CISCO SYSTEMS

White Paper

"A Day in the Life of a Fibre Channel Frame"
Cisco MDS 9000 Family Switch Architecture

Cisco Systems® released the first-generation Cisco® MDS 9000 Family of high-density, multiprotocol, intelligent storage area network (SAN) switches in December 2002. The switches delivered intelligent features and capabilities never seen before in the SAN switching marketplace, with higher port densities than any other SAN switch at the time.

With the introduction of second-generation linecards, supervisors, and chassis, Cisco further strengthens its leadership position in the SAN switching marketplace. The second-generation Cisco MDS 9000 Family linecards, supervisors, and chassis layer more intelligence into the SAN switching fabric, more than double the port densities offered by other SAN switch vendors, and provide investment protection with backward compatibility to existing chassis, linecards, and supervisors.

This white paper describes the day in a life of a Fibre Channel frame as it is switched within Cisco MDS 9000 Family switches. This paper provides a detailed walkthrough of the switch architecture, describing how various features and functions are implemented, and how the Cisco MDS 9000 Family offer investment protection across the entire range of supervisors, linecards, and chassis without loss of performance or functionality.

Figure 1. Cisco MDS 9000 Family Multilayer Switches

SYSTEM OVERVIEW

In December 2002, Cisco set the benchmark for high performance Fibre Channel Director switches with the release of the Cisco MDS 9509 Multilayer Director. Offering a maximum system density of 224 1/2-Gbps Fibre Channel ports, first-generation Cisco MDS 9000 Family linecards connected to dual crossbar switch fabrics with an aggregate of 80 Gbps full-duplex bandwidth per slot, resulting in a total system switching capacity of 1.44 Tbps.

In 2006, Cisco once again set the benchmark for high performance Fibre Channel Director switches. With the introduction of second-generation linecards, crossbars (supervisors), and the flagship Cisco MDS 9513 Multilayer Director, system density has grown to a maximum of 528 Fibre Channel ports operating at 1/2/4-Gbps or 44 10-Gbps Fibre Channel ports, increasing interface bandwidth per slot to 96-Gbps full-duplex, resulting in a total system switching capacity of 2.2 Tbps.

Supervisor-2 and second-generation, higher-density linecards enable the existing 9-slot Cisco MDS 9509 Multilayer Director to scale to a maximum system density of 336 1/2/4-Gbps Fibre Channel ports and enables the 6-slot Cisco MDS 9506 Multilayer Director to grow to a maximum of 192 1/2/4-Gbps Fibre Channel ports.

As well as increasing port speeds and density, Cisco also offers solid investment protection. All existing first-generation Cisco MDS linecards can operate alongside second-generation linecards, chassis, and supervisors. Conversely, second-generation Cisco MDS linecards can be used in conjunction with existing first-generation linecards, chassis, and supervisors.

CISCO MDS SYSTEM ARCHITECTURE

All Cisco MDS 9500 Series director switches are based on the same underlying crossbar architecture. Frame forwarding logic is distributed in application-specific integrated circuits (ASICs) on the linecards themselves, resulting in a distributed forwarding architecture. There is never any requirement for frames to be forwarded to the supervisor for centralized forwarding, nor is there ever any requirement for forwarding to drop back into software for processing—all frame forwarding is performed in dedicated forwarding hardware and distributed across all linecards in the system.
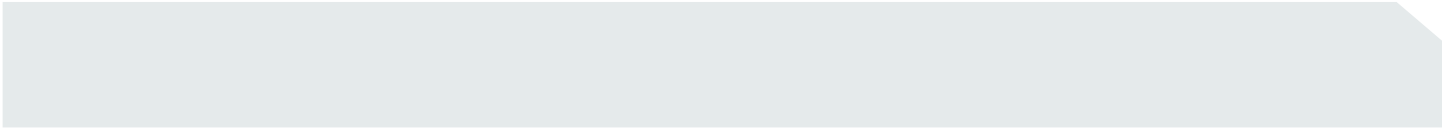
All advanced features, including virtual SANs (VSANs), VSAN Trunking, Inter-VSAN Routing (IVR) including Fibre Channel Network Address Translation (FC-NAT), Cisco PortChannels (port aggregation), quality of service (QoS), Fibre Channel Congestion Control (FCC), Switch Port Analyzer (SPAN), and Remote Switch Port Analyzer (RSPAN) are implemented within the hardware-based forwarding path and can be enabled without any loss of performance or additional latency. All of these advanced features were built into the Cisco MDS 9000 Family ASICs from the beginning. Second-generation Cisco MDS ASICs augment these capabilities with the addition of class of service (CoS) and port bandwidth reservation capabilities, with more hardware capabilities to be enabled in future Cisco MDS 9000 SAN-OS Software releases.

Forwarding of frames on the Cisco MDS always follows the same processing sequence:

1. Starting with frame error checking on ingress, the frame is tagged with its ingress port and VSAN, enabling the switch to handle duplicate addresses within different fabrics separated by VSANs.

2. Following the frame tagging, input access control lists (ACLs) are used to enforce hard zoning.

3. A lookup is issued to the forwarding and frame routing tables to determine where to switch the frame to and whether to route the frame to a new VSAN and/or rewrite the source/destination of the frame if IVR is being used.

4. If there are multiple equal-cost paths for a given destination device, the switch will choose an egress port based on the load-balancing policy configured for that particular VSAN.

5. If the destination port is congested or busy, the frame will be queued on the ingress linecard for delivery, making use of QoS policy maps to determine the order in which frames are scheduled.

6. When a frame is scheduled for transmission, it is transferred across the crossbar switch fabric to the egress linecard where there is a further output ACL, and the VSAN tag used internal to the switch is stripped from the front of the frame and the frame is transmitted out the output port.

Dual redundant crossbar switch fabrics are used on all director switches to provide low-latency, high-throughput, non-blocking, and non over-subscribed switching capacity between linecard modules.

All current linecard modules utilize crossbar switch fabric capacity with 1:1 redundancy. That is, there are always one active channel per crossbar and one standby channel. One active channel and one standby channel are used on each of crossbar switch fabrics, resulting in both crossbars being active (1:1 redundancy). In the event of a crossbar failure, the standby channel on the remaining crossbar becomes active, resulting in an identical amount of active crossbar switching capacity regardless of whether the system is operating with one or two crossbar switch fabrics.

From a switch architecture perspective, crossbar capacity is provided to each linecard slot as a small number of high-bandwidth channels. This provides significant flexibility for linecard module design and architecture and makes it possible to build both performance-optimized (non-blocking, non over-subscribed) linecards and host-optimized (non-blocking, over-subscribed) linecards, multiprotocol (Small Computer System Interface over IP [iSCSI] and Fibre Channel over IP [FCIP]) and intelligent (storage services) linecards. These crossbar channels enabled Cisco to 'future-proof' the switch for investment protection. They also enable port-speed increases (2 Gbps to 4 Gbps and 10 Gbps) without any need to replace existing linecard modules and supervisor modules.

To ensure fairness between ports, frame switching uses a technique called virtual output queuing (VOQ). Combined with centralized arbitration, this ensures that when multiple input ports on multiple linecards are contending for a congested output port, there is complete fairness between ports regardless of port location within the switch. It also ensures that congestion on one port does not result in blocking of traffic to other ports ("head-of-line blocking"). In addition, if priority is desired for a given set of traffic flows, QoS and CoS can be used to give priority to those traffic flows.

