane bandwidth expansion possible, forklift upgrade
Adopted since 2002 in the Fibre Channel space
XFM 1
XFM 2
XFM 3
XFM 4
XFM 5
XFM 6
X
X
X
X
X
X
Midplane
Slot Bandwidth
Queuing
egress forwarding
ingress forwarding
MAC/PHY
Linecards 1:N
User ports
Multiple N+1 Redundant Crossbar Fabric Modules
Same bandwidth upon crossbar fabric module failure
Easy backplane bandwidth expansion for higher port speeds
Adopted since 2013 in the Fibre Channel space

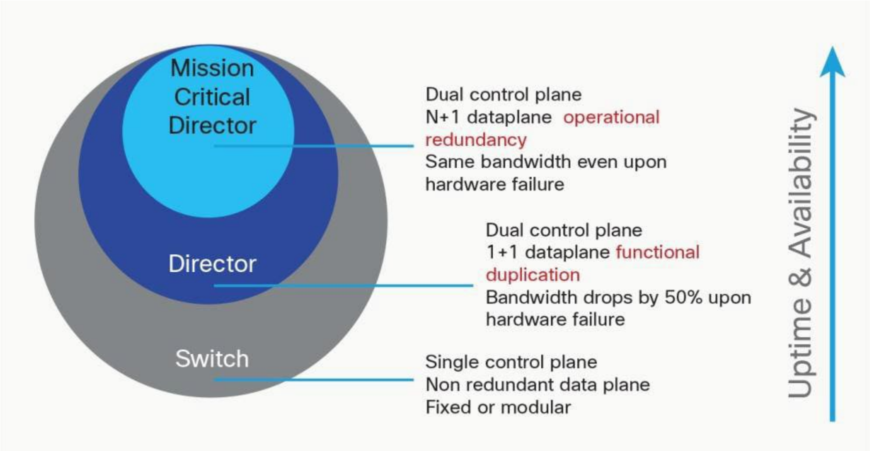
Classes of Fibre Channel Networking Devices and System Uptime

The concepts of single point of failure, functional duplication, and true operational redundancy can be understood with some analogies:

- In the human body, the heart is a single point of failure.
- A bird has two wings, which provide functional duplication. If the bird loses one wing, it cannot fly anymore. Therefore, two wings do not represent true redundancy. They instead represent just functional duplication.
- Airplanes have two engines. On take-off the plane needs both, so they are functionally duplicated, not redundant. When a plane is cruising, it can lose one engine and continue flying, so the two engines can be considered operationally redundant.

Fabric switches have a single point of failure, directors offer functional duplication, and mission-critical directors go beyond traditional directors to offer true operational redundancy under many deployment scenarios. As a result, the availability and uptime of these systems differ, as shown in Figure 4.

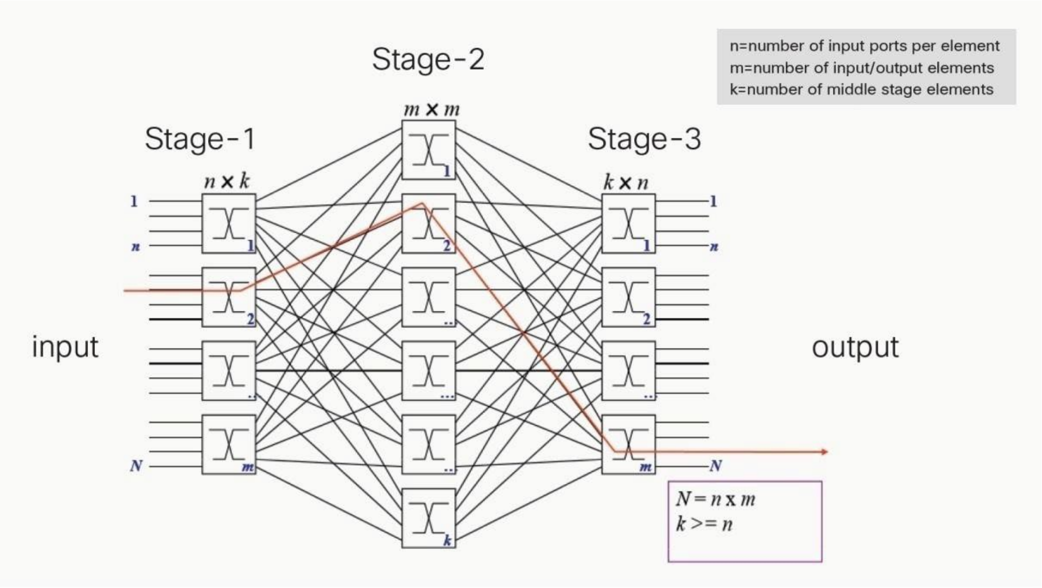
Figure 4. Classes of Fibre Channel Networking Devices and System Uptime


Mission Critical Director
Director
Switch
Dual control plane
N+1 dataplane operational redundancy
Same bandwidth even upon hardware failure
Dual control plane
1+1 dataplane functional duplication
Bandwidth drops by 50% upon hardware failure
Single control plane
Non redundant data plane
Fixed or modular
Uptime & Availability

Some Notes About Clos Architecture

One other notable detail in the architectural design of modular Fibre Channel networking devices is the specialization of crossbar fabric modules. The theory behind these elements has roots in the 1950s when Charles Clos introduced a technology to provide any-to-any communication within telephone networks. The term -fabric‖ was applied later because the pattern of links in this multistage switching architecture seemed like threads in woven cloth. Initially, Clos architectures were based on three stages, as shown in Figure 5: an ingress stage, a middle stage, and an egress stage. The concept was then extended to modular frame switches. Line cards implement the ingress and egress stages, and the crossbar fabric modules implement the middle stage. More recently, Clos architecture has been used to build entire networks (using a leaf-and-spine approach). By using sophisticated techniques of input and output queuing, arbitration, and more, a Clos fabric can be made non-blocking.

Figure 5. Three-Stage Clos Switching Network


n=number of input ports per element
m=number of input/output elements
k=number of middle stage elements
Stage-1
Stage-2
Stage-3
n x k
m x m
k x n
1
n
N
input
output
1
2
..
m
1
2
..
...
..
k
N = n x m
k >= n

In some older implementations, the middle-stage modules are built using the same ASICs as the line cards. This approach is described as a three-tier Clos mesh with switch-on-chip (SoC) technology. As a result, these generic ASICs could end up with more ports than necessary when they are deployed to interconnect the line cards among themselves. These stranded ports have been repurposed to interconnect directors, similar to the approach with standard Inter-Switch Links (ISLs). However, these interconnect links are proprietary, their reach is severely constrained, and they can be used to interconnect only directors of the same model from the same vendor. Both line cards and middle-stage modules include two or more identical ASICs. One implication of this approach is that the failure of such a middle-stage module affects both internal and external connectivity, with a 50 percent reduction in bandwidth. This design is often presented in marketing information as if more ports were available on the modular network device, but in reality it is just a reuse of wasted ASIC ports with several negative implications. Such an approach deviates from a pure Clos architecture, in which middle-stage switching elements should be used only to interconnect input and output elements, not to connect to the external world.

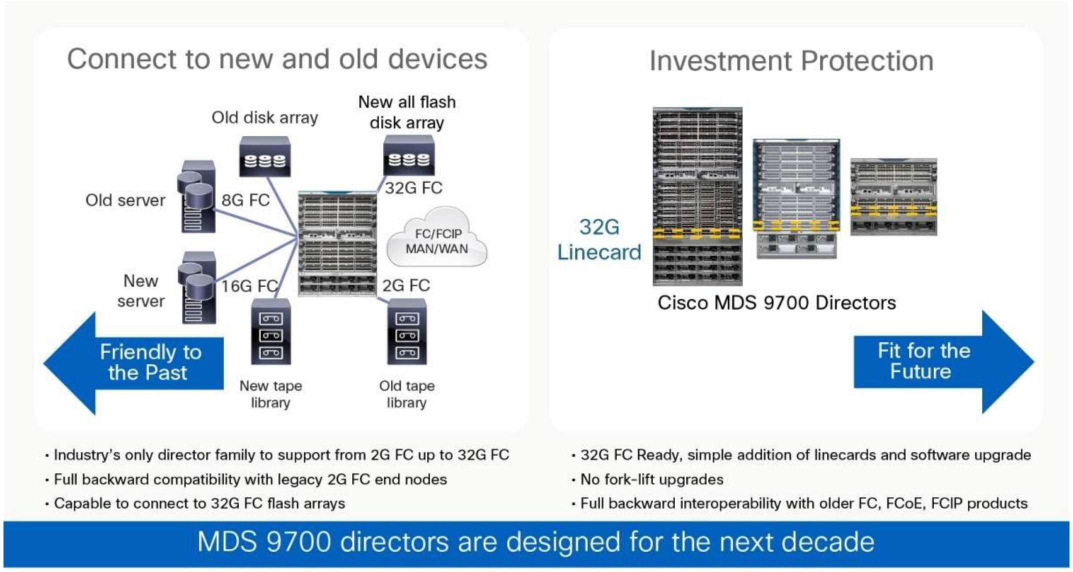
A more sophisticated approach was pioneered by Cisco more than a decade ago. In this case, the middle-stage modules use purpose-built specialized ASICs. Modern Cisco mission-critical directors also implement a crossbar multifabric design with specialized ASICs, with the middle stage solely dedicated to interconnecting line cards, with no proprietary interconnect links to other switching devices. Cisco believes that the SoC approach is best for fixed-configuration Fibre Channel fabric switches with relatively small port counts. Cisco considers this approach suboptimal for modular Fibre Channel directors and mission-critical directors, in which intelligent features and advanced functions are expected together with scalability and investment protection.

Cisco MDS 9700 Series Mission-Critical Directors

Cisco MDS 9000 Family storage networking devices have a long history of success, starting in 2002 with the first generation of products. The current generation of products include three fixed-configuration fabric switches (Cisco MDS 9148S and 9396S 16G Multilayer Fabric Switches and Cisco MDS 9250i Multiservice Fabric Switch) and three modular mission-critical directors (Cisco MDS 9706, 9710, and 9718 Multilayer Data Swicthes). Cisco MDS 9700 Series mission-critical directors can support 32-Gbps connectivity, and they are architecturally capable to support 128-Gbps Fibre Channel as well.

Despite pushing the limits toward higher port speeds, MDS 9700 Series mission-critical directors are also compatible with older equipment that customers may be reluctant to discard (Figure 6). In fact, MDS 9700 Series mission-critical directors are the only modular network devices in the industry that can provide port speeds starting from as low as 2-Gbps Fibre Channel and reaching all the way up to 32-Gbps Fibre Channel, with 128-Gbps connectivity technically possible with the same generation of ASICs.

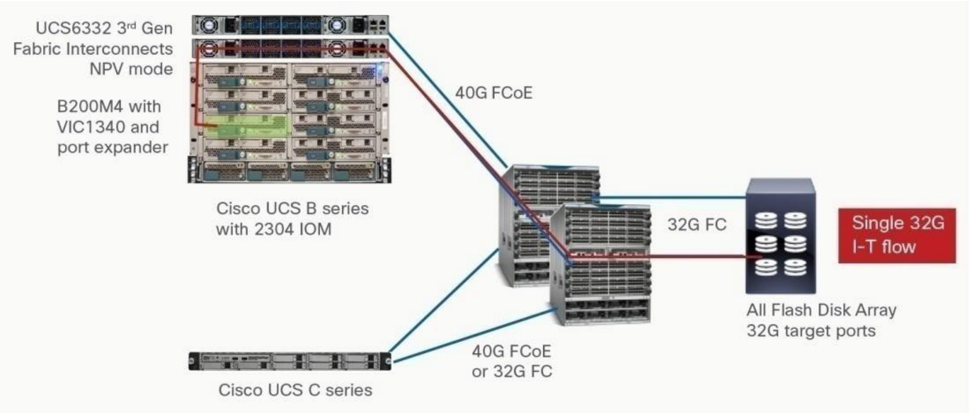
Figure 6. Cisco MDS 9700 Series Is Both Friendly to the Past and Fit for the Future



Connect to new and old devices
Old disk array
New all flash disk array
Old server
8G FC
32G FC
FC/FCIP MAN/WAN
New server
16G FC
2G FC
Friendly to the Past
New tape library
Old tape library
Investment Protection
32G Linecard
Cisco MDS 9700 Directors
Fit for the Future
• Industry's only director family to support from 2G FC up to 32G FC
• Full backward compatibility with legacy 2G FC end nodes
• Capable to connect to 32G FC flash arrays
• 32G FC Ready, simple addition of linecards and software upgrade
• No fork-lift upgrades
• Full backward interoperability with older FC, FCoE, FCIP products
MDS 9700 directors are designed for the next decade

Another unique benefit offered by the MDS 9700 Series is the flexibility of use and investment protection arising from its multiprotocol support. With Fibre Channel and FCoE capabilities combined in the same chassis, it is possible to establish initiator-to-target flows at 32 Gbps without sacrificing the consolidation advantages of unified fabric networking in blade servers. The deployment of Cisco Unified Computing System™ (Cisco UCS®) in conjunction with MDS 9700 Series mission-critical directors provides a powerful combination for organizations seeking extreme high-end performance when accessing their all-flash disk arrays (Figure 7).

Figure 7. Achieving Single Initiator–to–Target Flow at 32-Gbps Fibre Channel Speed



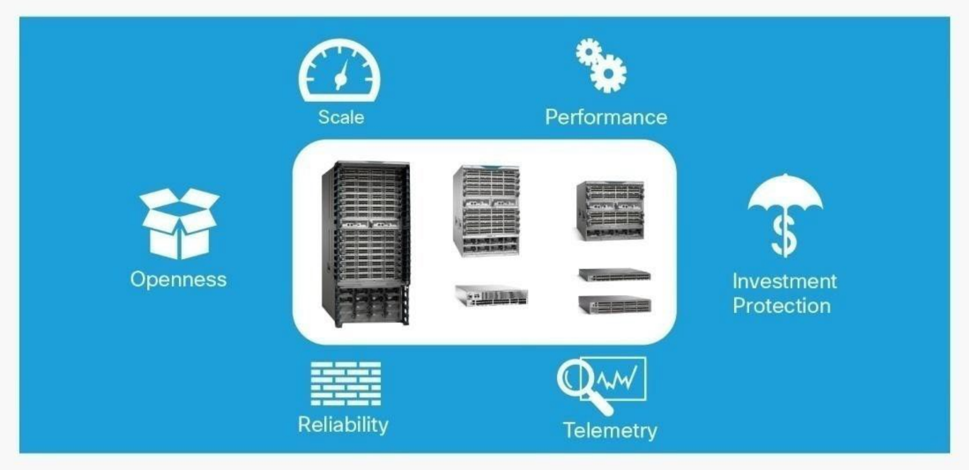
UCS6332 3rd Gen Fabric Interconnects NPV mode
B200M4 with VIC1340 and port expander
40G FCoE
Cisco UCS B series with 2304 IOM
32G FC
Single 32G I-T flow
All Flash Disk Array 32G target ports
40G FCoE or 32G FC
Cisco UCS C series

The multiprotocol support is further confirmed by the fact the MDS 9700 Series of mission-critical directors has no limit on the maximum number of 24/10-port SAN extension modules that can be hosted in a single chassis. This translates to an impressive 1.28 Tbps of FCIP throughput in a single device.

Cisco MDS 9000 Family Value Proposition

The value proposition of the MDS 9000 Family is based on several pillars, as shown in Figure 8. One of those pillars is reliability.

Figure 8. Cisco MDS 9000 Family Commitments



Both software and hardware reliability are critical to achieving a reliable storage networking device. At the foundation of hardware reliability are product design and component selection. Cisco designs and develops its own ASICs, bringing innovation, robust feature sets, and quality control to those crucial building blocks. Cisco also invests the resources and effort needed to select commodity components and suppliers to be sure that they comply with the highest quality standards in the industry.

Cisco's approach is all-encompassing, including product design, component selection, procurement, dual-vendor strategies, supply-chain management, component lifecycle management, reactive failure analyses, and proactive reliability monitoring. A specific highly accelerated life test (HALT) policy helps Cisco find and prevent manufacturing and design concerns at an early stage to avoid component failures in the field. Component- and system-level performance is evaluated beyond operational limits. Through the Cisco Extended Reliability Test (CERT) program, any reliability weaknesses are reported back to the engineering design team for correction. Electronic design-verification tests (EDVTs) and mechanical design-verification tests (MDVTs) are performed extensively on both operational and non-operational systems. The product quality engineers in the business unit in charge of product development and in the supply-chain organization have a complex set of decisions to make when selecting components. They take into account component specifications, component reliability history (both inside and outside Cisco), component lifecycle, and component cost. They also look at a supplier's on-time delivery capability and financial health, and even the geopolitical stability of the region in which the supplier manufactures the component. When possible, multiple suppliers are used.

Extreme care is also applied to internally developed ASICs to achieve the required capabilities and reliability outcomes, with every step fully controlled by Cisco engineering teams. Cisco products are designed using a well-refined product development methodology in association with a closed-loop corrective-action process. The engineers constantly monitor component quality and supplier performance to ensure that Cisco products can be produced in a manner that meets or exceeds customer expectations.

A similar approach is followed for software development. With an end goal of achieving both a robust feature set in a timely manner and code stability and reliability, Cisco uses modern development techniques coupled with strict quality-control policies, helping ensure that satisfaction remains high when customers deploy and use software associated with MDS 9000 Family products.

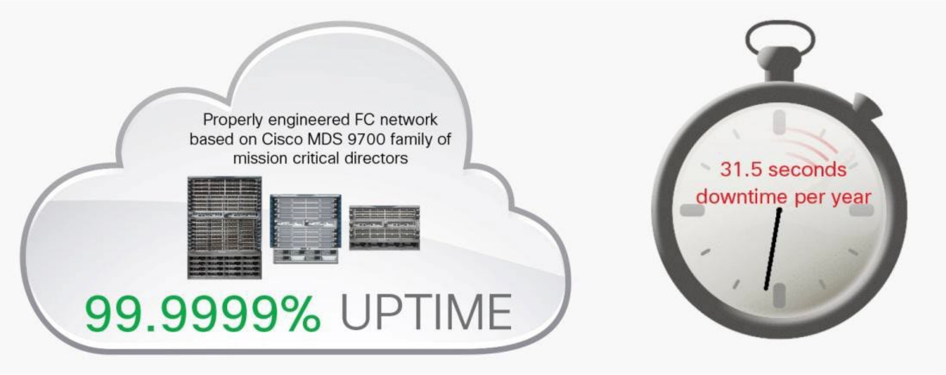
Six-Nines Uptime

According to a recent investigation by the Ponemon Institute published in a 2016 report, the average data center outage lasts 95 minutes, and the average cost per hour of unplanned downtime is US$540,000. For some companies in which the data center and data analysis are the main business, losses are even greater, in the range of a few million dollars. In fact, any downtime for IT services has a major impact on the financial health of any business. Organizations thus have increased their service-level agreement (SLA) requirements and expectations, which are today fundamental factors guiding the technological evolution of IT infrastructure, and storage networks in particular.

Several years ago, Cisco decided to stop developing traditional Fibre Channel directors and focus on mission-critical directors instead, capitalizing on Cisco's innovative technology and supporting the evolving market for greater availability. These new mission-critical directors are designed to meet the need for high availability, with reliability of individual components, software stability and self-recovery capabilities, diagnostic features, ease of use, automation, programmability, and troubleshooting tools. Repeat customer purchases of Cisco storage networking equipment demonstrate the high quality standard achieved by these products and the almost complete absence of service downtime when the directors are coupled with operational savviness and network design best practices.

The MDS 9000 Family products are purpose-built for storage networking and optimized for reliability. Five-nines availability (99.999 percent) has traditionally been the highest standard for reliable performance, but now Cisco mission-critical directors go beyond that and achieve six-nines availability. Figure 9 shows that with 99.9999 percent uptime, a properly engineered Fibre Channel network built using MDS 9700 Series mission-critical directors experiences only 31.5 seconds of unplanned downtime per year, meeting or exceeding the uptime requirements of today's most demanding environments.

Figure 9. Six-Nines Availability of a Properly Engineered Fibre Channel Network


Properly engineered FC network based on Cisco MDS 9700 family of mission critical directors
99.9999% UPTIME
31.5 seconds downtime per year

System Uptime Versus Kernel Uptime

On MDS 9700 Series mission-critical directors, system uptime refers to the time period starting when the chassis was powered on and had at least one supervisor module controlling its operation. Non-disruptive firmware upgrades and supervisor switchovers do not reinitialize the system uptime; hence, it keeps progressing across such events.

The kernel uptime refers to the time period starting when NX-OS was loaded on the supervisor module. An NX-OS upgrade will reset the kernel uptime, but not the system uptime.

The active supervisor uptime refers to the time period beginning when NX-OS was loaded on the active supervisor module. The active supervisor uptime can be less than the kernel uptime after a non-disruptive supervisor switchover.

The following example shows the system uptime for an MDS 9710 mission-critical director that was put in service in August 2013:

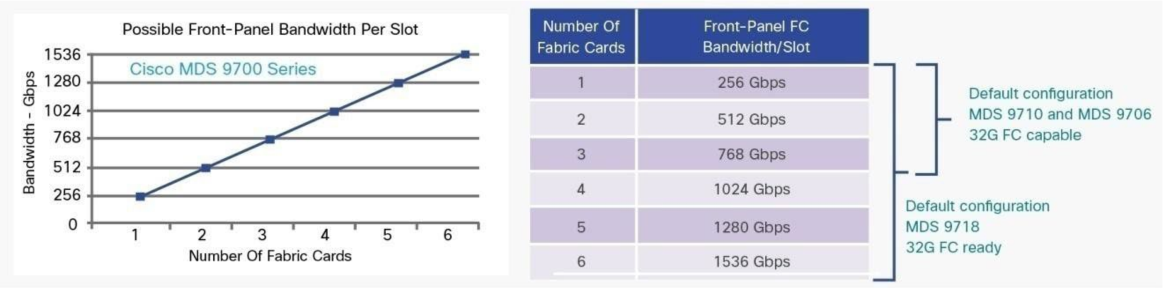
```
MDS9710-dirA# show system uptime
System start time: Fri Aug 24 09:00:02 2013
System uptime: 1201 days, 2 hours, 59 minutes, 9 seconds
Kernel uptime: 117 days, 1 hours, 22 minutes, 40 seconds
Active supervisor uptime: 117 days, 0 hours, 30 minutes, 32 seconds
```

N+1 Fabric Redundancy

MDS 9700 Family directors are the first in the industry to use the arbitrated crossbar multifabric architecture with true N+1 redundancy, offering exceptional high availability. The MDS 9710, introduced in 2013, was the first mission-critical director on the market providing true operational redundancy. The MDS 9706 and 9718 followed soon after, and have been built on the same architecture. All these mission-critical directors can host from three to six crossbar fabric modules in the back of their chassis. The last director developed, the MDS 9718, is sold only with all crossbar fabric modules inserted.

Each crossbar fabric module provides an equivalent of 256 Gbps of Fibre Channel front-panel bandwidth to each line card. With two crossbar fabric modules, enough bandwidth is available for wire-speed operation across all 48 ports on every line card when run at 8-Gbps Fibre Channel speed. By using three crossbar fabric modules, N+1 redundancy is achieved. Similarly, with 16-Gbps Fibre Channel operation, three crossbar fabric modules are needed to deliver the full bandwidth (768 Gbps) for all 48 ports on each line card, and four fabric modules provide N+1 redundancy. The same reasoning holds true for the 24/10-port SAN extension module. When the 24-port 40-Gbps FCoE line card is used, five crossbar fabric modules are required to achieve wire speed on all ports, and six crossbar fabric modules provide N+1 redundancy. Figure 10 summarizes these scenarios.

Figure 10. Bandwidth per Slot for Cisco MDS 9700 Series and Number of Crossbar Fabric Modules



Number Of Fabric Cards	Front-Panel FC Bandwidth/Slot
1	256 Gbps
2	512 Gbps
3	768 Gbps
4	1024 Gbps
5	1280 Gbps
6	1536 Gbps

Default configuration
MDS 9710 and MDS 9706
32G FC capable

Default configuration
MDS 9718
32G FC ready

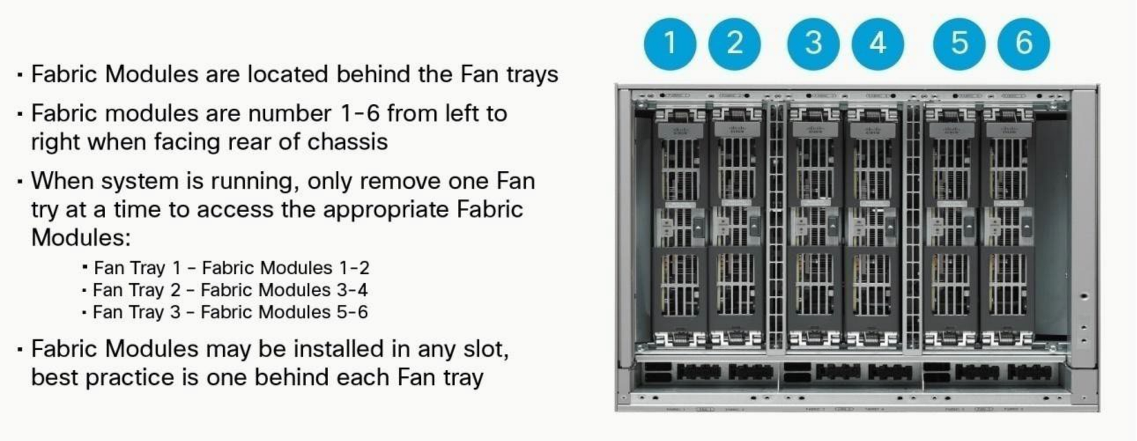
The modern design with multiple crossbar fabric modules provides two main benefits:

- True operational redundancy for higher uptime at full performance
- Easy scale-up of performance with module additions only, rather than replacement, leading to improved investment protection and no required downtime

For port speeds up to 40-Gbps for FCoE, the failure of one crossbar fabric module out of six has no effect whatsoever on the wire-speed performance of the mission-critical director. Moreover, depending on the port speed, an N+2 or N+3 redundancy schema can be deployed, if a customer should choose to do so.

Figure 11 shows the six crossbar fabric modules in the back of an MDS 9706 chassis. Crossbar fabric modules for the MDS 9710 and 9718 are taller but offer equivalent functions and internal circuitry.

Figure 11. Back of a Cisco MDS 9706 Chassis with Its Six Crossbar Fabric Modules Exposed



- Fabric Modules are located behind the Fan trays
- Fabric modules are number 1-6 from left to right when facing rear of chassis
- When system is running, only remove one Fan try at a time to access the appropriate Fabric Modules:
  - Fan Tray 1 – Fabric Modules 1-2
  - Fan Tray 2 – Fabric Modules 3-4
  - Fan Tray 3 – Fabric Modules 5-6
- Fabric Modules may be installed in any slot, best practice is one behind each Fan tray

The MDS 9700 Series is the industry’s first Fibre Channel modular platform to provide redundancy on all major infrastructure hardware components, including crossbar fabric modules, as detailed in Table 1.

Table 1. Redundancy Details for Cisco MDS 9700 Series Mission-Critical Directors

Component	Redundancy
Supervisor	1+1 active-standby
Power supply	N+1 and N+N grid redundancy
Crossbar fabric module	N+1 redundancy (typical)

Ease of Module Replacement

Ease of replacement of failed components was specifically considered in the design of the MDS 9700 Series mission-critical directors. Because availability improves when the time to restore service after a fault is short, all modules are designed to be hot swappable, and all include fast ejectors or handles and can be replaced in less than one minute when spare parts are available onsite. Modules are easily accommodated in available slots in the chassis and receive electricity from dual connectors hosted in the passive backplane. In this way, single points of failure are eliminated.

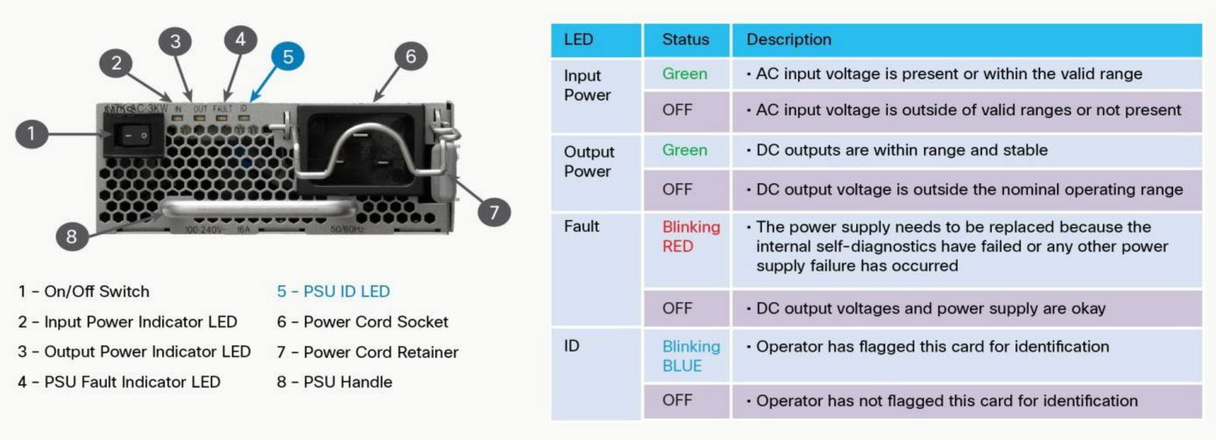
Locator ID on All Modules and Beacon LEDs

Before replacing a failed module, you need to identify it among its peers in the chassis and across different chassis in the data center. To help you, the MDS 9000 Family of Fibre Channel switches provides LEDs for easy identification of ports and modules. The beacon mode displays a flashing green light that helps you identify the port you are seeking. The locator ID LED helps you identify line cards, supervisors, power supplies, fans, or crossbar fabric units. The MDS 9700 Family is the only one to offers locator IDs for all system modules, not just some of them. The administrator can turn on the beacon mode or locator ID LED from the remote central management station, allowing the support engineer to quickly identify the component that requires attention. Enabling the beacon mode or locator ID LED has no effect on the operation of the interface or module.

Figure 12 shows the locator ID on the 3-kW AC power supply module. The locator ID LED for the power supply in slot 3 can be turned on from the Cisco command-line interface (CLI) by sending the following command:

```
switch(config)# locator-led powersupply 3
```

Figure 12. Locator ID LED on 3-kW AC Power Supply Module



1 - On/Off Switch	5 - PSU ID LED
2 - Input Power Indicator LED	6 - Power Cord Socket
3 - Output Power Indicator LED	7 - Power Cord Retainer
4 - PSU Fault Indicator LED	8 - PSU Handle

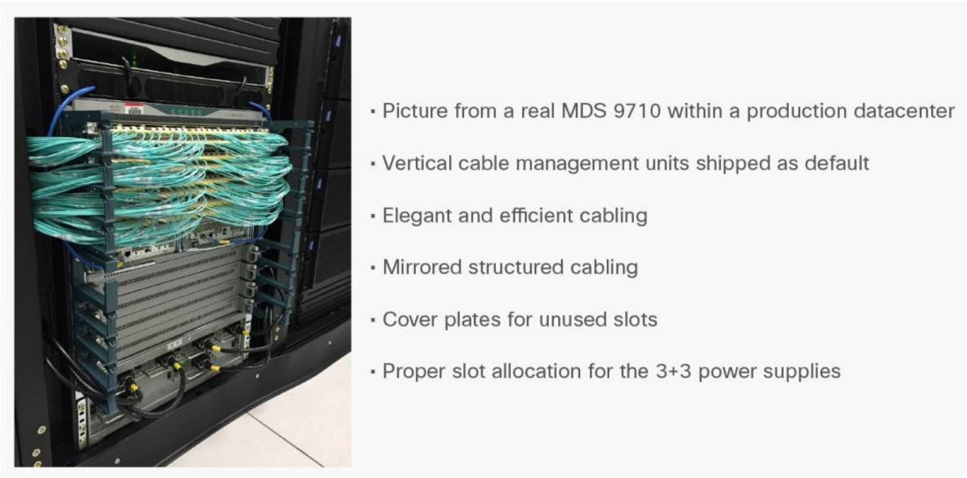
LED	Status	Description
Input Power	Green	• AC input voltage is present or within the valid range
	OFF	• AC input voltage is outside of valid ranges or not present
Output Power	Green	• DC outputs are within range and stable
	OFF	• DC output voltage is outside the nominal operating range
Fault	Blinking RED	• The power supply needs to be replaced because the internal self-diagnostics have failed or any other power supply failure has occurred
	OFF	• DC output voltages and power supply are okay
ID	Blinking BLUE	• Operator has flagged this card for identification
	OFF	• Operator has not flagged this card for identification

Racking and Cabling

The racking and cabling of networking equipment is a potential source of human errors that can negatively affect overall system availability. The challenge is greater with large modular devices because of their size and the quantity of cables they use. For this reason, MDS 9000 Family networking devices use standard Enhanced Small Form-Factor Pluggable (SFP+) and Quad SFP+ (QSFP+) form factors and LC fiber connectors. These offer the best combination of port density per rack unit and ease of cabling for individual ports. The use of proprietary form factors and custom cables to squeeze more ports into the same footprint would have negative consequences during the lifecycle of a solution.

The MDS 9700 Series also handles cables in a well-organized way, with vertical cable-management units to facilitate efficient cabling, as shown in Figure 13.

Figure 13. Cisco MDS 9710 Chassis with Its Vertical Cable Management Units



- Picture from a real MDS 9710 within a production datacenter
- Vertical cable management units shipped as default
- Elegant and efficient cabling
- Mirrored structured cabling
- Cover plates for unused slots
- Proper slot allocation for the 3+3 power supplies

In addition, cable management and module insertion are performed on the front of the chassis. This approach improves serviceability when you need to replace a failed module or insert a new line card to increase the port count.

Cooling and Energy Efficiency

Cooling is important to the proper functioning and life span of electronic components. The cooling system also needs to be flexible enough to accommodate possible differences in deployment locations. MDS 9148S switches are intended to be top-of-rack devices. They are designed to be housed inside racks of servers, and consequently their cooling system is optimized for that deployment mode, using port-side exhaust. The MDS 9396S switches can be used in top-of-rack or middle-of-row deployments. They therefore come with both port-side intake and port-side exhaust options, compatible with the cold-aisle, hot-aisle models used in most data centers.

In general, fabric switches can take a flexible approach to cooling because their lower port counts, component density, and power consumption do not impose excessive heat on the optical transceivers. But for modular devices, requirements are different. Large platforms, such as the MDS 9700 Series mission-critical directors, are typically hosted inside dedicated networking racks in middle-of-row or end-of-row deployments.

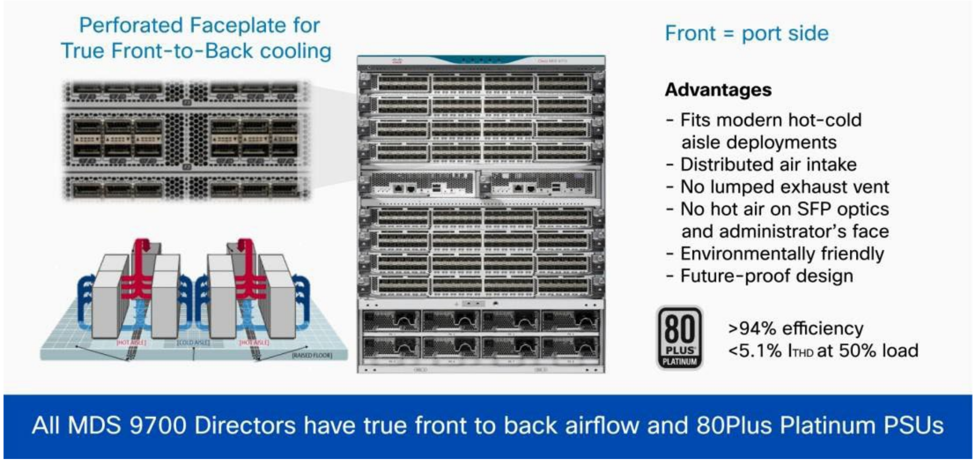
As a result, airflow direction is chosen to optimize the cooling and reliability of the device. All MDS 9700 Series mission-critical directors provide true front-to-back cooling. Cold air is taken in from the port side so that the SFP+ transceivers operate at a low temperature, helping eliminate random failures of these devices, which are sensitive to high operating temperatures. This front-to-back cooling direction also helps prevent air from blowing on the faces of operators when they make changes to port cables or perform some other maintenance activity.