All Cisco MDS 9500 Series director switches are fully redundant with no single point of failure. The system has dual supervisors, crossbar switch fabrics, clock modules, and power supplies and employs majority signal voting wherever 1:1 redundancy would not be appropriate.
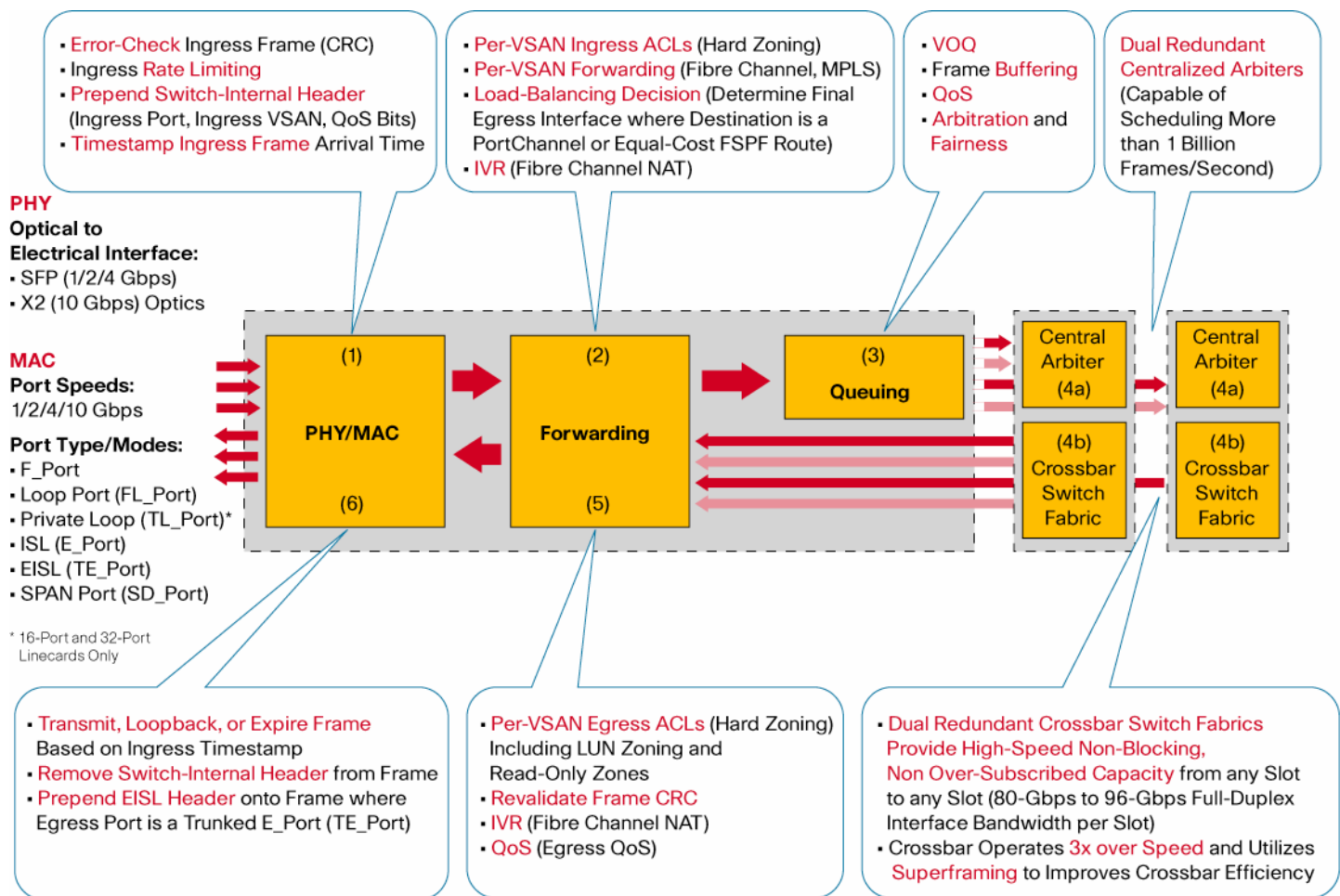
A side-mounted fan tray provides active cooling. Although there is a single fan tray, there are multiple redundant fans with dual fan controllers. All fans are variable speed, and failure of one or more individual fans causes all other fans to speed up to compensate. In the unlikely event of multiple simultaneous fan failures, the fan tray still provides sufficient cooling to keep the switch operational.

Redundant power supplies provide power to linecards in the form of a 42VDC power bus. Before linecards are powered up, they must first request power from the supervisor modules. These will only grant power to that linecard module if there is sufficient power within the system. The range of power supply options provides additional future compatibility for linecards, regardless of their power draw.

A DAY IN THE LIFE OF A FIBRE CHANNEL FRAME

All Cisco MDS 9000 Family linecards use the same process for frame forwarding. The various frame-processing stages are outlined in detail in this section (see Figure 2).

Figure 2. Frame Flow Within Cisco MDS 9000 Family Switches

- Error-Check Ingress Frame (CRC)
- Ingress Rate Limiting
- Prepend Switch-Internal Header (Ingress Port, Ingress VSAN, QoS Bits)
- Timestamp Ingress Frame Arrival Time

- Per-VSAN Ingress ACLs (Hard Zoning)
- Per-VSAN Forwarding (Fibre Channel, MPLS)
- Load-Balancing Decision (Determine Final Egress Interface where Destination is a PortChannel or Equal-Cost FSPF Route)
- IVR (Fibre Channel NAT)

- VOQ
- Frame Buffering
- QoS
- Arbitration and Fairness

Dual Redundant Centralized Arbiters (Capable of Scheduling More than 1 Billion Frames/Second)

PHY
Optical to Electrical Interface:
- SFP (1/2/4 Gbps)
- X2 (10 Gbps) Optics

MAC
Port Speeds:
1/2/4/10 Gbps

Port Type/Modes:
- F_Port
- Loop Port (FL_Port)
- Private Loop (TL_Port)*
- ISL (E_Port)
- EISL (TE_Port)
- SPAN Port (SD_Port)

* 16-Port and 32-Port Linecards Only

(1) PHY/MAC (6)

(2) Forwarding (5)

(3) Queuing

Central Arbiter (4a) — Central Arbiter (4a)

(4b) Crossbar Switch Fabric — (4b) Crossbar Switch Fabric

- Transmit, Loopback, or Expire Frame Based on Ingress Timestamp
- Remove Switch-Internal Header from Frame
- Prepend EISL Header onto Frame where Egress Port is a Trunked E_Port (TE_Port)

- Per-VSAN Egress ACLs (Hard Zoning) Including LUN Zoning and Read-Only Zones
- Revalidate Frame CRC
- IVR (Fibre Channel NAT)
- QoS (Egress QoS)

- Dual Redundant Crossbar Switch Fabrics Provide High-Speed Non-Blocking, Non Over-Subscribed Capacity from any Slot to any Slot (80-Gbps to 96-Gbps Full-Duplex Interface Bandwidth per Slot)
- Crossbar Operates 3x over Speed and Utilizes Superframing to Improves Crossbar Efficiency

(1) Ingress PHY/MAC Modules

Fibre Channel frames enter the switch front-panel optical Fibre Channel ports through Small Form-Factor Pluggable (SFP) optics for 1/2/4-Gbps Fibre Channel interfaces or X2 transceiver optical modules for 10-Gbps Fibre Channel interfaces. Each front-panel port has an associated physical layer (PHY) device port and a media access controller (MAC). On ingress, the PHY converts optical signals received into electrical signals, sending the electrical stream of bits into the MAC. The primary function of the MAC is to decipher Fibre Channel frames from the incoming bit stream by identifying start-of-frame (SoF) and end-of-frame (EoF) markers in the bit stream, as well as other Fibre Channel signaling primitives.

In conjunction with frames being received, the MAC communicates with the forwarding and queuing modules and issues return buffer-credits (R_RDY primitives) to the sending device to recredit the received frames. Return buffer credits are issued if the ingress-queuing buffers have not exceeded an internal threshold. If ingress rate limiting has been enabled, return buffer credits are paced to match the configured throughput on the port.