The back of chassis hosts three fan trays, but if a fan try is removed for maintenance purposes or in the event of a failure, the chassis can safely operate with just two trays. A single fan tray contains redundant fans, so that a single malfunctioning fan does not lead to the failure of an entire fan tray.

Power supply units, or PSUs, are 80 Plus certified with a Platinum rating (> 94 percent efficiency), reducing heat waste and power consumption. They have a total harmonic distortion value for current (I_{THD}) of less than 5.1 percent when they are used at maximum efficiency.

Figure 14 summarizes the cooling features of the chassis and its power efficiency.

Figure 14. Cisco MDS 9710 Chassis with Front-to-Back Cooling



Perforated Faceplate for True Front-to-Back cooling

Front = port side

Advantages

- Fits modern hot-cold aisle deployments
- Distributed air intake
- No lumped exhaust vent
- No hot air on SFP optics and administrator's face
- Environmentally friendly
- Future-proof design

>94% efficiency
<5.1% I_{THD} at 50% load

All MDS 9700 Directors have true front to back airflow and 80Plus Platinum PSUs

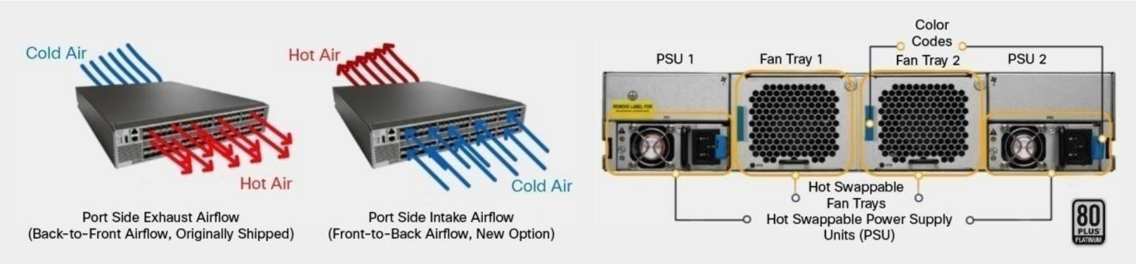
Color Coding of Modules

The MDS 9000 Family uses color coding in power supplies and fans where appropriate, so that airflow direction can easily be determined onsite. On MDS 9700 Series directors and MDS 9148S switches, color coding is not required because the airflow direction is fixed. But on MDS 9396S switches, which have reversible airflow options, color coding is helpful. Color coding is used on PSUs and fan trays to indicate the airflow direction. This color coding facilitates switch setup that conforms to the data center cooling methodology.

The airflow direction supported by PSUs and fan trays needs to match. Older PSUs supported only unidirectional airflow (port-side exhaust airflow) and were identified with a blue color. More recent PSUs, denoted with a white color on the side handle, support bidirectional airflow. The system automatically matches the airflow with that of the other PSU and system fans to avoid any conflicts. The MDS 9396S supports two fan trays with colored stickers on them: one with port-side intake airflow (red), and the other with port-side exhaust airflow (blue).

So the white-coded PSU supports bidirectional airflow (port-side exhaust or port-side intake), but fan trays can be either the port-side exhaust version (blue) or the port-side intake version (red). Figure 15 provides a graphical representation.

Figure 15. Color Coding on Cisco MDS 9396S



Cold Air
Hot Air
Port Side Exhaust Airflow
(Back-to-Front Airflow, Originally Shipped)
Hot Air
Cold Air
Port Side Intake Airflow
(Front-to-Back Airflow, New Option)
Color Codes
PSU 1
Fan Tray 1
Fan Tray 2
PSU 2
Hot Swappable Fan Trays
Hot Swappable Power Supply Units (PSU)
80 PLUS PLATINUM

Diagnostics and Troubleshooting Tools

The MDS 9700 Series is equipped with a large set of diagnostics and troubleshooting tools. Fibre Channel fabric connectivity requires multiple electrical and optical components to function correctly, including cables, transceivers, port ASICs, switching ASICs, and communication buses internal to the switches. If any of these components is faulty, it will affect I/O operations over the fabric.

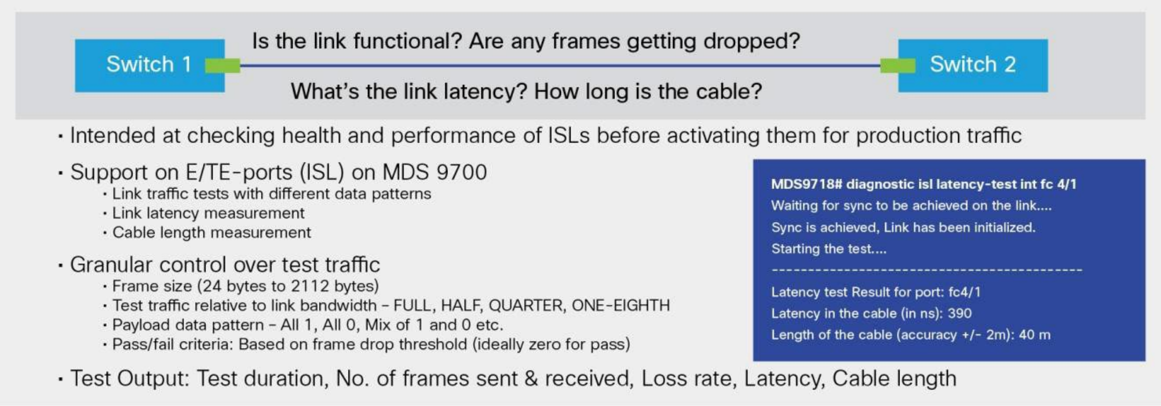
Today Fibre Channel is deployed in mission-critical networks, where resiliency and throughput are high-priority requirements. In such networks, early identification of any faults is critical to gaining customer trust. For this reason, the MDS 9700 Series provides a comprehensive set of system- and link-level diagnostics capabilities. This software-based suite of tools, hardware enabled for some tests, can dynamically verify whether everything is working as expected. Cisco Generic Online Diagnostics (GOLD) offers a complete suite of tests to verify that supervisor engines, switching modules, ASICs, communication buses, optics, and interconnections are functioning properly. GOLD tests can be run at initial system bootup, periodically at runtime, and on demand when invoked by the administrator.

The bootup diagnostics run during the bootup procedure and detect faulty hardware when a new networking device is brought online. These tests represent an evolution of the power-on self-test (POST) capabilities long present on MDS 9000 Family switches. They verify the checksum on the boot and switch firmware images, perform internal data loopback testing on all Fibre Channel ports, and perform access and integrity checks on management ports and nonvolatile memory components. During the diagnostics phase, the switch logs any errors encountered.

Runtime and on-demand tests are even more specific and implement health-monitoring diagnostics. Enabled by default, they verify the health of a live system at user-configurable periodic intervals. The health-monitoring diagnostic tests detect possible hardware errors and data-path problems without disrupting the data or control traffic.

ISL diagnostics are available to help check the health and performance of ISLs (E and TE ports) before the links are activated for production traffic, measuring frame round-trip latencies and cable lengths. Figure 16 shows the ISL diagnostics capability.

Figure 16. ISL Diagnostics Capability


Switch 1
Is the link functional? Are any frames getting dropped?
Switch 2
What’s the link latency? How long is the cable?

- Intended at checking health and performance of ISLs before activating them for production traffic
- Support on E/TE-ports (ISL) on MDS 9700
  - Link traffic tests with different data patterns
  - Link latency measurement
  - Cable length measurement
- Granular control over test traffic
  - Frame size (24 bytes to 2112 bytes)
  - Test traffic relative to link bandwidth – FULL, HALF, QUARTER, ONE-EIGHTH
  - Payload data pattern – All 1, All 0, Mix of 1 and 0 etc.
  - Pass/fail criteria: Based on frame drop threshold (ideally zero for pass)
- Test Output: Test duration, No. of frames sent & received, Loss rate, Latency, Cable length

```
MDS9718# diagnostic isl latency-test int fc 4/1
Waiting for sync to be achieved on the link....
Sync is achieved, Link has been initialized.
Starting the test....
--------------------------------------------
Latency test Result for port: fc4/1
Latency in the cable (in ns): 390
Length of the cable (accuracy +/- 2m): 40 m
```

Single-hop and multihop tests are also available for further analysis along the path, including link-saturation stress tests.

Host-to-switch connectivity (N and F ports) tests are also available to MDS 9000 Family customers as an enhancement to the current diagnostics suite. For host connectivity probing, the INCITS T11 FC-LS-3 standard refers to a specific implementation for beaconing the peer port for ease of identification. This solution is based on Link Cable Beacon Extended Link Service (LCB-ELS) with a read diagnostic parameters (RDP) command that is used to query N-port-related link- and port-level diagnostic parameters.

In addition to these intelligent diagnostics features, the MDS 9000 Family offers hardware-based slow-drain port detection and remediation capabilities that go well beyond the capabilities offered by competing products.

Data Integrity

Reliability is also related to how well a device performs its task. For a network switch, whose task is to interconnect nodes and allow them to communicate, introducing errors in the information exchange is not the expected behavior. For this reason, MDS 9000 Family switches implement an advanced feature set to help ensure data integrity on all data paths. To help ensure the reliable transport of data, MDS 9000 Family products use several error-detection mechanisms and provide error-correction capabilities whenever possible:

- Error detection and correction (ECC) on the supervisor memory
- Cyclic redundancy check (CRC) for frame integrity on the ingress port
- Internal CRC detection for frame integrity at the ingress port of the crossbar fabric module and the ingress port of the output line card, with automatic isolation of misbehaving components
- Automatic dropping of corrupted frames
- Forward error correction (FEC) on ISLs and F-ports

  Note: FEC can be considered a first line of defense with the capability to correct up to 11 bits out of a sequence of 2112 bits. If the corruption is more severe, MDS 9000 Family products do not let those frames flood the network.

- Syslog-based notifications to the administrator

Cisco Call Home and Smart Call Home

The Cisco Call Home feature is part of the suite of tools to accelerate problem resolution with real-time event notification to the relevant administrators. The Call Home feature provides email-based notification of critical system events triggered in software and hardware. It forwards to external entities both alarms and events, packaged with other relevant contextual information for better understanding. It can be thought of as a basic telemetry system, built on an event-based push model.

A versatile range of message formats are available for optimal compatibility with standard email services and XML-based automated parsing applications. Alert grouping capabilities and customizable destination profiles offer the flexibility needed to notify specific individuals or support organizations only when necessary. External entities can include, but are not restricted to, an administrator’s email account or pager, an in-house server, or a server at a service provider’s facility. The Call Home feature also provides message throttling capabilities. The list of deliverable Call Home messages also includes periodic inventory messages, port syslog messages, and remote monitor (RMON) alert messages.

If required, you can use the Cisco Fabric Services application to distribute the Call Home configuration to all other switches in the fabric to prevent inconsistencies.

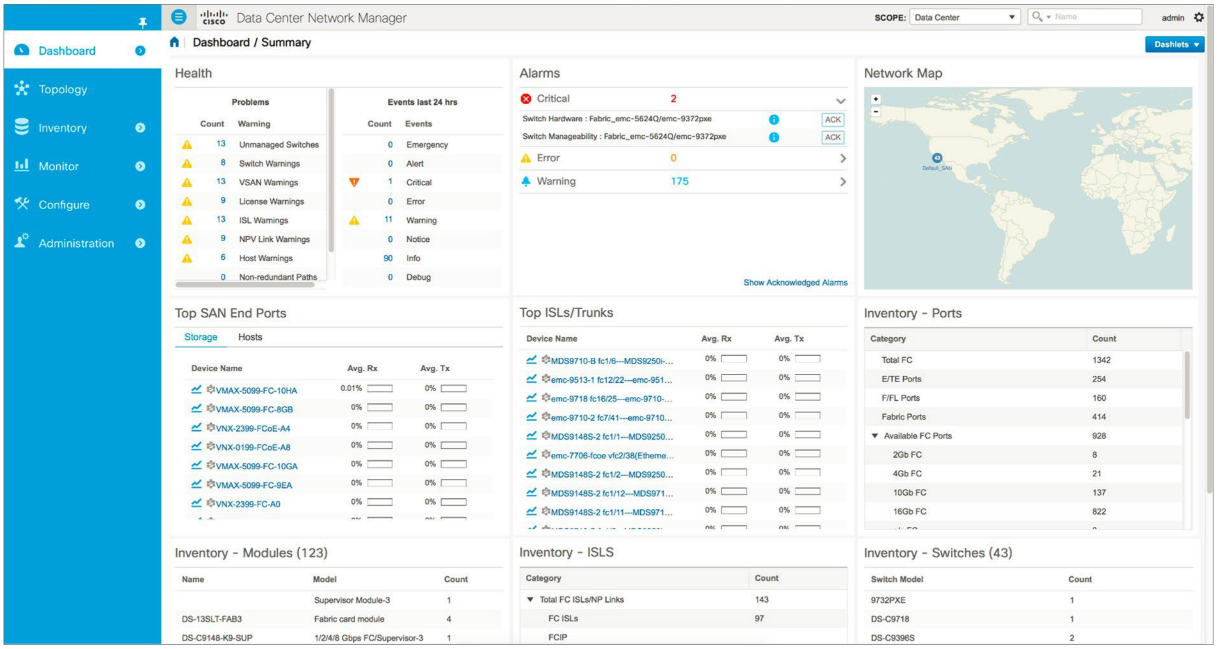
Cisco Smart Call Home builds on the Call Home feature, adding proactive diagnostics and fault management. In essence, Smart Call Home is a superset of the Call Home feature. The primary enhancement of Smart Call Home is the capability to use notification messages to automatically open technical-assistance tickets and resolve problems before they become critical. The Cisco Technical Assistance Center (TAC) maintains an inventory of events: faults and alerts are sent to the TAC, a case is created automatically, and users can access the Smart Call Home web portal. These capabilities help organizations quickly identify potential problems and resolve them quickly, while keeping an historical record for the networking devices. Preventive maintenance activities can be suggested, and availability improved as a result.

Management Tools

Human errors are reported to be the greatest cause of downtime in data centers. An easy-to-use GUI with embedded powerful monitoring tools can help improve uptime.

The MDS 9000 Family of storage networking devices is complemented by Cisco Prime™ Data Center Network Manager (DCNM), which provides a centralized and simple management tool for multiple fabrics. It builds on the inherent reliability of the MDS 9000 Family switches. It helps organizations quickly identify faults, run proactive diagnostics, automate configuration steps, and monitor performance from the source to the destination of data flows. Its ready-to-use, large-scale, and proven management capabilities provide visibility and control so that SLAs can be met. The DCNM dashboard, shown in Figure 17, is based on HTML5. It provides clear guidance, warning messages, and more across all phases of the storage network lifecycle. It enables administrators, for example, to simply drag and drop port world-wide names (PWWNs) for zoning instead of typing each one, avoiding the possibility of typing errors.

Figure 17. Cisco Prime DCNM 10 GUI Dashboard for SAN Administration



Health

	Problems	
	Count	Warning
⚠	13	Unmanaged Switches
⚠	8	Switch Warnings
⚠	13	VSAN Warnings
⚠	9	License Warnings
⚠	13	ISL Warnings
⚠	9	NPV Link Warnings
⚠	6	Host Warnings
	0	Non-redundant Paths

	Events last 24 hrs	
	Count	Events
	0	Emergency
	0	Alert
▼	1	Critical
	0	Error
⚠	11	Warning
	0	Notice
	90	Info
	0	Debug

Alarms

✖ Critical	2		
Switch Hardware : Fabric_emc-5624Q/emc-9372pxe		ⓘ	ACK
Switch Manageability : Fabric_emc-5624Q/emc-9372pxe		ⓘ	ACK
⚠ Error	0		›
Warning	175		›

Show Acknowledged Alarms

Network Map

Top SAN End Ports

Storage Hosts

Device Name	Avg. Rx	Avg. Tx
VMAX-5099-FC-10HA	0.01%	0%
VMAX-5099-FC-8GB	0%	0%
VNX-2399-FCoE-A4	0%	0%
VNX-0199-FCoE-A8	0%	0%
VMAX-5099-FC-10GA	0%	0%
VMAX-5099-FC-9EA	0%	0%
VNX-2399-FC-A0	0%	0%

Top ISLs/Trunks

Device Name	Avg. Rx	Avg. Tx
MDS9710-B fc1/6---MDS9250i-...	0%	0%
emc-9513-1 fc12/22---emc-951...	0%	0%
emc-9718 fc16/25---emc-9710-...	0%	0%
emc-9710-2 fc7/41---emc-9710...	0%	0%
MDS9148S-2 fc1/1---MDS9250...	0%	0%
emc-7706-fcoe vfc2/38(Etherne...	0%	0%
MDS9148S-2 fc1/2---MDS9250...	0%	0%
MDS9148S-2 fc1/12---MDS971...	0%	0%
MDS9148S-2 fc1/11---MDS971...	0%	0%

Inventory - Ports

Category	Count
Total FC	1342
E/TE Ports	254
F/FL Ports	160
Fabric Ports	414
▼ Available FC Ports	928
2Gb FC	8
4Gb FC	21
10Gb FC	137
16Gb FC	822

Inventory - Modules (123)

Name	Model	Count
	Supervisor Module-3	1
DS-13SLT-FAB3	Fabric card module	4
DS-C9148-K9-SUP	1/2/4/8 Gbps FC/Supervisor-3	1

Inventory - ISLS

Category	Count
▼ Total FC ISLs/NP Links	143
FC ISLs	97
FCIP	

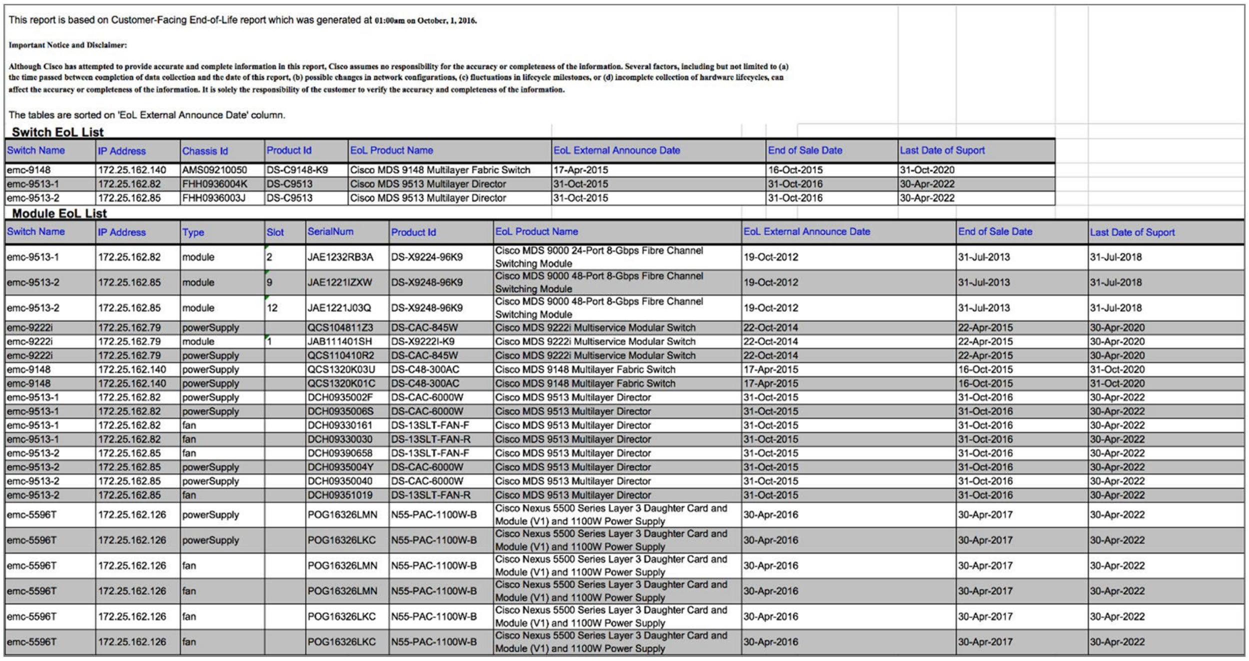
Inventory - Switches (43)

Switch Model	Count
9732PXE	1
DS-C9718	1
DS-C9396S	2

From DCNM, a southbound Storage Management Initiative Specification (SMI-S) interface can connect to storage arrays, and a northbound SMI-S interface allows integration with third-party management tools. OpenStack integration is achieved through a plug-in for the Cinder driver. Integration with external hypervisor solutions (such as VMware vCenter) is available through the web services API.