Before forwarding the frame, the MAC prepends an internal switch header onto the frame, providing the forwarding module with details such as ingress port, type of port, ingress VSAN, frame QoS markings (if the ingress port is set to trust the sender, blank if not), and a timestamp of when the frame entered the switch. The concept of an internal switch header is an important architectural element of what enables the multiprotocol and multitransport capabilities of Cisco MDS 9000 Family switches. The MAC also checks that the received frame contains no errors by validating its cyclic redundancy check (CRC).

(2) Ingress Forwarding Module

The ingress forwarding module determines which output port on the switch to send the ingress frame to (see Figure 3).

Figure 3. Ingress Frame Forwarding Logic
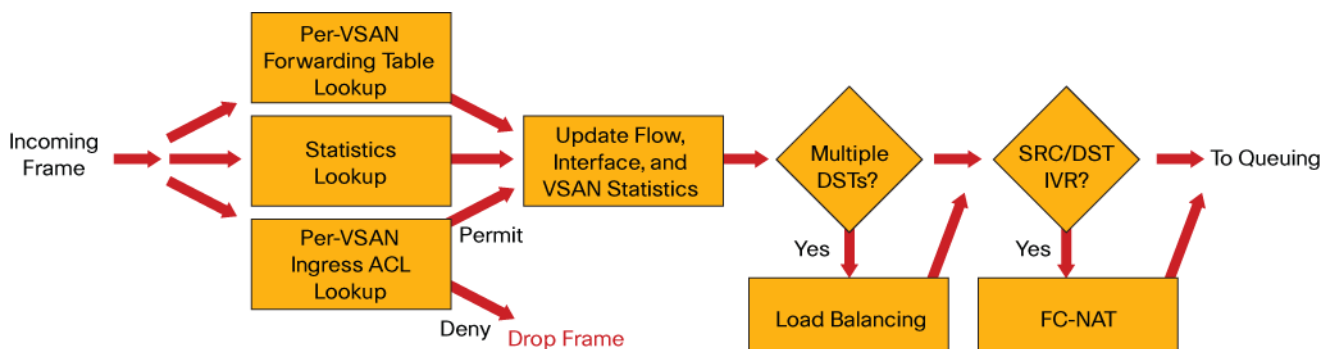
When a frame is received by the ingress forwarding module, three simultaneous lookups are initiated followed by forwarding logic based on the result of those lookups:

- The first of these is a per-VSAN forwarding table lookup determined by VSAN and destination address. The result from this lookup tells the forwarding module where to switch the frame to (egress port), based on the input port, VSAN, and destination address within the Fibre Channel frame. This table lookup also indicates whether there is a requirement for any IVR. If the lookup fails, the frame is dropped because there is no destination to which to forward to.
- The second lookup is a statistics lookup. The switch uses this to maintain a series of statistics about what devices are communicating between one another. Statistics gathered include frame and byte counters from a given source to a given destination.
- The third lookup is a per-VSAN ingress ACL lookup determined by VSAN, source address, destination address, ingress port, and a variety of other fields from the interswitch header and Fibre Channel frame header. The switch uses the result from this lookup to either permit the frame to be forwarded, drop the frame, or perform any additional inspection on the frame. This also forms the basis for the Cisco implementation of hard zoning within Fibre Channel.

If the frame has multiple possible egress ports (for example, if there are multiple equal-cost Fabric Shortest Path First [FSPF] routes or the destination is a Cisco PortChannel bundle), a load-balancing decision is made to choose a single physical egress interface from a set of interfaces. The load-balancing policy (and algorithm) can be configured on a per-VSAN basis to be either a hash of the source and destination addresses (S_ID, D_ID) or a hash also based on the exchange ID (S_ID, D_ID, OX_ID) of the frame. In this manner, all frames within the same flow (either between a single source to a single destination or within a single SCSI I/O operation) will always be forwarded on the same physical path, guaranteeing in-order delivery.

If traffic from a given source address to a given destination address is marked for IVR, then the final forwarding step is to rewrite the VSAN ID and optionally the source and destination addresses (S_ID, D_ID) of the frame.

The first-generation Cisco MDS forwarding complex is capable of forwarding up to 28.32 million frames per second. First-generation Cisco MDS linecards have either one or two forwarding complexes. The non-blocking, non over-subscribed 16-port linecard (part number DS-X9016) uses a pair of forwarding complexes with one per group of 8 front-panel Fibre Channel ports. This provides forwarding performance sufficient to sustain line-rate performance on all ports, even with minimum sized frames. The first-generation host-optimized over-subscribed non-blocking 32-port linecard (part number DS-X9032) uses a single forwarding complex for all 32 ports.

Second-generation Cisco MDS linecards increase both the port density and port speeds supported but require only a single forwarding complex per linecard. The second-generation Cisco MDS forwarding complex can forward up to 116 million frames per second, supporting up to 48 front-panel 1/2/4-Gbps Fibre Channel ports.

Both first-generation and second-generation Cisco MDS linecards support hard zoning for more than 2000 zones and 20,000 zone members. The forwarding adjacency tables support up to 128,000 unique destination addresses. Rewriting the VSAN, source Fibre Channel address, and destination Fibre Channel address is supported for up to 2000 IVR zones and 10,000 IVR zone members.

(3) Queuing Module

The queuing module is responsible for scheduling the flow of frames through the switch (see Figure 4). It uses distributed VOQ to prevent head-of-line blocking and is the functional block responsible for implementing QoS and fairness between frames contending for a congested output port. The queuing module is also responsible for providing frame buffering for ingress queuing if an output interface is congested.

Figure 4. Queuing Module Logic
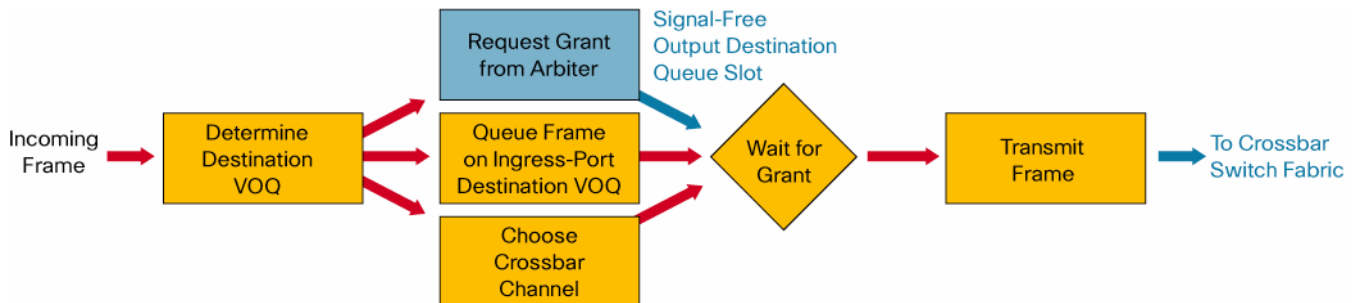
Frames enter the queuing module and are queued in a VOQ based on the output port to which the frame is due to be sent. Unicast Fibre Channel frames (frames with only a single destination) are queued in one virtual output queue. Multicast Fibre Channel frames are replicated across multiple virtual output queues for each output port to which a multicast frame is destined to be forwarded.

Four virtual output queues are available for each output port, corresponding to four QoS levels per port: one priority queue and three queues using deficit-weighted round robin (DWRR) scheduling (see Figure 5). Frames queued in the priority queue will be scheduled before traffic in any of the DWRR queues. Frames queued in the three DWRR queues are scheduled using whatever relative weightings are configured. The default relative weightings are 33/33/33; however, these can be set to any ratio at all.

Figure 5. Distributed Virtual Output Queuing

Ingress FC Port
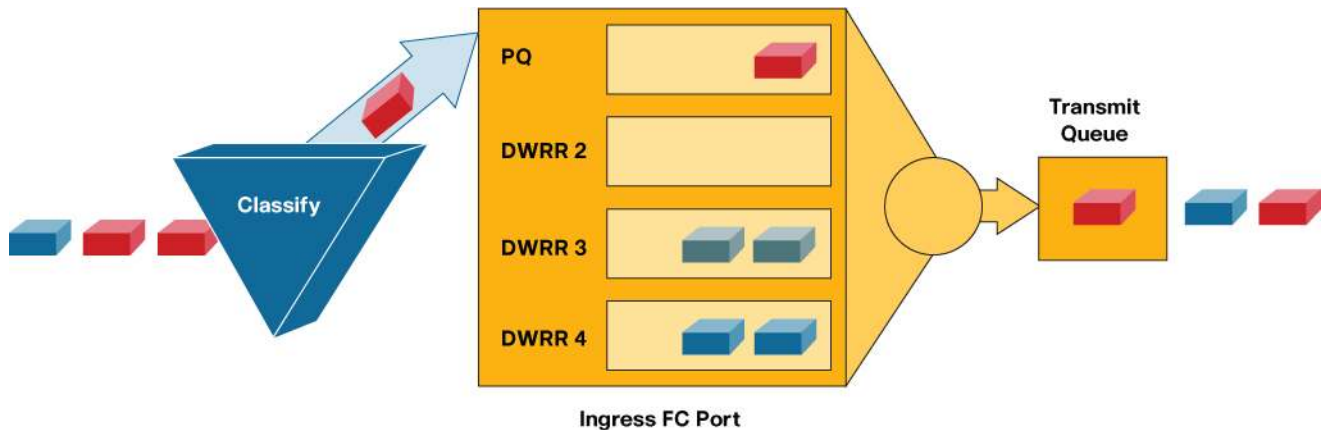
First-generation Cisco MDS linecards have 1024 available virtual output queues, allowing these linecards to address up to 256 destination ports with four QoS levels per port. Second-generation Cisco MDS linecards extend this to 4096 virtual output queues, increasing the number of addressable destination ports to a system architectural limit of 1024 ports per chassis, with four QoS levels per port.

In conjunction with frame queuing, the queuing module will issue a request to transmit the frame to the redundant central arbiters. The central arbiters are used by the system to help ensure fairness between frames in the event that the rate of traffic destined to a port exceeds the output transmission rate. The arbiter also prevents the crossbar congestion, thereby avoiding head-of-line blocking. Frames are queued (buffered) on ingress until the central arbiter issues a grant allowing the frames to be sent to the destination linecard/port.

The amount of buffering available varies depending on the type and generation of the linecard. First-generation Cisco MDS linecards provide up to 400 buffers per port (255 credited buffers and 145 performance buffers) on the 16-port, non over-subscribed, non-blocking (DS-X9016) linecard; 400 buffers shared across 8 ports (12 credited buffers per port) on the 32-port, over-subscribed, non-blocking (DS-X9032) linecard; and up to 3200 credited buffers per port on the MPS 14/2 (DS-X9302-14K9) linecard. Second-generation Cisco MDS linecards have 6000 buffers per module, allowing up to 4095 of these buffers to be dedicated to a single port. All buffers are capable of holding full-sized Fibre Channel frames (2148 byte frames).

After the central arbiter has issued a grant to the queuing module, the queuing module will transmit the frame across one of the four crossbar switch fabric channels.

(4a) Central Arbitration Module

The central arbiters are responsible for scheduling frame transmission from ingress to egress, granting requests to transmit from ingress forwarding modules whenever there is an empty slot in the output port buffers. They are capable of scheduling more than a billion frames per second.

Any time there is a frame to be transferred from ingress to egress, the originating linecard will issue a request to the central arbiters asking for a scheduling slot on the output port of the output linecard. If there is an empty slot in the output port buffer, the active central arbiter will respond to the request with a grant. If there is no free buffer space, the arbiter will wait until there is a free buffer. In this manner, any congestion on an output port will result in distributed buffering across ingress ports and linecards. Congestion does not result in head-of-line blocking because queuing at ingress utilizes virtual output queues, and each QoS level for each output port has its own virtual output queue. This also helps ensure that any blocking does not spread throughout the system (see Figure 6).

Figure 6. Head-of-Line Blocking Mitigated Using Virtual Output Queuing
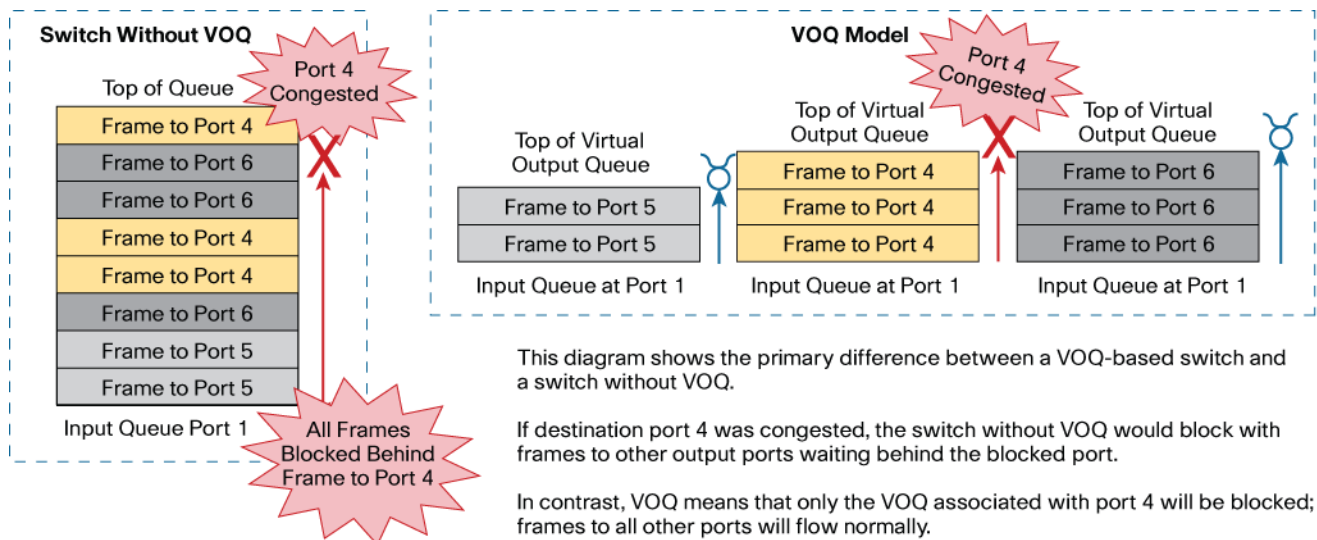
Because switchwide arbitration is operating in a centralized manner, fairness is also guaranteed; in the event of congestion on an output port, the arbiter will grant transmission to the output port in a fair (round-robin) manner.

Two redundant central arbiters are available on each Cisco MDS director switch, with the redundant arbiter snooping all requests issued at and grants offered by the primary active arbiter. If the primary arbiter fails, the redundant arbiter can resume operation without any loss of frames or performance or suffer any delays caused by switchover.
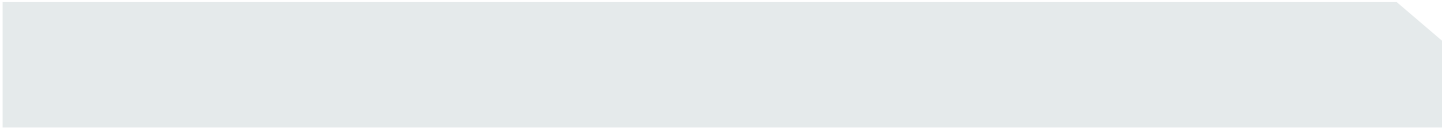
(4b) Crossbar Switch Fabric Module

Dual redundant crossbar switch fabrics are used on all director switches to provide low-latency, high-throughput, non-blocking, and non over-subscribed switching capacity between linecard modules.