From the DCNM GUI, you can retrieve prebuilt and custom reports, which can help you ensure high availability for the storage fabric network. For example, you generate a report that shows where any MDS 9000 Family Fibre Channel switch in the fabric is within its lifecycle. One DCNM prebuilt report (Figure 18) shows both the end-of-life and the end-of-support dates for switches and modules. With this information, storage network administrators can quickly identify devices that are approaching the end of their useful life, plan for their refresh or replacement, and avoid running a switch for which vendor support can no longer be obtained.

Figure 18. Cisco Prime DCNM 10 Prebuilt Report for End-of-Life and End-of-Support for SAN Devices



This report is based on Customer-Facing End-of-Life report which was generated at **01:00am on October, 1, 2016.**

Important Notice and Disclaimer:

Although Cisco has attempted to provide accurate and complete information in this report, Cisco assumes no responsibility for the accuracy or completeness of the information. Several factors, including but not limited to (a) the time passed between completion of data collection and the date of this report, (b) possible changes in network configurations, (c) fluctuations in lifecycle milestones, or (d) incomplete collection of hardware lifecycles, can affect the accuracy or completeness of the information. It is solely the responsibility of the customer to verify the accuracy and completeness of the information.

The tables are sorted on 'EoL External Announce Date' column.

Switch EoL List

Switch Name	IP Address	Chassis Id	Product Id	EoL Product Name	EoL External Announce Date	End of Sale Date	Last Date of Suport
emc-9148	172.25.162.140	AMS09210050	DS-C9148-K9	Cisco MDS 9148 Multilayer Fabric Switch	17-Apr-2015	16-Oct-2015	31-Oct-2020
emc-9513-1	172.25.162.82	FHH0936004K	DS-C9513	Cisco MDS 9513 Multilayer Director	31-Oct-2015	31-Oct-2016	30-Apr-2022
emc-9513-2	172.25.162.85	FHH0936003J	DS-C9513	Cisco MDS 9513 Multilayer Director	31-Oct-2015	31-Oct-2016	30-Apr-2022

Module EoL List

Switch Name	IP Address	Type	Slot	SerialNum	Product Id	EoL Product Name	EoL External Announce Date	End of Sale Date	Last Date of Suport
emc-9513-1	172.25.162.82	module	2	JAE1232RB3A	DS-X9224-96K9	Cisco MDS 9000 24-Port 8-Gbps Fibre Channel Switching Module	19-Oct-2012	31-Jul-2013	31-Jul-2018
emc-9513-2	172.25.162.85	module	9	JAE1221IZXW	DS-X9248-96K9	Cisco MDS 9000 48-Port 8-Gbps Fibre Channel Switching Module	19-Oct-2012	31-Jul-2013	31-Jul-2018
emc-9513-2	172.25.162.85	module	12	JAE1221J03Q	DS-X9248-96K9	Cisco MDS 9000 48-Port 8-Gbps Fibre Channel Switching Module	19-Oct-2012	31-Jul-2013	31-Jul-2018
emc-9222i	172.25.162.79	powerSupply		QCS104811Z3	DS-CAC-845W	Cisco MDS 9222i Multiservice Modular Switch	22-Oct-2014	22-Apr-2015	30-Apr-2020
emc-9222i	172.25.162.79	module	1	JAB111401SH	DS-X9222I-K9	Cisco MDS 9222i Multiservice Modular Switch	22-Oct-2014	22-Apr-2015	30-Apr-2020
emc-9222i	172.25.162.79	powerSupply		QCS110410R2	DS-CAC-845W	Cisco MDS 9222i Multiservice Modular Switch	22-Oct-2014	22-Apr-2015	30-Apr-2020
emc-9148	172.25.162.140	powerSupply		QCS1320K03U	DS-C48-300AC	Cisco MDS 9148 Multilayer Fabric Switch	17-Apr-2015	16-Oct-2015	31-Oct-2020
emc-9148	172.25.162.140	powerSupply		QCS1320K01C	DS-C48-300AC	Cisco MDS 9148 Multilayer Fabric Switch	17-Apr-2015	16-Oct-2015	31-Oct-2020
emc-9513-1	172.25.162.82	powerSupply		DCH0935002F	DS-CAC-6000W	Cisco MDS 9513 Multilayer Director	31-Oct-2015	31-Oct-2016	30-Apr-2022
emc-9513-1	172.25.162.82	powerSupply		DCH0935006S	DS-CAC-6000W	Cisco MDS 9513 Multilayer Director	31-Oct-2015	31-Oct-2016	30-Apr-2022
emc-9513-1	172.25.162.82	fan		DCH09330161	DS-13SLT-FAN-F	Cisco MDS 9513 Multilayer Director	31-Oct-2015	31-Oct-2016	30-Apr-2022
emc-9513-1	172.25.162.82	fan		DCH09330030	DS-13SLT-FAN-R	Cisco MDS 9513 Multilayer Director	31-Oct-2015	31-Oct-2016	30-Apr-2022
emc-9513-2	172.25.162.85	fan		DCH09390658	DS-13SLT-FAN-F	Cisco MDS 9513 Multilayer Director	31-Oct-2015	31-Oct-2016	30-Apr-2022
emc-9513-2	172.25.162.85	powerSupply		DCH0935004Y	DS-CAC-6000W	Cisco MDS 9513 Multilayer Director	31-Oct-2015	31-Oct-2016	30-Apr-2022
emc-9513-2	172.25.162.85	powerSupply		DCH09350040	DS-CAC-6000W	Cisco MDS 9513 Multilayer Director	31-Oct-2015	31-Oct-2016	30-Apr-2022
emc-9513-2	172.25.162.85	fan		DCH09351019	DS-13SLT-FAN-R	Cisco MDS 9513 Multilayer Director	31-Oct-2015	31-Oct-2016	30-Apr-2022
emc-5596T	172.25.162.126	powerSupply		POG16326LMN	N55-PAC-1100W-B	Cisco Nexus 5500 Series Layer 3 Daughter Card and Module (V1) and 1100W Power Supply	30-Apr-2016	30-Apr-2017	30-Apr-2022
emc-5596T	172.25.162.126	powerSupply		POG16326LKC	N55-PAC-1100W-B	Cisco Nexus 5500 Series Layer 3 Daughter Card and Module (V1) and 1100W Power Supply	30-Apr-2016	30-Apr-2017	30-Apr-2022
emc-5596T	172.25.162.126	fan		POG16326LMN	N55-PAC-1100W-B	Cisco Nexus 5500 Series Layer 3 Daughter Card and Module (V1) and 1100W Power Supply	30-Apr-2016	30-Apr-2017	30-Apr-2022
emc-5596T	172.25.162.126	fan		POG16326LMN	N55-PAC-1100W-B	Cisco Nexus 5500 Series Layer 3 Daughter Card and Module (V1) and 1100W Power Supply	30-Apr-2016	30-Apr-2017	30-Apr-2022
emc-5596T	172.25.162.126	fan		POG16326LKC	N55-PAC-1100W-B	Cisco Nexus 5500 Series Layer 3 Daughter Card and Module (V1) and 1100W Power Supply	30-Apr-2016	30-Apr-2017	30-Apr-2022
emc-5596T	172.25.162.126	fan		POG16326LKC	N55-PAC-1100W-B	Cisco Nexus 5500 Series Layer 3 Daughter Card and Module (V1) and 1100W Power Supply	30-Apr-2016	30-Apr-2017	30-Apr-2022

Preconfigured Monitoring Templates

From the DCNM management system, interfaces can be monitored according to preconfigured templates reflecting multiple years of best practices in operating large Fibre Channel networks in an extremely reliable way. These templates can be pushed to an entire fabric with a single mouse click, and monitoring thresholds can be defined based on port type (host-facing port or ISL). Instead of having to implement monitoring on individual switches one at a time, administrators can apply to multiple switches and achieve fabricwide consistency. Monitoring templates allow administrators to quickly identify anomalies and respond so that proper operation is preserved on the SAN.

This same tool can also provide hints for capacity-planning purposes, showing which ports are fully utilized and where additional ports may be required to prevent future bottlenecks.

Configuration File Management with Backup and Restore Capabilities

DCNM also can be used as a centralized configuration file management solution, with backup and restore support both on individual switches and at the fabric level. Administrators can safely make changes to device configurations because the option to restore a well-known and previously backed-up file is always available. This capability can help reduce the impact of human errors on the overall availability of the SAN and significantly reduce human-generated downtime. Here are some examples the of uses of this capability:

- A zoning change causes the fabric to misbehave. Even though the NX-OS Software on the MDS 9000 Family switches provides several checks to help prevent inadvertent zoning changes and options to roll back configurations, mistake can still occur. The capability to revert to a working condition addresses this scenario.
- A configuration change on one device causes problems on another device in the fabric. For example, a change made on a director may exceed the scale limits on a fabric switch in the same fabric. Rolling back to a working configuration is faster than troubleshooting the network.
- A change in security policies (for example, role-based access control [RBAC] or authentication, authorization, and accounting [AAA] configuration) makes the fabric inaccessible to users entitled to use it. A quick fix is needed.

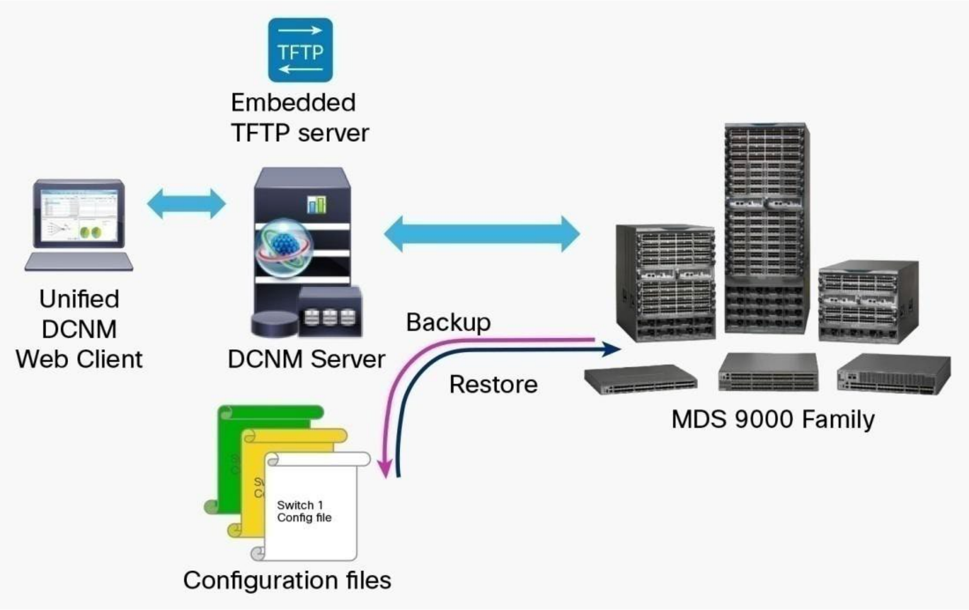
For IT operations staff, configuration file backup is a general best practice and is not specific to SAN devices. Scheduled regular backup operations protect against human errors as well as against hardware failures.

The traditional approach to configuring switch backups is to enable an external TFTP, SFTP, or FTP server; write a script to provide the backup commands on the switches; and then use a scheduling tool such as cron to run these commands. The MDS 9000 Family switches provide an internal scheduler to accomplish this task. The configuration file can then be restored on the switch following this simple procedure.

```
Myswitch# copy tftp://192.168.0.12/Myswitch.cfg bootflash:
  Trying to connect to tftp server......
  Connection to server Established. Copying Started.....
  TFTP get operation was successful
Myswitch# dir
  16153 Nov 05 14:31:27 2016 Myswitch.cfg
Myswitch#write erase
Myswitch#reload
Myswitch#copy bootflash://Myswitch.cfg running-config
Myswitch#copy run start
```

By using the DCNM GUI, the storage network administrator can obtain additional advantages. The administrator does not need to use an external TFTP, SFTP, or FTP server because the DCNM software contains an embedded TFTP server. The backup files will be stored in the DCNM database. The scheduling function is also part of DCNM. The GUI allows the administrator to easily configure regular backup operations for some or all switches in the fabric and restore files as needed (Figure 19). More typically, the running configuration file can be confronted and compared to the backup file for identification of deviations. Remediation can happen by making configuration changes to a single switch and have the changes propagate throughout the fabric via Cisco Fabric Services protocol. Alternatively DCNM allows for pushing CLI commands out to multiple switches at the same time.

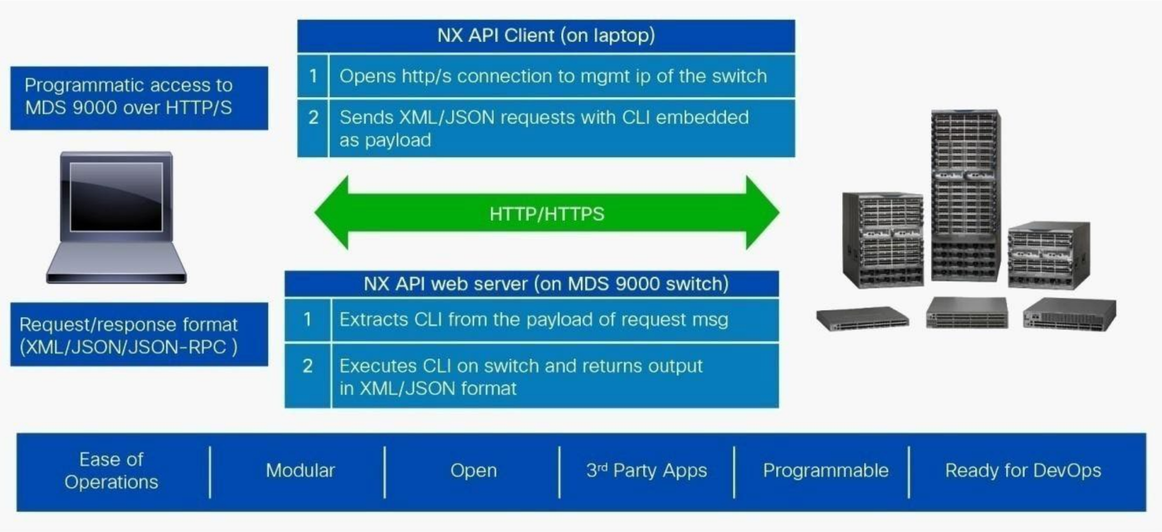
Figure 19. Cisco MDS 9000 Family Configuration Files Backup and Restore Solution



Programmability and Automation

MDS 9000 Family switches have long provided Simple Network Management Protocol Version 3 (SNMP v3) interfaces to connect to external management tools (including DCNM). More recently, an onboard Python interpreter was introduced. With increasing numbers of organizations pursuing digital transformation, automation has become a top priority for Cisco and its customers. Automation reduces the number of manual tasks that need to be performed, accelerates delivery of services, and helps prevent unplanned downtime. To promote automation, starting with NX-OS Release 7.3, every MDS 9000 Family device offers the Cisco NX-API, a representational state transfer (REST) API, which supports complete and direct programmability of individual networking elements without the limitations of intermediate software layers. The REST API framework enables direct programmatic access to the switches over HTTP/HTTPS. It provides a uniform programmatic interface between the client and the server that is stateless and cacheable. Resources are manipulated (passed, modified, and parsed) between the client and the server, as shown in Figure 20, typically using JSON or XML message formats, which are easy to parse and therefore to automate.

Figure 20. Cisco MDS 9700 Series Ease of Automation with REST API on Cisco NX-OS Software



Programmatic access to MDS 9000 over HTTP/S
Request/response format (XML/JSON/JSON-RPC)
NX API Client (on laptop)
1 Opens http/s connection to mgmt ip of the switch
2 Sends XML/JSON requests with CLI embedded as payload
HTTP/HTTPS
NX API web server (on MDS 9000 switch)
1 Extracts CLI from the payload of request msg
2 Executes CLI on switch and returns output in XML/JSON format
Ease of Operations
Modular
Open
3rd Party Apps
Programmable
Ready for DevOps

This approach allows application developers to remotely send Cisco CLI commands and receive responses. All Cisco CLI command types (**show** and **configure**) can be sent, and any REST-based tools can be used to interact with the devices (API Blueprint, Swagger, RAML, etc.). NX-API supports the most-touched SAN components: VSAN, zoning, device alias, etc. A crowd-sourced repository of NX-API-based utilities and scripts is provided for fabric-level configuration, zone provisioning and automation, health and performance monitoring, troubleshooting, and custom operations.

You can enable the NX-API feature on MDS 9000 Family switches with this command:

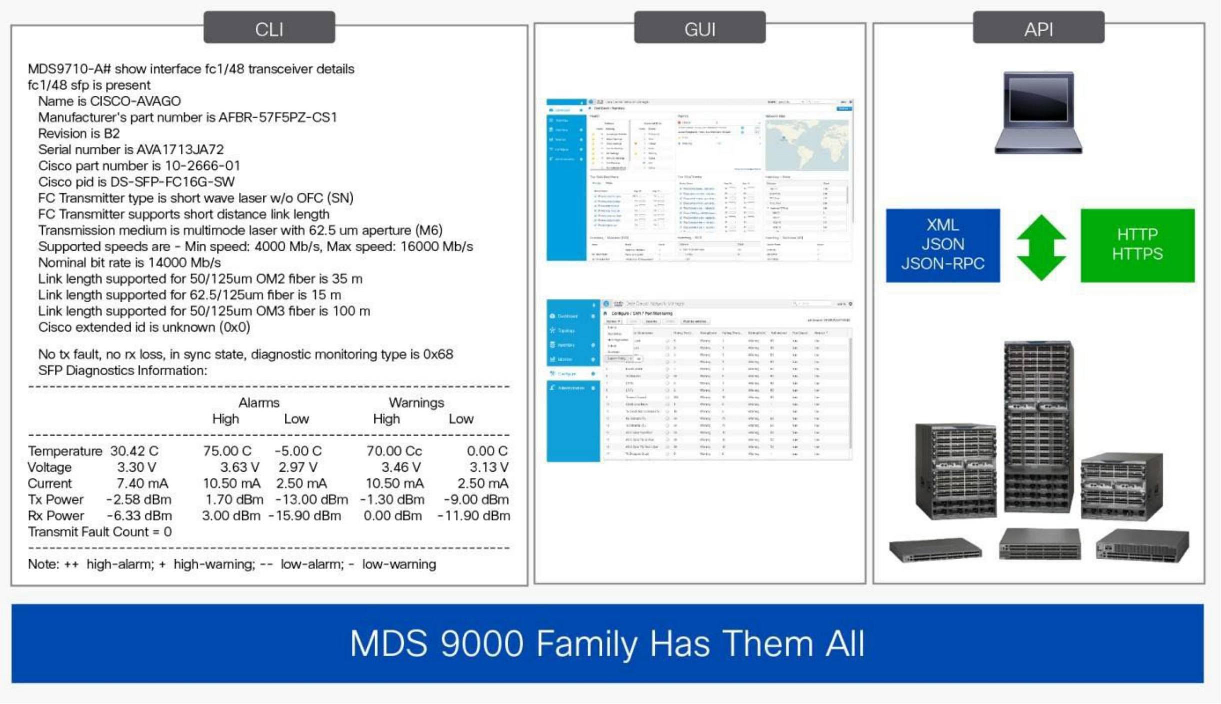
```
Switch(config)#feature nxapi
```

After the feature is enabled, you can control the switches using a web browser. On Cisco devices, CLI commands are run on the device itself. NX-API improves the accessibility of these Cisco CLI commands by making them available outside the switch by using HTTP/HTTPS. Because the application is not on the device, NX-API allows users to control multiple MDS 9000 Family devices through one single script. Using NX-API, a user can manage a Fibre Channel fabric built on MDS 9000 Family network elements using existing web-based technologies, without the need to learn new languages or technologies. Note, too, that NX-API also is available on all Cisco Nexus® networking products, allowing administrators to conveniently write code for the entire data center network.

Not all customers are moving with the same speed toward programmability and automation. The traditional Cisco CLI still provides a very convenient way to manage and operate MDS 9000 Family switches. The GUI, offered by DCNM, is quickly becoming an essential tool for many storage network administrators. It is intuitive, provides visibility between virtual machines and logical unit numbers (LUNs), and includes an embedded analytics engine for capacity planning and advanced troubleshooting. The REST NX-API provides excellent new capabilities for those willing to handle infrastructure as code.

The MDS 9000 Family of storage networking products thus supports all types of interfaces, as depicted in Figure 21, to help users wherever they are in their journey.

Figure 21. CLI, GUI, and API: Cisco MDS 9000 Family Has Them All



```
MDS9710-A# show interface fc1/48 transceiver details
fc1/48 sfp is present
   Name is CISCO-AVAGO
   Manufacturer's part number is AFBR-57F5PZ-CS1
   Revision is B2
   Serial number is AVA1713JA72
   Cisco part number is 10-2666-01
   Cisco pid is DS-SFP-FC16G-SW
   FC Transmitter type is short wave laser w/o OFC (SN)
   FC Transmitter supports short distance link length
   Transmission medium is multimode laser with 62.5 um aperture (M6)
   Supported speeds are - Min speed: 4000 Mb/s, Max speed: 16000 Mb/s
   Nominal bit rate is 14000 Mb/s
   Link length supported for 50/125um OM2 fiber is 35 m
   Link length supported for 62.5/125um fiber is 15 m
   Link length supported for 50/125um OM3 fiber is 100 m
   Cisco extended id is unknown (0x0)

   No tx fault, no rx loss, in sync state, diagnostic monitoring type is 0x68
   SFP Diagnostics Information:
----------------------------------------------------------------------------
                                  Alarms                  Warnings
                              High        Low          High        Low
----------------------------------------------------------------------------
Temperature  30.42 C      75.00 C     -5.00 C       70.00 Cc       0.00 C
Voltage       3.30 V       3.63 V      2.97 V        3.46 V        3.13 V
Current       7.40 mA     10.50 mA     2.50 mA      10.50 mA       2.50 mA
Tx Power     -2.58 dBm     1.70 dBm  -13.00 dBm    -1.30 dBm     -9.00 dBm
Rx Power     -6.33 dBm     3.00 dBm  -15.90 dBm     0.00 dBm    -11.90 dBm
Transmit Fault Count = 0
----------------------------------------------------------------------------
Note: ++  high-alarm; +  high-warning; --  low-alarm; -  low-warning
```

In-Service Software Upgrade and Downgrade

In addition to using redundant hardware components, the MDS 9000 Family of mission-critical directors is protected by an outstanding high-availability software framework. Features include process restartability and nondisruptive supervisor switchover. The switchover capability also helps ensure nondisruptive in-service software upgrade (ISSU) and in-service downgrade (ISSD).

Mission-critical directors in the MDS 9700 Series have two supervisor modules: in slots 9 and 10 (MDS 9718), slots 5 and 6 (MDS 9710), or slots 3 and 4 (MDS 9706). When the mission-critical director powers up and both supervisor modules are present, the supervisor module that comes up first enters the active mode, and the supervisor module that comes up second enters the standby mode. The standby supervisor module constantly monitors the active supervisor module. If the active supervisor module fails, the standby supervisor module takes over without any impact on user traffic. A switchover can also be initiated manually when a stable standby supervisor module is available (high-availability standby state). The standby supervisor module automatically synchronizes its image with the running image on the active supervisor module.

The ISSU and ISSD capabilities are extremely important for system availability. Any software problems can be corrected without affecting traffic. New features can be made available for use without any disruption. ISSU is as important as hardware reliability in improving overall system uptime.

Process Restartability

Process restartability contributes to the high availability of MDS 9000 Family switches. It helps ensure that process-level failures do not cause system-level failures. It also restarts failed processes automatically, making the MDS 9700 Series exceptionally robust. The supervisor module continuously 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 increases reliability because supervisor failover 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. Even in the rare event that a supervisor module is reset, complete synchronization between the active and standby supervisor modules helps ensure stateful failover with no disruption of traffic. This vital capability functions on infrastructure that is internal to the switch.

The individual processes running on a switch can be displayed by the following command:

```
switch# show process
PID State PC Start_cnt TTY Process
-------- -------- ------------ ---------------------
868 S 2ae4f33e 1 - snmpd
869 S 2acee33e 1 - rscn
870 S 2ac36c24 1 - qos
871 S 2ac44c24 1 - port-channel
872 S 2ac7a33e 1 – ntp
- ER - 1 – mdog
- NR - 0 – vbuilder
```

Per-VSAN Fabric Services

To help limit fault domains, ease management tasks, and reduce costs, Cisco introduced virtual SAN, or VSAN, technology for the MDS 9000 Family of switches. VSANs help data centers achieve multiple isolated environments without the expense of having to build physically separate fabrics. Each separate virtual fabric runs on top of the same physical infrastructure and 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 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 (VSAN trunking ports, however, can transport multiple VSANs). Therefore, whatever end node is connected to a physical port is a member of that port's VSAN.

VSAN technology is fully supported on various Cisco data center products, including Cisco Nexus switches and the Cisco UCS platform. VSANs offer a great deal of flexibility to the user. In theory, up to 4096 VSANs can be configured per physical infrastructure, though a more practical number is 2 to 16 VSANs, with storage vendor certification achieved for up to 32 VSANs for switches and up to 80 VSANs for mission-critical directors.

Each VSAN can be selectively added to or deleted from an ISL trunk to control the VSAN's reach. In addition, special traffic counters are provided to track statistics per VSAN.

In addition to providing strict isolation with hardware enforcement, 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 log-in server, is created and enabled across those switches that are configured to carry the new VSAN. This replication of services enables organizations 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 in VSAN 10 does not affect the fabric in any way in VSAN 20.

VSANs also provide a way to interconnect isolated fabrics in remote data centers over a common long-haul infrastructure. Because frame tagging is performed in hardware and is included in every frame traversing the EISL, traffic can be transported using dense wavelength-division multiplexing (DWDM), coarse wavelength-division multiplexing (CWDM), or FCIP. 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.

A VSAN represents a logical fault domain, in contrast to zones, which provide only a security perimeter for node communication.

Role-Based Access Control

Security is not a consideration normally associated with high availability. However, as previously mentioned, one of the leading causes of network downtime is human error. For example, a user may mistakenly run a command without fully realizing the implications of that command.

The MDS 9000 Family supports a role-based security methodology to help ensure that only authorized individuals have access to critical functions in the fabric. With role-based access control (RBAC), each user is assigned to a role that has a specific privilege level in the fabric. This privilege level dictates the commands 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 specificity of this permission system can be very precise: two levels deep within the command hierarchy.

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 Fabric Shortest Path First (FSPF) debug commands.

Roles can be defined and assigned locally within a switch by using CLI commands, or centralized from a RADIUS, TACACS, or Lightweight Directory Access Protocol (LDAP) server for easier management. Default roles include network administrator (full access) and network operator (read-only access), and custom roles can be provisioned.

RBAC can be applied on a per-VSAN basis. In this way, some users can have access to a specific VSAN, while others do not. By using this specific RBAC capability and VSANs, some organizations share the same physical SAN for Fibre Channel and IBM Fibre Connection (FICON) traffic.

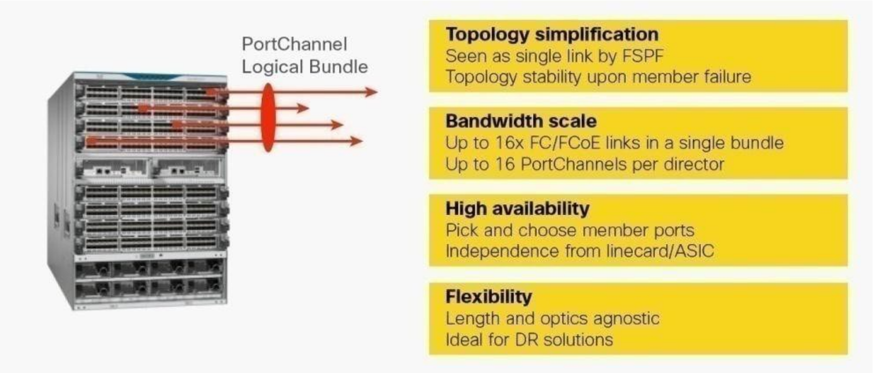
Port Channels, Virtual Router Redundancy Protocol, FSPF, and Port Tracking

The high-availability software framework also supports port channels, to protect ISLs, and Virtual Router Redundancy Protocol (VRRP), to protect management ports.

High availability is implemented at the fabric level using robust and high-performance ISLs. The port channel capability allows you aggregate up to 16 physical links into one logical bundle. The bundle can consist of any speed-matched ports in the chassis, even on different line cards and different ASICs, helping ensure that the bundle can remain active in the event of a port, ASIC, or module failure.

ISLs in a port channel can have significantly different lengths and resulting latencies. This capability is valuable in campus and metropolitan area network (MAN) environments, because logical links can be spread over multiple physical paths, helping ensure uninterrupted connectivity even if one of the physical paths is disrupted. Figure 22 shows the benefits of port channels.

Figure 22. Benefits of Cisco MDS 9000 Family Port Channel Technology: Simplicity, Performance, and High Availability



Topology simplification
Seen as single link by FSPF
Topology stability upon member failure

Bandwidth scale
Up to 16x FC/FCoE links in a single bundle
Up to 16 PortChannels per director

High availability
Pick and choose member ports
Independence from linecard/ASIC

Flexibility
Length and optics agnostic
Ideal for DR solutions

As required by standards, the FSPF protocol comes to rescue whenever frames need to be rerouted around a failed ISL or port channel. FSPF is a link-state routing protocol that directs traffic along the shortest path between the source and destination according to the link cost. For multihop networks, the protocol analyzes all paths from a switch to all the other switches in the fabric, calculates the path cost by adding the cost of all links traversed along that path, and then chooses the path with the lowest cost. If a failure is detected by FSPF, a new shortest route for traffic is calculated and the routing table is updated and shared among all switches in the fabric. The collection of link states and their cost for all the switches in a fabric constitutes the topology database. The FSPF protocol runs natively in NX-OS on MDS 9000 Family devices and allows SANs to be self-healing.

The port tracking feature enhances SAN extension resiliency. If an MDS 9000 Family switch detects a WAN or MAN link failure, it takes down the associated disk-array link if port tracking is configured. Then the array can redirect the failed I/O operation to another link without waiting for an I/O timeout. Otherwise, disk arrays must wait seconds for an I/O timeout to recover from a network link failure. This feature contributes to network availability by reducing amount of time needed to detect link failures and restore operations for optimal service.

Summary of Capabilities

The MDS 9700 Series of mission-critical directors has set the new benchmark for enterprise-class reliability for storage networking. It was designed from the beginning to deliver reliability and high availability at scale through nondisruptive software upgrades, stateful process restart and failover, redundancy of all infrastructural components, management and diagnostics tools, automation and programmability, smart call home functions, fault-domain containment, security postures, and a focus on all details including serviceability. Table 2 summarizes the main capabilities. The list is not exhaustive.

Table 2. High-Availability Capabilities in the Cisco MDS 9700 Series

Capability	Description
Reliability and availability	• Hot-swappable 1+1 redundant supervisor modules • Hot-swappable N+1 redundant crossbar fabric modules • Hot-swappable N+N redundant power supplies • Hot-swappable fan trays and redundant fans for each fan tray, with integrated temperature and power management • Hot-swappable SFP+ optics (2/4/8/10/16/32-Gbps Fibre Channel and 10 Gigabit Ethernet) and QSFP+ optics (40-Gbps FCoE) • Hot-swappable switching modules (line cards) • Passive backplane, redundant IDPROM units • Online, nondisruptive software upgrades (ISSU) and downgrades (ISSD) • Stateful nondisruptive supervisor module failover with watchdog daemons • Stateful process restart • Any module, any port configuration for port channels • FSPF protocol • Fabric-based multipathing • Per-VSAN fabric services • RBAC • POST • Online end-to-end diagnostics, health check, and troubleshooting suite • Port tracking • VRRP for management ports • DCNM GUI, monitoring templates, centralized backup of configuration files
Serviceability	• Modular design with hot-swappable components • Flash memory on supervisors to store at least two software images • USB port on supervisor modules for speedy file management, including configuration and firmware upgrades • Configuration file management • Nondisruptive software upgrades for Fibre Channel interfaces • Cisco Call Home and Smart Call Home • Logging system • Power-management LEDs • Port beaconing • System LEDs • Locator ID LEDs on all system modules, including power supplies and fans • Module ejectors and handles for ease of replacement • Vertical cable management brackets • Port-side to rear-side airflow direction • Internal temperature and power sensors everywhere • SNMP v3 support and SNMP traps for alerts • Network boot (power-on auto-provisioning [POAP]) and USB-key boot • Deep programmability with REST API on networking devices • SNMP, SMI-S, Python, and Cinder support for integration with management tools • Online end to end diagnostics, health check and troubleshooting suite • Onboard failure logging (OBFL) for persistent NVRAM date/time based diagnostic information

All these capabilities are built in to the MDS 9700 Series to help administrators keep the storage network healthy, stable, and consistent. This achievement reflects the Cisco development team’s ―right the first time‖ approach, rather than the more common ―break and then fix‖ approach.

Fabric-Level Reliability

When port requirements exceed the capability of a single switching device, or when server and disk components are spread across multiple racks and data center rooms, multiple Fibre Channel switches need to be connected to form a fabric. In this case, the reliability of individual switches will affect the availability of the entire fabric. Also, appropriate fabric topologies will be more resilient and stable than those that do not follow best practices. In this scenario, you must consider fabric-level reliability.

One advantage of mission-critical directors is their modularity. They allow organizations to start with a relatively low port count and then scale up as needed with a convenient pay-as-you-grow model. By choosing the appropriate chassis size, enterprises with requirements ranging from 192 to 768 ports per fabric have simplified their SANs to a single network element.