In first-generation Cisco MDS line cards and Supervisor-1 modules, the crossbar switch fabric provides 80-Gbps full-duplex system switching capacity per slot. Second-generation Cisco MDS line cards, Supervisor-2 modules, and Cisco MDS 9513 switch fabric cards provide up to 96-Gbps full-duplex interface bandwidth per slot when operating in enhanced mode and 80 Gbps full-duplex when operating in standard mode. Supervisor-2 modules and the Cisco MDS 9513 switch fabric cards can operate in either enhanced or standard mode on a per-channel basis, depending on whether the line-card module is first-generation or second-generation. All second-generation line cards can also operate in either mode, always choosing the highest rate possible.

All current linecard modules utilize crossbar switch fabric capacity with 1:1 redundancy. That is, there are always one active channel per crossbar and one standby channel. One active channel and one standby channel are used on each of crossbar switch fabrics, resulting in both crossbars being active (1:1 redundancy). In the event of a crossbar failure, the standby channel on the remaining crossbar becomes active, resulting in an identical amount of active crossbar switching capacity regardless of whether the system is operating with one or two crossbar switch fabrics.

From a switch architecture perspective, crossbar capacity is provided to each linecard slot as a small number of high-bandwidth channels. This provides significant flexibility for linecard module design and architecture and makes it possible to build both performance-optimized (non-blocking, non over-subscribed) linecards and host-optimized (non-blocking, over-subscribed) linecards, multiprotocol (Small Computer System Interface over IP [iSCSI] and Fibre Channel over IP [FCIP]) and intelligent (storage services) linecards. These crossbar channels enabled Cisco to 'future-proof' the switch for investment protection. They also enable port-speed increases (2 Gbps to 4 Gbps and 10 Gbps) without any need to replace existing linecard modules and supervisor modules. A high-level overview of the primary architectural design decisions for utilizing centralized crossbar switch fabrics is shown in Figure 7.

The crossbar itself operates three times overspeed—that is, it operates internally at a rate three times faster than its endpoints. The overspeed helps ensure that the crossbar remains non-blocking even while sustaining peak traffic levels. Another feature to maximize crossbar performance under peak frame rate is a feature called superframing. Superframing improves the crossbar efficiency where there are multiple frames queued originating from one linecard with a common destination. Rather than having to arbitrate each frame independently, multiple frames can be transferred back to back across the crossbar.

All existing first-generation linecards are supported with second-generation crossbar switch fabric modules. Conversely, all second-generation linecards are supported in chassis with first-generation crossbar switch fabric modules. There is no requirement to upgrade a system to a second-generation crossbar switch fabric where the system is populated with first-generation linecard modules.

Crossbar switch fabric modules depend on clock modules for crossbar channel synchronization. In the case of a clock module failing, they are field replaceable. While failure of a clock module is disruptive to the system, the clock modules are little more than oscillators and currently have a real-world mean time between failure (MTBF) exceeding 300 years, so this is unlikely to be a problem. These are the same oscillators that have been used on Cisco Catalyst 6506 and 6509 switches since 1999.

Figure 7. Crossbar Versus 2-Tier Switch on Chip (SoC)
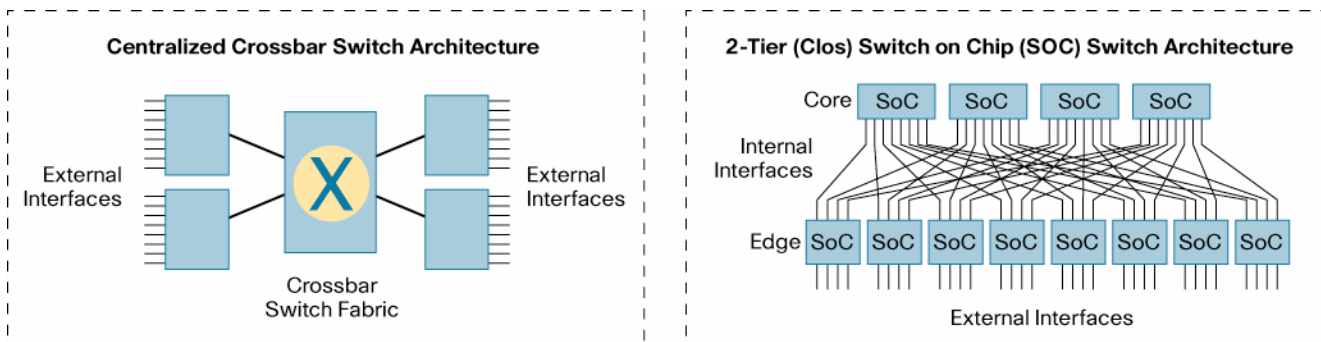
Director-class Fibre Channel switches are predominantly built using one of two architectures: two-tier switch-on-chip (SoC) and buffered crossbar.

It is both less costly and quicker to build a switch using two-tier SoC mesh than buffered crossbar—the idea here is that all resources can be put into building a single chip with a maximum number of ports on it, without the need to build multiple functional areas such as port, crossbar, and queuing functional groups.

The basic idea behind two-tier SoC is that a single chip can be deployed in both edge *ports* of a switch (with external-facing and internal-facing ports) as well as be deployed in the core of the switch (internal ports only).

This diagram shows the architecture of two theoretical 32-port switches. At first look, both offer an identical number of external ports (32) and identical levels of performance.

The primary differences start to become apparent when it comes to future port speeds and densities:

- A crossbar-based switch architecture with a large amount of internal (cross bar) connectivity can potentially support future speed increases transparently (for example, 2-Gbps to 4-Gbps to 10-Gbps) as well as potential increases in port density (for example, increasing line-card density from 32 ports per line card to 48 ports per line card).
- The increased speeds and density can typically be supported without upgrading the crossbar switch fabrics and obsoleting existing line cards.
- In contrast, a two-tier SoC switch architecture generally requires updating all of the components (edge line cards and core internal modules) in order to support higher speeds and densities.
- Adding support for disparate speeds (for example, 4-Gbps and 10-Gbps) where there is not an even multiplier between the two speeds and/or adding multiprotocol support further complicates this.
- This is not to say that it is not possible but rather that increasing speeds and densities typically results in a switch that is internally blocking and oversubscribed and/or requires a comprehensive upgrade.

(5) Egress Forwarding Module

The primary role of the egress forwarding module is to perform some additional frame checks, make sure that the Fibre Channel frame is valid, and enforce some additional ACL checks (logical unit number [LUN] zoning and read-only zoning if configured). On second-generation Cisco MDS linecards, there is also additional IVR frame processing (if the frame requires IVR) and outbound QoS queuing (see Figure 8).

Figure 8. Egress Frame Forwarding Logic
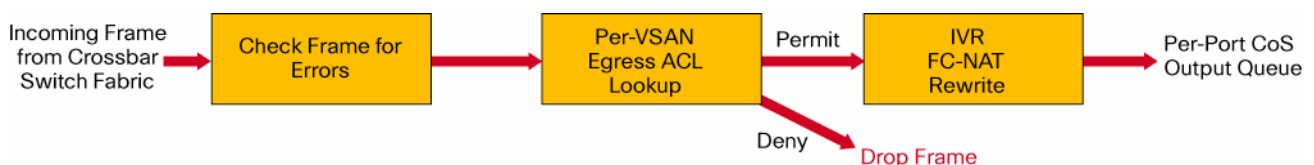
Before the frame arrives, the egress forwarding module has signaled the central arbiter that there is output buffer space available for receiving frames. When a frame arrives at the egress forwarding module from the crossbar switch fabric, the first processing step is to validate that the frame is error free and has a valid CRC.