Mission-critical directors also provide greater flexibility than fixed switches, because their chassis can house different blade types. These can provide FCoE connectivity to extend storage access to disk arrays natively equipped with FCoE ports or to establish high-speed ISLs over a reduced number of links. In this scenario, the 40-Gbps FCoE option is slowly replacing the previous generation of 10-Gbps FCoE line cards.

Another benefit of mission-critical directors is their support for SAN extension through line cards with a combination of Fibre Channel and FCIP ports, with their encryption and compression features. FCIP is the technology of choice for connecting two mission-critical directors over a WAN for long-distance data replication. Fibre Channel over wavelength-division multiplexing (WDM) may be preferred when distances are shorter, in the metropolitan-area range.

Although organizations can build relatively large fabrics using only fixed-configuration switches, the need for ISLs to connect and build fabrics leads to additional complexity when the number of ports exceeds a few hundred. Cabling becomes a concern, and management tasks require more advanced skills and appropriate tools. Also, switches are equipped with a single controller processor, which cannot govern the login and logout of too many end nodes in the fabric at one time and cannot scale beyond a certain number of zones, fabric logins (FLOGIs), Fibre Channel network switch (FCNS) entries, and other parameters. As a general rule, customers building a SAN fabric that will scale to more than seven switches should consider using mission-critical directors instead, eventually in a core-edge topology. Some customers will want to consider modular platforms as soon as they need more than 96 ports per fabric.

When helping ensure fabric-level reliability, you also need to consider the cable plant that transports the data. A poor implementation can result in system outages, unpredictable application performance, and higher operating expenses. First, you need to comply with the maximum supported distances and loss budgets for the specific transceivers in use. In general, the higher the bit rate, the lower the maximum distance on a given fiber type. Second, old fiber types (in particular, OM1 and OM2) may be suboptimal for newer bit rates, and you should consider replacing them with more recent types to help ensure that the cable plant is appropriate for present and future needs. Occasionally, the adoption of 40-Gbps FCoE connectivity with its innovative BiDi QSFP+ transceivers may allow you to preserve investments in existing multimode cable plants (in particular, OM4 cable types) by covering distances of up to 150 meters. This distance is not possible with standard 32-Gbps and even 16-Gbps Fibre Channel transceivers (unless you use the newest violet-colored OM4+ fiber type).

A structured cabling approach is normally recommended and is well described in the TIA-942 data center standard. Structured connectivity topologies with low-loss, small-size connectors (LC type) help ensure compliance with industry standards while protecting investments in infrastructure by supporting new technologies.

Table 3 provides additional information about Fibre Channel and FCoE speeds, fiber types, and maximum distances.

Table 3. Fibre Channel Optics, Speed, Distance, and Media Type

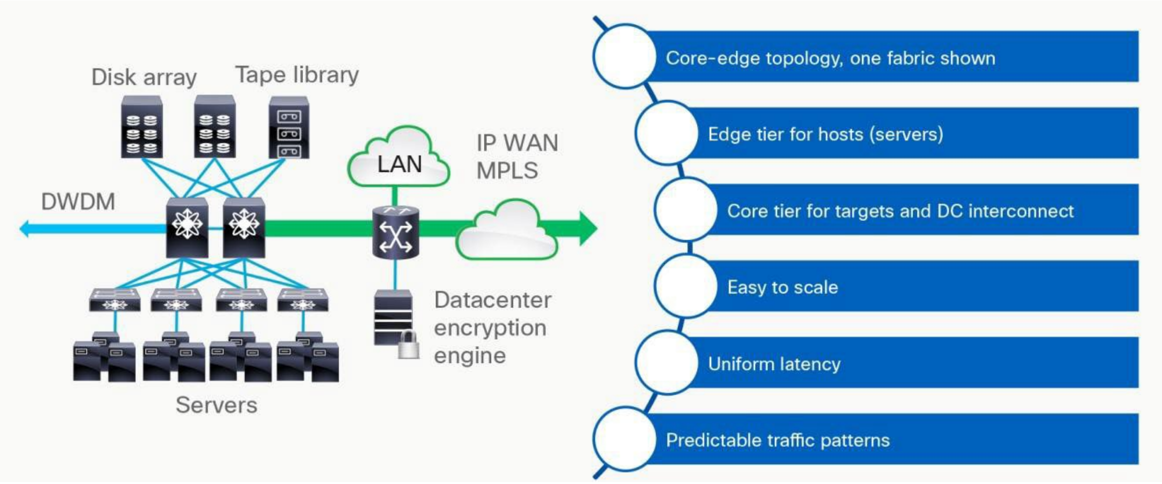
Description	Part Numbers	Distance and Media Type		
8G FC SW SFP+ Module	DS-SFP-FC8G-SW	150M OM3	190M OM4	~225M OM4+
16G FC SW SFP+ Module	DS-SFP-FC16G-SW	100M OM3	125M OM4	~150M OM4+
32G FC SW SFP+ Module	DS-SFP-FC32G-SW	70M OM3	100M OM4	~125M OM4+
10GBASE-SR SFP+ Module	SFP-10G-SR	300M OM3	400M OM4	~500M OM4+
40GBASE-SR QSFP+ Module	QSFP-40G-SR4	100M OM3	150M OM4	~200M OM4+
40GBASE-CSR QSFP+ Module	QSFP-40G-CSR4	300M OM3	400M OM4	~450M OM4+
40GBASE-QSFP+ BiDi Module	QSFP-40G-SR-BD	100M OM3	150M OM4	~200M OM4+
Active optical cable assembly	QSFP-H40G-AOCxM (x=1, 2, 3, 5, 7, or 10)	1, 2, 3, 5, 7, or 10M		

Although Fibre Channel networking devices are essentially plug-and-play, and the protocol itself can function properly in a variety of topologies, you always should carefully evaluate the design and deployment strategy. To provide reliable and efficient delivery of data, your SAN topology should follow best practices based on SAN industry standards and considerations specific to your devices and their vendors. You also need to consider physical-environment factors such as power, cooling, and rack layout, as well as network connectivity and software configuration. Focusing on these preliminary planning steps will help ensure that your SAN, when it is deployed, meets your current and future business objectives, including your requirements for availability, deployment simplicity, performance, future business growth, and cost.

SAN topology is described according to the way that the switches are interconnected. Some examples are port-expansion, ring, core-edge, edge-core-edge, and fully meshed topologies. Taking into account the most recent 16- and 32-Gbps Fibre Channel switching products on the market and Cisco’s extensive experience over many years in this business, Cisco recommends a core-edge topology to optimize performance, management, and future scalability. This deployment model provides the best choice over alternatives when costs and troubleshooting features are considered. At a high level, the core-edge topology has a variable number of edge switches for end-node connectivity, and a smaller number of core switches (typically mission-critical directors) for collecting traffic from edge switches and sending it to disk arrays or remote data centers, as shown in Figure 23.

In other words, this is a two-tier design in which initiators are connected on the edge tier and targets on the core tier. The core tier is also used to connect devices with data center–wide scope, such as DWDM systems, encryption engines, storage virtualization engines, and tape libraries.

Figure 23. Fibre Channel Core-Edge Topology


Disk array
Tape library
LAN
IP WAN
MPLS
DWDM
Datacenter
encryption
engine
Servers
Core-edge topology, one fabric shown
Edge tier for hosts (servers)
Core tier for targets and DC interconnect
Easy to scale
Uniform latency
Predictable traffic patterns

This design creates a scenario in which every server is one hop away from the disk array, keeping latency uniform. It also provides easy scalability by adding more edge switches connected to free ports on core devices. Scaling was not as easy a few years ago, when switches were limited to 48 ports and directors to 384 ports. Today, with 96-port switches readily available and mission-critical directors able to accommodate up to 768 ports, the scale limitations of core-edge topologies are much more relaxed, particularly when only directors are used. Remote data centers can be reached through dedicated ports on core devices, through either native Fibre Channel ports or FCIP links.

Management and operational capabilities are also optimized, because the core is not exposed to frequent configuration changes, and actions at the edge have a reduced fault domain. In some circumstances, edge switches can be configured in N-port virtualization (NPV) mode to reduce domain ID sprawl and reduce management points. This approach is common when blade servers are used. The traffic pattern is clear and deterministic—from edge to core—simplifying ISL oversubscription calculation and troubleshooting. With this topology, an exception to the one-hop-away rule may occur if two storage arrays are connected to different core switches. In this case, the storage-to-storage local replication traffic will traverse two hops. This situation can be remediated by connecting the two core devices within a single fabric through an ISL.

The core-edge topology is not the only possible deployment model, but it is the first you should consider when a single device cannot meet your port-count requirements and you need a fabric of multiple switches. By avoiding local switching of traffic, this approach makes scalability and troubleshooting easier. The separation of attachment points for storage arrays and servers helps reduce the risk of incidental cross-propagation of human mistakes. By using large modular directors in the core, this approach allows a high level of scalability without the need to deploy multiple fabrics or an edge-core-edge topology, inevitably more complicated to manage because of the quantity of ISLs in place. With core-edge topologies, you can build fabrics with 2500 ports or more with few SAN design considerations required.

One important aspect of any SAN topology is its resiliency and redundancy. The main objective for storage architects is to remove any single point of failure. Resiliency refers to the self-healing capability of the network: its ability to automatically recover from a failure and still continue to function. Redundancy refers to the duplication of components, network elements, and even an entire fabric to eliminate any single point of failure in the network. Redundancy is the key to high availability and enterprise-class installations. For organizations wanting to achieve business continuance under both foreseeable and unforeseeable circumstances, the elimination of any single point of failure should be a top priority. This is why a share-nothing approach should be used at the highest level of fabric design: the complete network should be redundant, with two completely separate fabrics and no network equipment in common. The use of logical segmentation on top of a single physical fabric can protect from human errors and other software-related issues, but it cannot provide the same degree of availability as two physically separated infrastructures.

Servers and storage devices should be connected to both physical fabrics. Data traffic should flow across both networks transparently in either active-active or active-passive mode, depending on the settings applied to the multipath I/O (MPIO) solution. MPIO is responsible for helping ensure that if one path on a host fails, an alternative path is readily available. Ideally, the two networks should be identical, but during migrations, differences in the way the networks are designed and in the products used to build them are common. Generally these two networks, identified as SAN A and SAN B, are in the same location. However, to provide greater robustness at the facility level, they are sometime kept in separate data center rooms. Enterprises may also rely on a secondary data centers to achieve business continuance or disaster recovery, depending on the distance between data centers and the recovery point objective (RPO). The use of two fabrics locally and two separate locations within the territory provides an excellent approach for achieving complete redundancy.

In addition to redundant fabrics operating in parallel, high availability within each individual fabric is required as a best practice to use redundant links between switches. A minimum of two ports, on two different line cards, should be used for ISLs, and they should be part of a logical bundle in a Fibre Channel port channel. With this setup, the storage network administrator can take down one of the member links for diagnostic purposes without disrupting the traffic on the remaining Fibre Channel port channel members.

This level of redundancy on ISLs would also prevent fabric segmentation even if a link shuts down under a failure condition. Active mode should be preferred as a setting for port channels so that recovery occurs automatically without the need to explicitly enable and disable the port-channel member ports at either end of the link.

SAN extension line cards should be redundant within a single mission-critical director, and traffic should be load-shared among them. MDS 9700 Series mission-critical directors can be filled up with SAN extension line cards with no limitation on the number that can be accommodated. The MDS 9718 chassis can deliver up to 1.28 Tbps of FCIP traffic. Members of the same Fibre Channel port channel should be placed on different line cards, on different ASICs, or in different port groups whenever possible. Creating a logical bundle across ports served by the same ASIC will not have any positive effect on network availability.

To improve operational ease and achieve greater availability, configuration and cabling should be consistent across the fabric. For example, do not configure ISLs on the top-left ports in one chassis and on the bottom-right ports in another chassis: mirrored configurations are recommended.

Here is a list of best practices recommended for proper and reliable SAN design to help ensure application availability:

- Avoid a single point of failure by using share-nothing redundant fabrics.
- Use MPIO-based failover for server-to-storage connectivity using redundant fabrics.
- Use redundancy features built into individual fabrics.
- Use mirrored cabling and configurations for ease of operation and troubleshooting.
- Use a core-edge topology with separate storage and server tiers for independent expansion.
- Core switches should be mission-critical directors whenever economically viable.
- Spread port channel members across line cards and use the active-mode setting.
- Use the highest performing and most resilient switch in the fabric (typically a core mission-critical director) as the principal switch. Point to it for zoning changes and use it as the seed for management tools.
- Always use static domain IDs.
- Enable and configure optional features that help prevent user misconfiguration using checking, alerting, and monitoring functions.

Cisco fabrics have resiliency features built in to the NX-OS operating system, the software that runs on all MDS 9000 Family switches, whose self-healing capabilities can quickly overcome most failures and repair the network. For example, when a link between switches fails, the Fibre Channel port channel technology will immediately move all traffic flowing through that port channel to the surviving member links. If an entire Fibre Channel port channel fails (very unlikely), the FSPF process will immediately recalculate the distribution of all traffic flows. All these functions require a second route to be available, making use of fabric redundancy.

NX-OS also includes other capabilities that help make networks built using MDS 9000 Family devices extremely resilient and highly available. For example, processes can be gracefully shut down and restarted. VSANs isolate traffic flows at the hardware level, to the point that a misconfiguration in the zoning database in a VSAN will not affect the others. When an individual port is administratively shut down, the process occurs gracefully, with buffers cleared and no packets lost.

A highly available storage network needs to be paired with a highly available storage infrastructure. Data must be available when it is accessed. Several approaches can be used for this purpose, including Redundant Array of Independent Disks (RAID) implementations, multiple copies of data spread across a clustered system, data replication over distance, and tape backup.

Conclusion

The design of highly available Fibre Channel and FCoE networks is not an easy task. The challenges of providing IT services 24 hours a day, 7 days a week, with no downtime or maintenance windows is causing IT managers to turn to mission-critical directors as the preferred hardware. When coupled with a solid software implementation, careful selection of fabric topology, and adherence to best practices and recommendations, mission-critical directors offer the highest level of availability. MDS 9700 Series 32-Gbps Fibre Channel–capable storage networking products offer an intrinsically superior design. Their arbitrated crossbar multifabric architecture with true N+1 redundancy represents an industry first in the Fibre Channel business and an unsurpassed benchmark for predictable performance and high availability.

In addition to hardware redundancy, the MDS 9700 Series provides highly resilient software with an innovative high-availability feature set designed to eliminate downtime in the storage network. Serviceability aspects have also been carefully addressed to reduce the time needed to restore operation after a system failure. The era of six-nines availability is now here. Welcome aboard.

For More Information

For additional information, see:

- http://blogs.gartner.com/andrew-lerner/2014/07/16/the-cost-of-downtime/
- http://www.snia.org/sites/default/orig/DSI2014/presentations/StorPlumb/CraigCarlson_Gen_6_Fibre_Channel%20_v02.pdf
- http://fibrechannel.org/wp-content/uploads/2015/11/FCIA_Roadmap_092215.pdf
- http://www.availabilitydigest.com/private/0206/benchmarking.pdf
- http://www.cisco.com/c/dam/en/us/products/collateral/storage-networking/mds-9700-series-multilayer-directors/esg_wp_cisco_next_gen_mds_apr_2013.pdf
- http://www.cisco.com/c/en/us/products/collateral/interfaces-modules/storage-networking-modules/prod_white_paper0900aecd8044c7e3.html
- http://www.cisco.com/c/en/us/products/storage-networking/mds-9700-series-multilayer-directors/datasheet-listing.html
- http://www.cisco.com/c/en/us/products/collateral/storage-networking/mds-9396s-16g-multilayer-fabric-switch/datasheet-c78-734525.html
- http://www.cisco.com/c/en/us/products/collateral/storage-networking/mds-9148s-16g-multilayer-fabric-switch/datasheet-c78-731523.html
- http://www.cisco.com/c/en/us/products/collateral/storage-networking/mds-9250i-multiservice-fabric-switch/data_sheet_c78-727493.html
- http://www.emersonnetworkpower.com/en-US/Resources/Market/Data-Center/Latest-Thinking/Ponemon/Pages/default.aspx
- http://www.cisco.com/c/en/us/support/docs/storage-networking/mds-9000-nx-os-san-os-software/118952-technote-mds9k-00.html
- http://www.cisco.com/c/en/us/support/docs/storage-networking/mds-9500-series-multilayer-directors/117621-configure-MDS-00.html
- http://www.cisco.com/c/dam/en/us/products/collateral/storage-networking/mds-9700-series-multilayer-directors/white-paper-c11-736963.pdf
- http://www.cisco.com/c/en/us/products/collateral/storage-networking/mds-9700-series-multilayer-directors/white-paper-c11-729697.html
- http://www.cisco.com/c/en/us/products/storage-networking/mds-9700-series-multilayer-directors/white-paper-listing.html
- http://www.cisco.com/c/en/us/products/collateral/storage-networking/mds-9700-series-multilayer-directors/white-paper-c11-734381.pdf
- http://www.cisco.com/c/en/us/products/collateral/interfaces-modules/transceiver-modules/data_sheet_c78-660083.html
- https://www.seagate.com/files/docs/pdf/datasheet/disc/cheetah-15k.7-ds1677.3-1007us.pdf

Appendix: Reliability and Availability

Concepts: Reliability Versus Availability

Often the words ―availability‖ and ―reliability‖ are incorrectly used interchangeably. In reality, they indicate two very different concepts. They can be applied to individual components or parts or an entire system.

System availability is derived from the availability of the system's parts and components using specific standards to calculate this metric. System availability can be defined as the probability that a system will be operating properly when it is requested for use. It can also be defined as the probability that the system will not be in a failure state or under unscheduled maintenance when its use is requested. For example, a car is available when you turn the key and can start driving. The intervals when a tire is flat and when the battery is discharged contribute to its unavailability period.

Now consider the concept of reliability. At first glance, you might think that if a system has high availability, then it also has high reliability. However, this is not necessarily the case, even though there is a strong relationship between the two characteristics. Reliability represents the probability that a system will perform its mission under well specified conditions and for a desired period of time. The higher the probability, the better. In other words, reliability indicates the chances that a system will be operational at a specific time.