If the frame is valid, the egress forwarding module will issue an ACL table lookup to see if any additional ACLs need to be applied in permitting/denying the frame to flow to its destination. ACLs applied on egress include LUN zoning and read-only zoning ACLs.

The next processing step is to finalize any Fibre Channel frame header rewrites associated with IVR.

Finally, the frame is queued for transmission to the destination port MAC with queuing on a per-CoS basis matching the DWRR queuing and configured QoS policy map.

(6) Egress PHY/MAC Modules

Frames arriving into the egress PHY/MAC module first have their switch internal header removed. If the output port is a Trunked E_Port (TE_Port), then an Enhanced ISL (EISL) header is prepended to the frame.

Next, the frame timestamp is checked, and if the frame has been queued within the switch for too long it is dropped. Dropping of expired frames is in accordance with Fibre Channel standards and sets an upper limit for how long a frame can be queued within a single switch. The default drop timer is 500 milliseconds and can be tuned using the "fcdroplatency" switch configuration setting. Note that 500 milliseconds is an extremely long time to buffer a frame and typically only occurs when long, slow WAN links subject to packet loss are combined with a number of Fibre Channel sources. Under normal operating conditions, it is very unlikely for any frames to be expired.

Finally, the frames are transmitted onto the cable of the output port. The outbound MAC is responsible for formatting the frame into the correct encoding over the cable, inserting SoF and EoF markers and inserting other Fibre Channel primitives over the cable.

IP SAN Extension—Fibre Channel over IP

Among the distinctive features of Cisco MDS 9000 switches are the multiprotocol capabilities enabled through integrated IP services and multiprotocol services modules. FCIP is one such capability and is used to extend Fibre Channel SAN connectivity anywhere IP network connectivity is available with the necessary bandwidth to transport SAN traffic.

FCIP is a means of providing a SAN extension for IP infrastructure, enabling storage applications such as asynchronous data replication, remote tape vaulting, and host initiator to remote pooled storage to be deployed irrespective of latency and distance. FCIP tunnels Fibre Channel frames over an IP link, using TCP to provide a reliable transport stream with a guarantee of in-order delivery.

In addition to standard FCIP, Cisco MDS 9000 Family switches offer a number of enhancements on top of standards-based FCIP to improve functionality and usability:

- **IVR**—Enables IP SAN extension without compromising fabric stability and reliability by merging fabrics
- **FCIP TCP traffic shaping**—Shaping storage traffic to meet bandwidth limits of the IP WAN
- **TCP enhancements**—Modifications to improve IP SAN extension performance
- **Compression (hardware based on MPS-14/2 linecards, software based on IPS-4/8 linecards)**—Conserves IP WAN bandwidth
- **FCIP Write Acceleration (FCIP-WA)**—Reduces the number of round trips associated with write operations, thereby improving latency
- **FCIP Tape Acceleration (FCIP-TA)**—Improves remote tape backup performance by minimizing the effect of latency and distance
- **Encryption and decryption**—Hardware-assisted IP Security (IPSec) AES128 crypto for transporting sensitive data over untrusted WAN links
- **Cisco PortChannels**—Bundling up to 16 FCIP tunnels as a single logical link, utilizing multiple physical Gigabit Ethernet interfaces and modules for higher performance and resiliency
- **Up to three FCIP tunnels per Gigabit Ethernet interface**—Enables multisite IP SAN extension

The flow of frames and order of processing for FCIP with these enhancements is shown in Figure 9.

Figure 9. FCIP Processing Logic
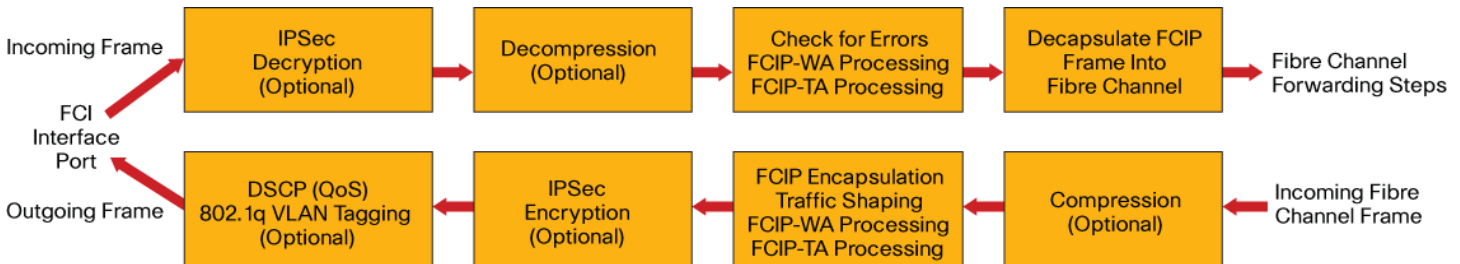
Compression and traffic shaping always occur before encryption. Traffic shaping always occurs after compression. This enables deployment of IP SAN extension, encryption, compression, and traffic shaping in a single system.

After all FCIP processing has completed for received frames, frames are processed using the standard Fibre Channel frame forwarding logic.

All Cisco MDS 9000 Family IP modules (IPS-4, IPS-8, MPS-14/2) are capable of sustaining high throughput on Gigabit Ethernet ports, regardless of distance or latency and whether features such as FCIP Write Acceleration, FCIP Tape Acceleration, or IVR are enabled. In most cases, when following best-practice deployment, 100 percent line-rate traffic can be sustained even where connectivity is over tens of thousands of miles. Note that Cisco MDS 9000 Family IP services Gigabit Ethernet ports provide no Ethernet switching capabilities—they provide only connectivity as FCIP tunnel endpoints and iSCSI gateway connectivity.

Cisco PortChannels

PortChannels are an ISL aggregation feature that can be used to construct a single logical ISL between switches from up to 16 physical ISLs. This is useful in terms of both providing high-throughput connection between switches and providing a highly resilient connection. Traffic across a PortChannel bundle is automatically load-balanced according to the per-VSAN load-balancing policy, providing for very granular traffic distribution across all physical interfaces within a PortChannel bundle. The Fibre Channel in-order delivery requirement is always preserved.

With PortChannels, physical interfaces can be added to and removed from a PortChannel bundle without interrupting the flow of traffic. This enables nondisruptive configuration changes to be made in production environments.

There are no restrictions on the location of ports or modules used for building PortChannels. Because the Cisco MDS 9000 switch architecture uses a consistent forwarding path for frames from ingress to egress ports, any port on any module can be used for PortChannels without concern for negative effects caused by different ports or modules offering different latency characteristics.

Lossless Networkwide In-Order Delivery Guarantee

The Cisco MDS switch architecture guarantees that frames can never be reordered within a switch. This guarantee extends across an entire multiswitch fabric, provided the fabric is stable and no topology changes are occurring.

Unfortunately, when a topology change occurs (e.g. a link failure occurs), frames might be delivered out of order. Reordering of frames can occur when routing changes are instantiated, shifting traffic from what might have been a congested link to one that is no longer congested. In this case, newer frames transmitted down a new noncongested path might pass older frames queued in the previous congested path.

In order to protect against this, all Fibre Channel switch vendors have a fabricwide configuration setting called "networkwide in-order delivery" (network-IOD) guarantee. This stops the forwarding of new frames when there has been an ISL failure or a topology change. No new forwarding decisions are made during the time it takes for traffic to flush from existing interface queues. The general idea here is that it is better to stop all traffic for a period of time than to allow the possibility for some traffic to arrive out of order. This mechanism is a little crude because it guarantees that an ISL failure or topology change will be a disruptive event to the entire fabric. Because of this, all SAN switch vendors default to disabling network-IOD except where its use is mandated by channel protocols such as IBM Fiber Connection (FICON).

As of Cisco MDS 9000 SAN-OS Software Release 3.0, Cisco MDS 9000 Family switches offer a new feature called "enhanced networkwide in-order delivery." The enhancement uses VOQ, and knowledge of what specific routes are affected by a topology change, to stop only that traffic headed along paths affected by the topology change. This enhancement is accomplished using unique hardware support built into the VOQ, where route changes can be instigated in a nondisruptive manner.

A similar technique is used with Cisco PortChannel bundles. New links can be added to a PortChannel bundle nondisruptively and with a guarantee that frames will not be reordered. Likewise, interfaces can be nondisruptively removed from PortChannel bundles while guaranteeing in-order delivery and no dropped frames.
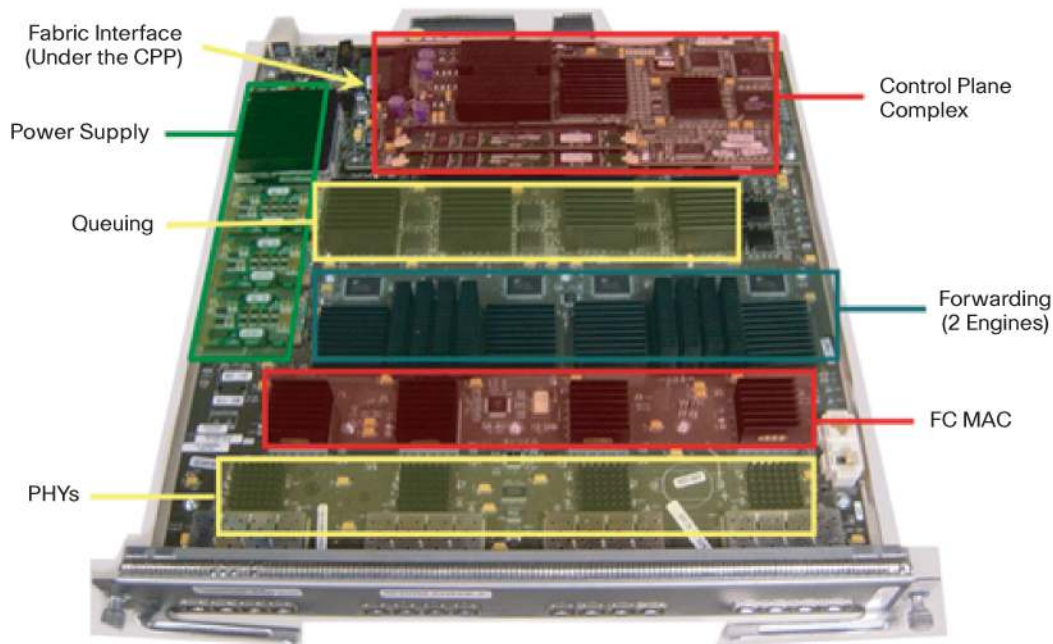
A LOOK AT THE HARDWARE

First-Generation Fibre Channel Linecards

16-Port Non Over-Subscribed, Non-Blocking Linecard

Figure 10 shows a 16-port non over-subscribed, non-blocking linecard.

Figure 10. 16-Port 1/2-Gbps Non Over-Subscribed, Non-Blocking Linecard

The first-generation Cisco performance linecard (part number DS-X9016) provides 16 non over-subscribed, non-blocking front-panel 1/2-Gbps Fibre Channel ports. All ports are capable of sustaining line-rate traffic simultaneously.

Each front-panel Fibre Channel port has its own 2-Gbps channel to one of two forwarding complexes. The linecard has two forwarding complexes, each servicing 8 front-panel ports.

Queuing provides 400 ingress buffers for each port. These are split: 255 buffers are credited (linked to buffer credits on front-panel ports), and 145 buffers are uncredited performance buffers used to absorb upstream congestion. Connectivity to the crossbar switch fabric is through all four crossbar channels, with only two of these active at any given time.

Control-plane functions on the linecard are performed locally on a linecard local control-plane processor.

32-Port Host-Optimized Over-Subscribed, Non-Blocking Linecard

Figure 11 shows a 32-port host-optimized over-subscribed, non-blocking linecard.

Figure 11. 32-Port 1/2-Gbps Host-Optimized Over-Subscribed, Non-Blocking Linecard
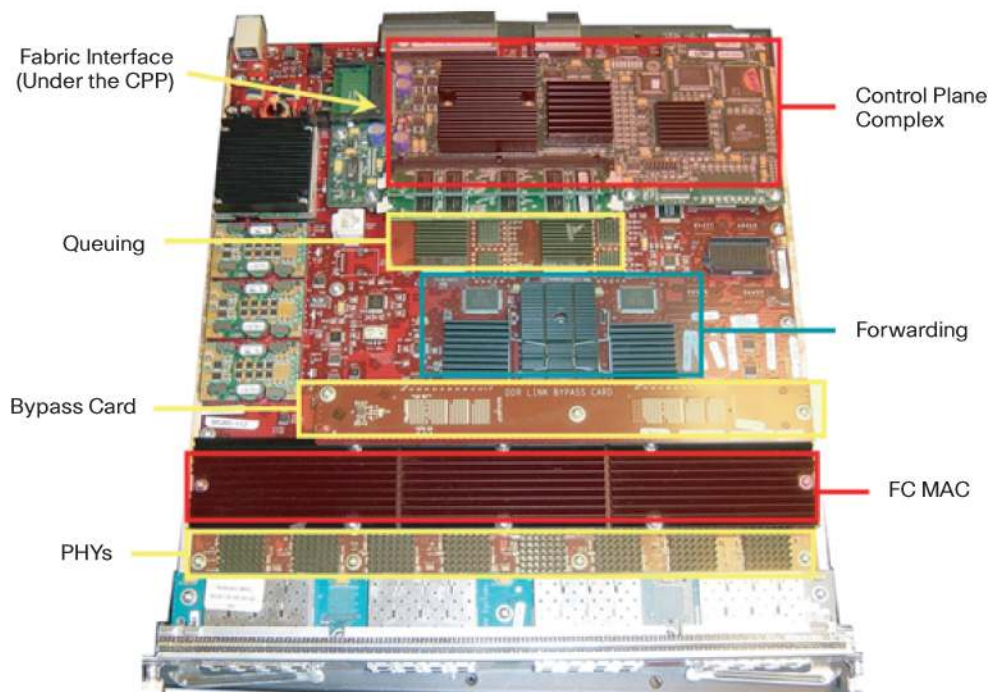
The first-generation Cisco host-optimized linecard (part number DS-X9032) provides 32 over-subscribed, non-blocking front-panel 1/2-Gbps Fibre Channel ports.

Front-panel ports are divided into 4-port groups. Each group shares 2 Gbps throughput between the MAC and forwarding modules. 1/2-Gbps Fibre Channel is encoded at 8b/10 (one start and stop bit for every 8 bits of data), so maximum usable throughput per port is 1.7 Gbps. After removing the inter-frame gap (IFG), 4 2-Gbps ports sharing 2 Gbps usable full-duplex capacity to the forwarding module equates to an oversubscription ratio of 3.25:1. Individual front-panel ports are capable of sustaining line-rate traffic, provided other ports in the same group are not attempting to do so at the same time. Because this linecard is targeted for host connectivity, the oversubscription on this card is not normally a limiting factor, because an overall SAN design typically has an oversubscription ratio ranging from 7:1 to 12:1 between host ports and storage ports.