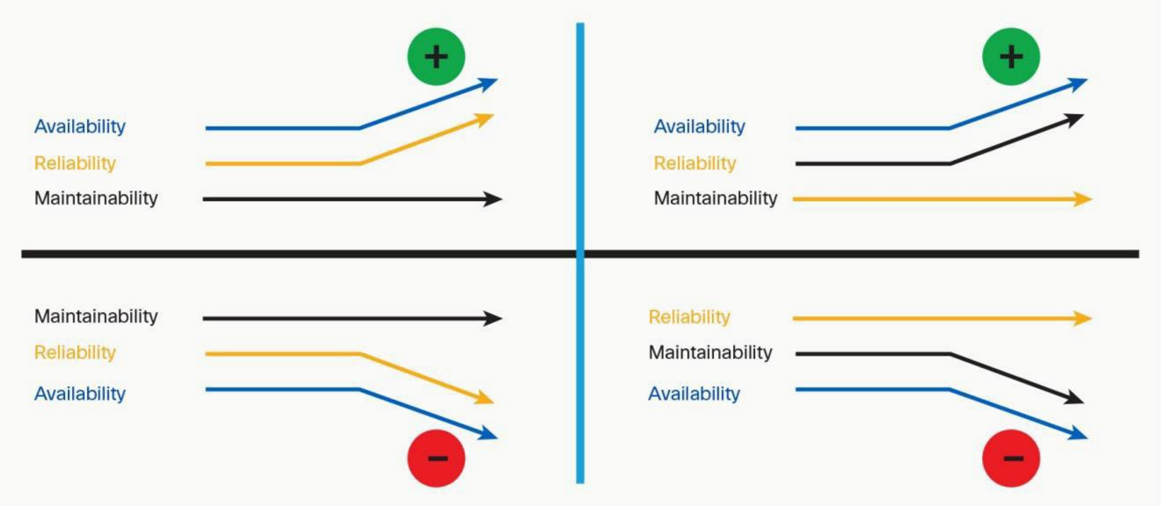
Note that reliability does not reflect the amount of time that will be needed to get a unit under repair back in a working condition. In the car example, for instance, you could have an extremely reliable car that almost never fails. However, if the amount of time needed to repair it when it does fail is very long, its availability will be negatively affected.

Reliability is essentially an intrinsic property of the system in question, reflecting the quality of its components and the quality of the overall architectural design, whereas availability also takes into account external factors such as logistics (are spare parts readily available?) and administration skills (did someone made improper use of it?). Most IT systems fail only occasionally, so it is important to think of reliability in statistical terms rather than look at an individual system and extrapolate from that.

Some people like to think of reliability as the number of times the system fails, whereas availability is inversely dependent on the amount of unscheduled downtime that results from the failures. A system is highly reliable if the number of failures is low. A system is highly available if the amount of unscheduled downtime is low.

You can see now that availability is a function of reliability, but it also depends on maintainability (often treated as synonymous with serviceability). Maintainability refers to the amount of time needed to repair the system. A system with high maintainability requires only a short time to perform maintenance. Figure 24 shows the relationship of reliability, maintainability, and availability.

Figure 24. Relationship of Reliability, Maintainability, and Availability



As can be seen in Figure 24, holding reliability constant, even at a high value, does not necessarily imply high availability. As the time to repair increases, maintainability goes down and availability decreases. Even a system with low reliability could have high availability if the time to repair is very short. Consider two extreme examples. In one case, a system runs perfectly well for almost 365 days a year and then suddenly fails. Because it needs a custom-made component, it takes one year to repair. This system has extremely high reliability but very poor availability over a two-year time frame. In the second case, a system breaks every day, but it takes only one second to repair. As a result, reliability is poor, but availability is reasonably high. In fact, over the same period of two years, this second system would be available all the time except for the 730 seconds when it is under maintenance. In the real world, when high availability for a system is desired, you should select a highly reliable system as the foundation and make sure that maintenance cycles are as short as possible.

Availability can be defined in different ways according to the types of downtimes considered in the analysis. In general, IT managers are interested mainly in operational system availability. Operational system availability includes all sources of experienced downtime: administrative downtime, logistics downtime, etc. Operational availability is affected by situations that are outside the system itself, such as logistics downtime from a lack of spare parts and downtime for preventive maintenance.

Although any downtime is unwelcome, it is unscheduled downtime that is more problematic for customers. They can plan for maintenance downtime as necessary, and this downtime is accepted as a condition for a good system. And for electronic devices that support hot-pluggable modules and in-service software upgrades, scheduled downtime is often absent. However, unscheduled downtime resulting from problems and poor reliability can infuriate customers. There is a major difference in customer perception of scheduled and unscheduled downtime, and a major difference in customer satisfaction.

Consider the following scenario. If the single power supply for your home router fails, to remediate the problem you would remove it and replace it with a new power supply of the same model and type. The entire maintenance cycle could require an unscheduled downtime measured in minutes. However, if a new power supply is not available due to logistics problems, you could experience downtime lasting for days. Consequently, operational system availability would be low.

Operational availability is what really matters to IT managers because that is what they actually experience. It can be viewed as availability based on actual events that happened to the system (—a posterioriǁ availability). All other definitions of availability are less representative because they are predictions or estimations (—a prioriǁ availability) based on models of system failure and downtime distribution.

The responsibility for operational availability is shared by the manufacturer and the user. Improper use of a system can degrade operational availability and cannot be attributed to low product quality or a poor support organization. For example, operating high-tech electronic devices intended for data center use in a dusty environment can result in system failures. Human errors in configuring a system can also lead to reduced operational availability. In both examples, product reliability may be excellent, but not operational availability.

The equation for operational availability is this:

$$A_{op} = \frac{Uptime}{Operating\ cycle}$$

Here, the operating cycle is the overall operation time period being investigated (typically one year), and uptime is the total amount of non consecutive time that the system was functioning during the operating cycle. Availability is often expressed as a percentage, with a value of 99.999 percent or more associated with networking devices that meet the highest standards.

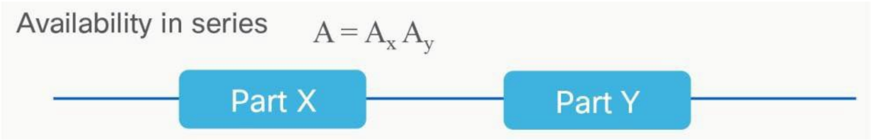
In some cases, during a failure condition, a system may still be running but with degraded performance. For example, a car may have a broken window. This is not preventing its use, but the damage may not provide an optimal driving experience during freezing weather. In IT, there is a growing trend to treat a significant performance degradation in IT resources as a failure. For example, several market analyses indicate that a user accessing a webpage that takes longer than 5 seconds to respond will leave that page and go elsewhere. This slow response time may be the consequence of a network link failure that results in insufficient bandwidth to serve the application's needs. Or all resources may be formally operational but in a degraded mode. That is why some organizations treat a device with significantly reduced performance as a failed device and try to work around it by rerouting traffic to alternative paths. Practically, because organizations generally like to use whatever resources are available close to their limits plus a little safety margin, most IT managers say that if performance degrades by more than 25 percent, they will consider the system to be failed. This perspective is important to understand when considering architectural differences among networking devices and the impact of failures of critical infrastructural components.

Calculating System Availability and Downtime

Specific techniques are used to calculate system availability from the availability information of a system's components. System availability is calculated by modeling the system as an interconnection of parts, in a series or in parallel.

Two parts are considered to be operating in a series if the failure of one part makes the combination inoperable. For example, the controller in a network switch is a single device that includes two parts: a CPU and a RAM unit. If one of these two components fails, the switch will go offline. In mathematical terms, the availability of the device is the product of the availability of the two parts, as shown by the equation in Figure 25.

Figure 25. Availability Calculation for a Series of Two Components



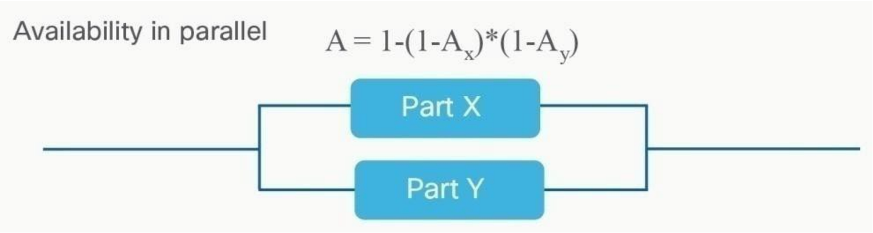
An implication of this equation is that the combined availability of two components in a series is always lower than the availability of its individual components. This is why we say that a chain is as strong as its weakest link. Good system designers work hard to select components that provide the highest availability because a single component with poor availability will compromise the entire system under review. This is why many products that appear to be similar from the outside can actually have much different levels of availability.

Two parts are considered to be operating in parallel if the failure of one part causes the other part to take over the operations of the failed part. The combination is considered failed only when both parts fail, and it is considered operational when either part is operational. For example, consider two power supplies that provide energy to a computing server. If one of the two fails, the remaining one takes over the full load and keeps the server fully operational. Under normal operation, the two power supplies would likely share the load, but if either of them fails, the remaining one can sustain the entire load.

The control units of modular switches, often called supervisors, take a somewhat different approach to parallel redundant components. The control units are redundant, but they do not share the load. They operate using an active-standby model. The hardware unit that performs the functions under normal conditions is the active one, and the redundant unit is in standby mode. The standby unit monitors the active unit at all times, and it will become active and take over operations if the active unit fails. Because the standby unit has to take over under fault conditions, it must stay synchronized with the active unit.

Regardless of whether the load -sharing or the active-standby approach is used, the combined availability of the two parts in parallel is equal to 1 minus the unavailability of both parts. The mathematical calculation for this combined availability is shown in the equation in Figure 26.

Figure 26. Availability Calculation for Two Components in Parallel



An implication of this equation is that the combined availability of two components in parallel is always much higher than the availability of the individual components. Therefore, even if Part X with a very low availability is used, the overall availability of the system with two identical parts running in parallel (Part X = Part Y) will be much higher. Thus, parallel operation provides a powerful mechanism for creating a highly reliable system even with the use of components and assemblies whose complexity leads them to achieve a lower availability rating. For this reason, all mission-critical systems are designed with redundant components. When a component is redundant, the system can continue without it. A single one of the two equivalent components can do the job in its entirety, without loss of function.

Redundancy also usually costs more, so design engineers should find the appropriate balance between the use of serial and parallel approaches. Typically, the greatest percentage increase in system reliability is achieved by deploying the least reliable component in a redundant parallel configuration. The use of more components in parallel usually boosts system availability.

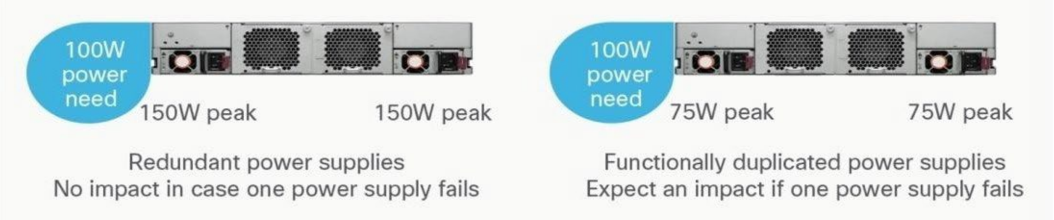
For instance, in the power supply example, a top-of-rack network switch typically requires a single power supply to operate. This arrangement is referred to as N=1, where N means necessary. Because the power supply is vital to the entire system (and often its availability is in the 99.98 percent range), a best practice is to deploy two of them. This arrangement is called 1+1 redundancy, or N+1 redundancy, where N=1. With this approach, the system will tolerate the failure of a single power supply.

You could also, if you wanted, deploy three power supplies in the switch, achieving N+2 redundancy, where N=1. In this case, the system could tolerate the simultaneous failure of two power supplies and stay fully operational. This kind of implementation is seldom seen in the real world in electronic devices because of its cost. However, it is often used in the IT industry for mechanical components. For example, the hard-disk drive (HDD) protection schema known as RAID 6 is an instance of an N+2 redundancy model. Obviously, the more redundancy that is built in to a system, the greater the expected availability. In reality, however, you need to balance cost and complexity trade-offs, so although redundancy can improve availability, excessive redundancy is considered a waste of resources.

One additional distinction is the difference between true redundancy and functional duplication. A system is considered to have redundancy when two elements perform the same function but only one is strictly necessary for proper operation. The second element is intended to take over all functions of the first in the event that the first element fails. The redundancy schema in use can be active-standby or load sharing, depending on other design considerations. In the active-standby model, one unit performs all the work, and the other unit is essentially idle, waiting for a failure to occur to the active unit. In the case of the load-sharing model, both units are operational, and each supports half of the required load. In the event of a failure, the surviving unit takes over the entire load. Whatever the schema, one unit is necessary for proper operation, and one is redundant. For example, a network switch may consume 100 watts (W) of power. A power supply of 150W can be selected to power the switch with some safety margin to accommodate variable environment conditions and traffic loads. To provide redundancy, a second, identical power supply is used. Therefore, for a typical power consumption of 100W, the switch is equipped with two power supplies of 150W each. This is a truly redundant configuration for the power supplies.

If, instead, functional duplication is used, the system still has two identical units than can perform the same functions. However, now both units are required to achieve proper system operation, and the only possible schema is load sharing. Individually, the units cannot sustain the entire workload. If one of the two load-sharing units fails, the remaining one will keep working but will not be able to support the entire workload for the system. Hence, performance degradation will occur. In practical terms, the network switch that consumes 100W could be equipped with two power supplies of 75W each, providing some safety margin to accommodate variable environment conditions and traffic loads. Under normal operation, the combined power supplies would be able to deliver all the required power for the switch. However, if one power supply fails, the switch will not be able to get all the power it needs and would eventually enter a state of graceful performance reduction, disabling some ports or capping their bandwidth. It would still be up and running, and it could still be configured and monitored, but its capabilities would be reduced. In such a scenario, during a failure the system cannot be considered fully available, but nor is it completely unavailable. This intermediate situation is often referred to as partial operational availability, and it is shown in Figure 27.

Figure 27. Redundancy and Functional Duplication for Power Supply Units


100W power need
150W peak
150W peak
Redundant power supplies
No impact in case one power supply fails
100W power need
75W peak
75W peak
Functionally duplicated power supplies
Expect an impact if one power supply fails

Humans have some organs with functional duplication. We have two legs, and they perform the same function, but they are both necessary to walk properly. We have two eyes, and they perform the same function, but they are both necessary for a 114-degree binocular field of view, and 3D vision, and depth perception. These are examples of functional duplication, not true redundancy. A system with functional duplication can actually be considered truly redundant only if the expectations for its capabilities are halved. For example, if you can accept that in a 24-port switch, only 12 are operational when the switch receives 50W of power, you can turn functional duplication into true redundancy by reducing your expectations for the switch's functions. The drawback to this approach is that you will need two switches to sustain the load you are expecting on them.

As you can see, functional duplication is necessary but not sufficient to achieve true redundancy. Functional duplication can be used as a means to help reduce costs: the system contains no superfluous hardware—everything is needed. However, if a hardware failure occurs during a busy period, system performance will be suboptimal until the failed module is replaced, with all the negative consequences of that scenario.

System availability indicates the percentage of time that the system is operational. Availability typically is specified in nines notation. For example, three-nines availability corresponds to 99.9 percent availability. Five-nines availability corresponds to 99.999 percent availability. Downtime per year is a more intuitive way of understanding availability. It is equivalent to 1 minus availability. Table 4 shows availability and the corresponding downtime.

Table 4. Availability and Corresponding Downtime Values

Availability	Downtime
90% (1-nine)	36.5 days/year
99% (2-nines)	3.65 days/year
99.9% (3-nines)	8.76 hours/year
99.99% (4-nines)	52 minutes/year
99.999% (5-nines)	5 minutes/year
99.9999% (6-nines)	31 seconds/year
99.9999% (7-nines)	3.1 seconds/year

Hardware Reliability: Failure Rate, FIT, MTTF, MTBF, and MTTR

IT professionals often talk about uptime, downtime, availability, and system failures. However, hardware reliability cannot be properly understood without an understanding of quantitative measures such as failure rate, failures in time (FIT), mean time to failure (MTTF), mean time between failures (MTBF), and mean time to restore (MTTR). This complex topic is the specific focus of reliability engineering subject-matter experts, and definitions sometimes overlap. The goal of this section is simply to clarify these terms and their use in relation to system availability.

First, you need to consider what exactly qualifies as a failure. Clearly, if the system is down, it has failed. When your car cannot start, it has failed. But what about a system running in degraded mode? For example, if your car has a major problem in the fuel pump and can run at only 1 mile per hour, would you consider it failed or not?

Technically speaking, a failure is declared when a system does not meet its desired objectives. But this definition involves a degree of subjectivity. Who determines the desired objective? Should an IT system be considered failed only when downtime or a complete outage time occurs, or also when the system is running slowly even though it is technically still up and available? The minimum performance requirements for the system need to be defined.

In general, performance degradation exceeding 25 percent is considered to be a failure. Organizations that make purchasing decisions for new network devices expect to have full use of those devices. If a device that supports 16 ports operating at 16-Gbps Fibre Channel connectivity is purchased, the organization will not want it to be able to operate only 8 ports at 16-Gbps Fibre Channel connectivity. In that case, some hosts will not be able to communicate, and the IT manager likely will consider the switch failed and seek remediation.

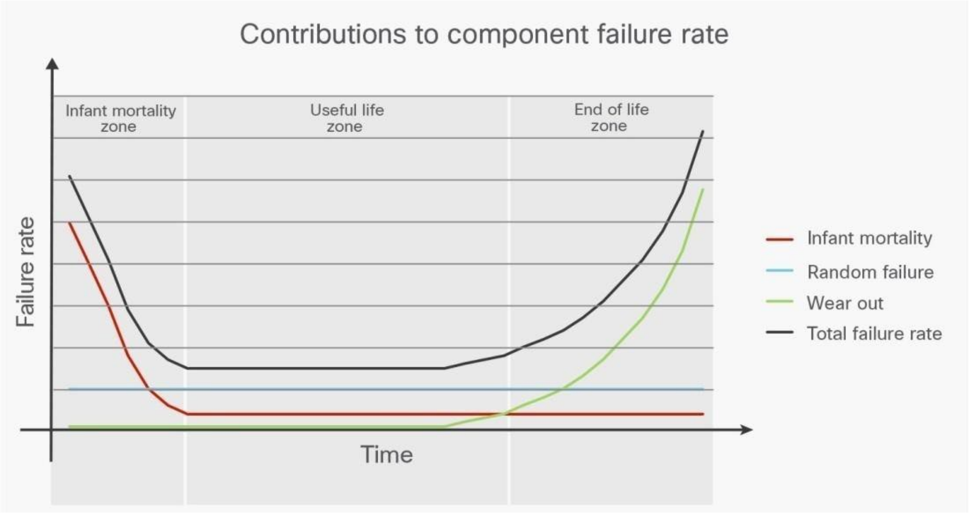
Similarly, the same device could have all 16 ports operational, but with half the bandwidth available to each. In this second example, however, all connected hosts can still communicate, even though significantly more slowly than expected. Some IT managers would consider such a switch operational despite its degraded performance, and others would consider it failed. With today's expectation of availability 24 hours a day, 7 days a week, an increasing number of IT managers would consider the switch to be failed. The only way to consider the switch to be operational would be to expect the traffic load on the switch to always be less than 50 percent of its full capability. However, this approach has cost implications if you need to purchase twice the number of devices to transport the desired amount of data. For example, you might purchase a 192-port 16-Gbps Fibre Channel modular platform with an aggregate bandwidth of 3 Tbps but use it to carry only 1.5 Tbps of traffic. Such a high underutilization rate would allow you to consider the platform operational even if a failure reduces its bandwidth by half. But how many IT managers would consider this a smart decision?

The failure rate parameter is frequently used in reliability engineering. It is often denoted by the Greek letter λ (lambda). It represents the frequency with which an engineered system or component fails, expressed in failures per unit of time. The failure rate of a system varies with time, depending on when in the lifecycle of the system you calculate it. For example, a car tends to have a higher failure rate when it is four years old than when it is new for reasons tied to the age of components. Hardware failures during a product lifecycle can be grouped into four main classes:

- Design failures: Inherent design flaws in the system can lead to failures. For example, an active component that cannot be properly cooled by an undersized heat sink is destined to fail soon.
- Infant mortality: Newly manufactured hardware has a greater tendency to fail. In this type of failure, the weaker units fail early, usually because of manufacturing problems such as poor soldering, leaking capacitors, etc. These failures are normally caught during in-factory system burn-in tests. However, sometimes problems remain in systems leaving the factory, becoming systems dead on arrival (DOA).
- Random failures: This type of failures occurs over the entire life of a hardware module. Random failures can affect the entire system. Redundancy provides a way to recover from these failures.
- Wear-out: When a hardware module reaches the end of its useful life, chances of failure increase due to degradation of components. For example, the wear-out in power supplies is usually caused by the breakdown of electrical components as a result of physical wear and electrical and thermal stress. This type of failures can be reduced by refreshing the entire system every few years. In the IT sector, a 4-year renewal cycle is common. This cycle also helps IT managers incorporate in industry's continuous technological innovation. Preventive maintenance is a recent approach for some components such as flash storage, in which wear-out begins in the early phase of the lifecycle but can be monitored.

The graph in Figure 28 shows the contribution of the different classes of failures to the total failure rate. Design failures are not shown and are not included in this discussion. The figure also shows the three generally accepted phases of a product lifetime.

Figure 28. Contributors to Component Failure Rate and the Bathtub Shape



Hardware failures are typically characterized by a bathtub curve. As you can see, the failure rate is not constant, but generally goes through three phases over the device’s lifetime. The chances of a hardware failure are high during the initial life of the module. In the flat region of the bathtub curve, known as the useful life of the product, the failure rate is fairly low and approximately constant. This is the value used for reliability calculations, as specified in the Telcordia Technical Reference TR-332 document ―Reliability Prediction Procedure for Electronic Equipment.‖ After the end of life is reached, the failure rate increases again. Sometimes, for easier understanding, the failure rate is expressed as a percentage per unit of time (for example, 0.2 percent per year).

Similar to the failure rate and directly correlated with it, the FIT value of a device is the number of failures that can be expected in 1 billion (10^9) device hours of operation. With the unit of time set very high, to 10^9 hours, components with extremely high reliability have a FIT value that is an integer, which can be used easily in calculations. Remember that reliability engineering is based on statistics: nobody can really wait 1 billion hours (114,000 years) to see whether a device will fail. The practical approach is to analyze a pool of identical devices over a given time period to see how many of them fail. For example, a population of devices could consist of 1000 units that are run for 1 million hours, or 1 million devices that are in operation for 1000 hours each, or some other combination. The FIT parameter is used particularly by the electronics industry and in the Telcordia TR-332 recommendation, and it is normally considered more intuitive than alternatives.

MTTF is the predicted elapsed time to failure for a nonrepairable component during operation. MTTF can be empirically determined in the field as the arithmetic average of time to failure for a pool of components of the same kind. In other words, MTTF is calculated as the number of total hours of service for all devices divided by the number of devices. Because it is a statistical parameter, the confidence level of the result directly depends on the sample size. The MTTF value can be determined empirically when a reasonable quantity of devices has been shipped and operated for some time.

For example, consider the population of Cisco QSFP-40GE-SR-BD transceivers that have been sold. Those transceivers are also used on Cisco MDS 9000 Family 24-port 40-Gbps FCoE line cards. In the first 12 month of shipping, approximately 450.000 units of that specific transceiver went into operation. This translates to an estimated total of 1.9 x 10^9 hours of operation. Out of all shipped transceivers, a tiny fraction were returned to Cisco with a failure condition. The field-calculated MTTF value for those transceivers was calculated as 2.2 x 10^6 hours (equivalent to 250 years). In other words, on average, you would need to have 250 transceivers in operation for one year if you want to see one of them fail. As you can see, the MTTF value is approximately 50 times longer than the expected lifetime of the transceiver itself. Nobody would expect a single transceiver to be operational after as many years as its MTTF value.

Now consider another example. A typical 3.5-inch 15,000-rpm HDD might present an MTTF value of 1,600,000 hours, or more than 180 years, equivalent to a 625 FIT rate. But no one expects a given HDD to last this long. In fact, the disk replacement rate is much greater than the disk failure rate. Practically, if you have 180 HDDs up and running, you can expect one of them to fail in the course of one year.

The theoretical MTTF value can also be calculated based on specific lab testing under high-stress conditions such as extreme temperature cycles. When these accelerated life tests are used, techniques based on the Weibull distribution and graphical plots provide accurate failure analysis and risk predictions from only a small number of samples. In particular, the Arrhenius equation evaluates the way that increased temperature accelerates the age of a product compared to its normal operating temperature. Reliability engineers often use reliability software to calculate a product's MTTF according to various methods and standards, such as the Telcordia Technical Reference TR-332.

For complex systems that use multiple components and that can thus be repaired, the MTBF parameter is used instead of MTTF. The MTBF value is calculated assuming that a failure event takes the system out of service. For example, if a system has redundant power supplies and one fails, the system continues to work. The failure of the power supply unit is not considered a system failure in this case. However, if the system chassis catches on fire, it will be considered failed. In general, MTBF is the uptime between two failure states of a repairable system during operation.

Note that MTBF is an ensemble characteristic, and it applies to populations of identical devices. It is not a sample characteristic that applies to one specific device. Therefore, estimating the lifetime of a specific device from its MTBF value would be highly misleading. For example, the MTBF for humans is in the range of 900 years, not their lifetime. A power supply can have an MTBF value of 200,000 hours, but that does not mean that one sample of that power supply model should be expected to last for approximately 200,000 hours. The lifetime for one sample of that power supply model will more likely be in the range of 43,000 hours, so much shorter than the MTBF value. Electronic components in general are designed so that wear-out does not occur during the expected useful life of units (exceptions just confirm the rule). Efforts are made to help ensure that the MTBF value is much longer than the expected lifetime of a device and that any failures during that time period are random.

The failure rate is strictly related to the MTBF value according to the equation MTBF = 1 / λ. This statement is valid when the failure rate can be assumed to be constant, in the flat region of the bathtub curve known as the useful life period, rather than during the initial and late phases of the system lifecycle. So MTBF is the inverse of the failure rate during the constant failure rate phase. Therefore, you should not use MTBF to extrapolate the service life of a component, which will typically be much less than what the MTBF value suggests, mainly due to the much higher failure rates in the end-of-life wear-out phase of the bathtub curve. For accuracy, the MTBF value should be measured in device hours instead of in hours only, but most documents ignore the statistical nature of this parameter and quantify it only in hours.

Just as MTBF and failure rate are related, so are the MTBF and FIT values. The relationship of the FIT value to the MTBF value can be expressed as MTBF = 1 billion x 1 / FIT. A device with FIT value of 1 has an MTBF value of 1 billion hours. Although the Telcordia TR-332 standard uses the FIT value as the primary parameter for its calculations, other standards tend to use the MTBF value.

Knowledge of the MTBF value allows you to calculate the probability that any one particular module from a pool of identical units will be operational at a time equal to the MTBF value. In other words, you can calculate the reliability of a module at a specified time. You can use the following exponential equation, which is valid under the assumption that failures occur randomly (during the useful life period):

$$\mathrm{Reliability} = \mathrm{R(t)} = \mathrm{e}^{-\mathrm{t}/\mathrm{MTBF}}$$

The probability that any one particular module will survive to its calculated MTBF value is about 36.8 percent.
In fact, when t = MTBF, then R(t) = e^{-1} = 0.3677. When the MTBF value is much longer than the expected lifetime, which is typical for electronic devices, it is interesting to know the chances that a unit will be operational at the end of its expected lifetime. Assume that the MTBF value is 10 times the expected lifetime, which is a valid assumption for IT equipment in general. In this case, the probability that any particular unit will be operational at the end of its expected lifetime is R(t) = $e^{-1/10}$ = 0.9, or 90 percent. And with an MTBF value of up to 100 times the expected lifetime, R(t) = $e^{-1/100}$ = 0.99, or 99 percent.

MTTR is a basic measure of the maintainability of repairable items. It represents the average time required to restore the correct operational state for a failed device. Five main factors contribute to this number:

- Detection time: This time includes diagnostic analyses and identification and localization of the failed component.
- Setup time: This time includes the identification of the appropriate repair action and any preliminary activities.
- Procurement time: This time is the sum of the administrative delay time (ADT) and logistic delay time (LDT) required to obtain any needed components.
- Effective repair time: This is the time needed to manually perform the repairs.
- Recovery time: This time includes reconfiguration (if necessary) and cleanup of alarms.

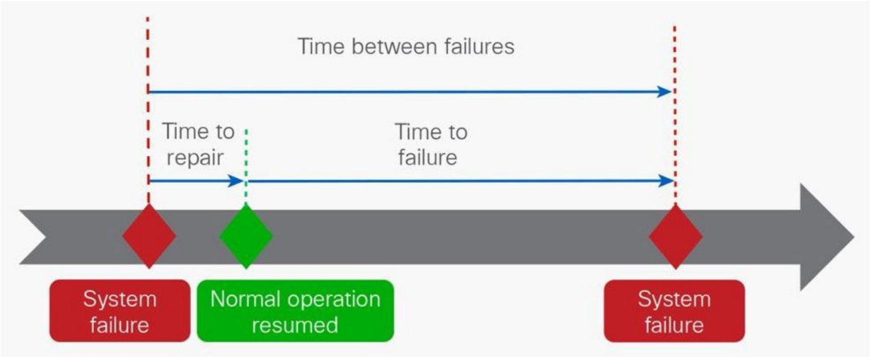
Some people exclude the procurement time from MTTR and introduce the concept of mean downtime (MDT) as the sum of the MTTR plus the procurement time. This practice is not described in this document.

In high-tech electronic equipment, a failed component will not be repaired until the failure is detected. For this reason, fault-tolerant equipment, such as Fibre Channel networking devices, implement self-diagnostic procedures to quickly identify and report failures for individual functional components. For a reparable electronic device, the repair action is often identified as the replacement of the failed module with a new one, at times preceded by some preliminary activity such as removing some screws or cover panels. The hot-swap capability claimed by system vendors is important in this context. The effective repair time in itself may be very short, taking perhaps just one minute. Recovery time may be needed to get the new module recognized and to perform some reconfiguration activities. The time needed to get a system fully recovered is also greatly influenced by the onsite availability of spare parts. This procurement time is normally the time commitment to which the vendor and customer agree when signing a support contract for the system.

The MTTR, including the lead time for parts and administrative and logistic activities, thus is part of a maintenance contract because it has a direct impact on the operational availability of the system. This is why a vendor's support service, such as Cisco's award-winning worldwide support organization, is so important to system operational availability. The shorter the MTTR, the greater the operational availability.

Mathematically, MTBF is the sum of MTTR and MTTF, as shown in Figure 29. Practically, for any network device today, MTTF is so much longer than MTTR that the numerical difference between MTBF and MTTF is within the range of statistical error.

Figure 29. Relationship of MTTF, MTBF, and MTTR



Hardware reliability and MTTR determine the availability of a system. Availability of a hardware or software

$$A_{op} = \frac{\text{MTBF}}{\text{MTBF} + \text{MTTR}}$$

module can be obtained with the following formula:

If MTTR = 0 (which is impossible), availability would be 100 percent. Practically, what matters is the ratio of MTTR to MTBF: the lower the value, the greater the availability. Thus, an MTTR value of 24 hours, for example, may or may not be acceptable, depending on the MTBF value and the desired availability target.

Software Reliability, Stability, and Fault Domains

The same availability concepts that apply to hardware apply to software as well, though they may be more challenging to understand.

One difference is that software failures are design or implementation problems, not wear and tear problems. Software doesn't wear out with use. Software is a set of instructions for a piece of hardware to follow. It has no moving parts, so nothing can physically deteriorate. Software fails because of design errors. So whereas a hardware module will fail from time to time as parts wear out and the device ages, software will work fine until it encounters inputs or conditions it was not designed to handle. Then it will fail completely (crash) or result in unexpected and unintended behavior. New code tends to be more prone to these problems, but they can be avoided through proper design using a modular architecture to reduce the inevitable entropy as changes are incorporated into the code.

Software failures can be characterized by the software defect density for a system, and a numerical value can be obtained by keeping track of the software defect history. Defect density depends on the following factors:

- Software process used to develop the design and code
- Complexity of the software
- Size of the software
- Experience of the team developing the software
- Percentage of code reused from a previous stable project
- Rigor and depth of testing before product shipment.

Defect density is typically measured in the number of defects per thousand lines of (source) code (defects divided by the KLOC value).

In addition to applying to hardware, the concept of MTBF applies to software. In this case, the MTBF value can be numerically determined by multiplying the defect rate by KLOCs processed per second. Similarly, MTTR for a software module can be computed as the mean time needed to reboot after a software fault is detected.

Software developers also use other metrics to track the quality of software. One metric, the defect detection percentage, measures the percentage of severity-level 1, 2, and 3 customer-found defects (CFDs) are found in relation to the total number of severity-level 1, 2, and 3 defects found over a one-year period. This metric is interesting because externally found defects are more expensive to fix than internally found defects, and a poor value for this metric suggests a need for deeper testing and more time spent on quality control and assurance.

Note that in high-tech electronic devices that include both hardware and software for their operation, the software and the hardware for any specific module operate in a series, because the module cannot function if the hardware or the software is not operational. The importance of software quality and reliability is thus apparent.

Often for software, the terms –reliability‖ and -stability‖ are used interchangeably. The quality model presented in the first part of the ISO and IEC 9126-1 standard classifies software quality as a structured set of characteristics and subcharacteristics, as shown in Table 5.

Table 5. ISO and IEC 9126-1 Quality Model

Characteristic	Subcharacteristic
Function	Suitability
	Accuracy
	Interoperability
	Security
	Functional compliance
Reliability	Maturity
	Fault tolerance
	Recoverability
	Reliability compliance
Usability	Understandability
	Learnability
	Operability
	Attractiveness
	Usability compliance
Efficiency	Time behavior
	Resource utilization
	Efficiency compliance
Maintainability	Analyzability
	Changeability
	Stability
	Testability
	Maintainability compliance
Portability	Adaptability
	Installability
	Coexistence
	Replaceability
	Portability compliance

Reliability defines the capability of the software to maintain its services under defined conditions for defined periods of time. One aspect of this characteristic is fault tolerance: the capability of a system to withstand component failures.

Stability characterizes the software's sensitivity to change and the negative impact that system changes may have on the software. Stability is a subcharacteristics of maintainability. The maintainability characteristic addresses the capability to identify and fix a fault within a software component. Maintainability is affected by the readability and complexity of code as well as by modularization. The capability to verify (or test) the software is also very important for stability.

The firmware and software combination running on a network device is a key contributor to the overall operational system availability. One of the most important achievements related to availability is ISSU. This capability, added in the early 2000s on Fibre Channel switches, allows nondisruptive software upgrades to improve functions and address any software errors in previous versions. Administrators can load new software without having to shut down the switch or reinitialize it. Traffic keeps flowing unaffected. ISSU provides users with greater availability and reliability for their SANs. This capability is available on switches and director-class devices, both traditional and mission-critical, but there are subtle implementation differences because a director-class device benefits from a dual supervisor, whereas a switch has a single control-plane unit. As a result, ISSU on switches imposes some constraints on the fabric in which they reside. The fabric must be in a stable operating condition, without changes of any sort during the upgrade cycle.

Fault domains are used to help make sure that a software problem has limited impact on the entire solution. For example, consider logical virtual partitioning technology, pioneered by Cisco in 2003 as virtual SAN (VSAN) technology and ratified as a standard by the INCITS T11 committee in 2004 as virtual fabric technology. With this partitioning technology, a single Fibre Channel fabric can be treated like multiple logical independent fabrics. Every virtual partition has its own processes for the name server, zoning, quality of service (QoS), and more. The failure of one virtual partition will not compromise operation for the others, providing a high level of fault containment.

Although they have no direct effect on hardware and software reliability, advanced manageability tools are essential to achieve high availability for a solution. They can help quickly identify possible misbehavior and fix it. They also can help storage administrators predictably manage configuration changes, increase configuration integrity, and reduce the risk of improper operation due to human error.

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 C11-738426-00 02/17