Queuing provides 12 ingress buffers for each port. Connectivity to the crossbar switch fabric is through all four crossbar channels, with two of these active at any given time.

Control-plane functions on the linecard are performed locally on a linecard local control-plane processor.

Figure 11 shows a card marked as a "bypass card." Note that the 32-port module was designed to support intelligent storage services and is the card on which the Cisco MDS 9000 32-Port Storage Services Module (SSM) linecard is based.

Cisco MDS 9000 32-Port Storage Services Module

Figure 12 shows the Cisco MDS 9000 32-Port Storage Services Module.

Figure 12. Cisco MDS 9000 32-Port 1/2-Gbps Storage Services Module
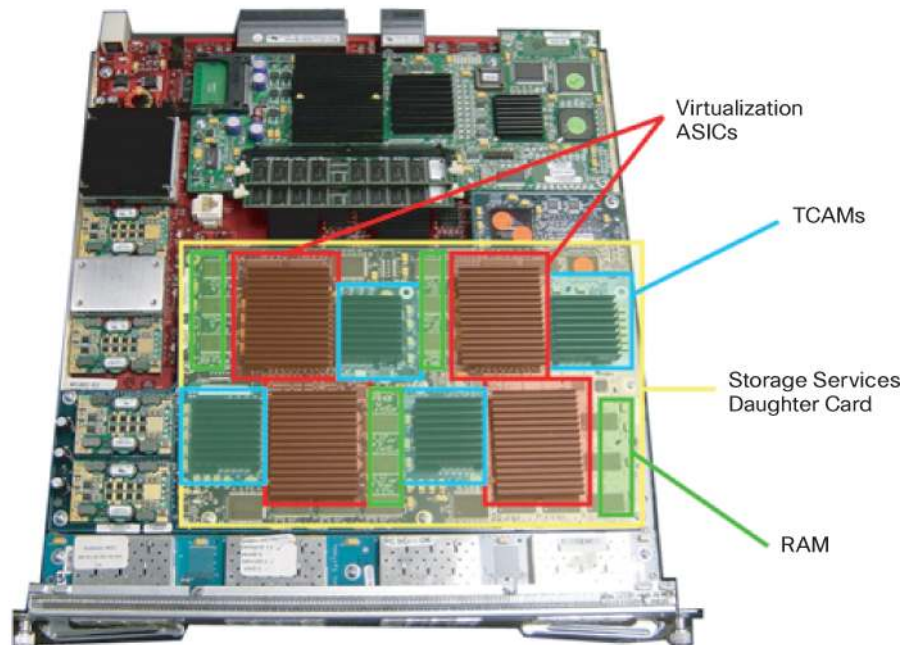
The SSM linecard is based on the same design as the standard 32-port linecard, but also features a storage services daughter card where the bypass card exists on the standard module. Storage services supported include storage virtualization, volume management, reliable replication writes (Cisco SAN Tap), Fibre Channel Write Acceleration, and Network-Accelerated Serverless Backup (NASB).

Front-panel ports are configured in the same manner as the host-optimized ports on the standard 32-port linecard.

The virtualization daughter card is directly connected to the forwarding complex with 20-Gbps full-duplex usable bandwidth. Each of the four virtualization ASICs on the virtualization daughter card contains two cores. Multiple cores and virtualization ASICs can be used in parallel, with the same virtual LUN instantiated on multiple virtualization ASICs, even across multiple SSMs.

Second-Generation Fibre Channel Linecards

Although there is a significant reduction in the number of ASICs on a linecard, second-generation linecards use the same frame forwarding processing logic.

All second-generation Cisco MDS linecard modules are based on the same board design, with variations in the number and type of front-panel ports and the number of MAC/PHY complexes on a linecard. Frames entering the switch arrive at a combined MAC/PHY module, which can service 3, 6, or 12 1/2/4-Gbps front-panel Fibre Channel ports or one 10-Gbps Fibre Channel port, depending on the linecard.

Figure 13 shows 12/24/48-port 1/2/4-Gbps Fibre Channel linecards and a 4-port 10-Gbps Fibre Channel linecard.

Figure 13. 12/24/48-Port 1/2/4-Gbps Fibre Channel Linecards 4-port, Four 10-Gbps Fibre Channel Linecard
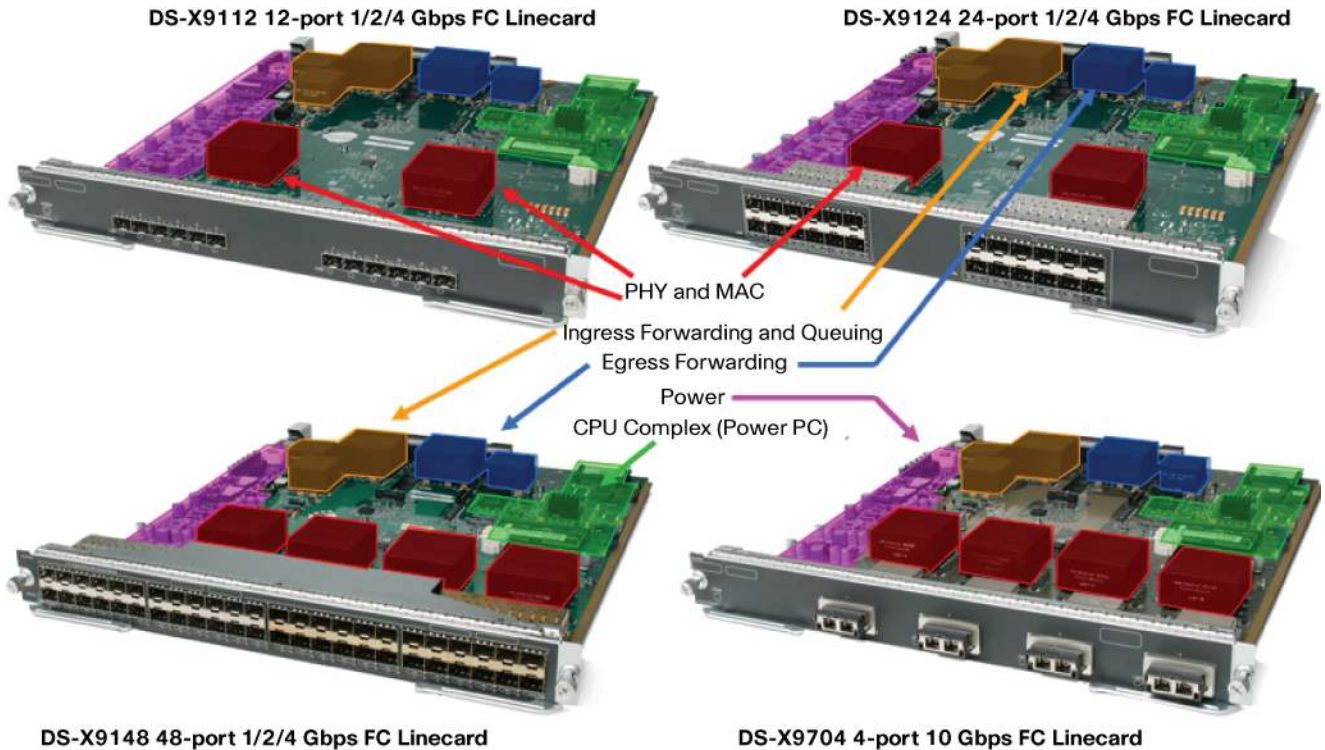
Variations in oversubscription for different linecards at different speeds are caused by the number of MAC/PHY complexes and the number of front-panel ports each MAC/PHY is servicing. Front-panel ports are divided into port groups. A rectangle on the front bezel of each linecard shows these port groups, as shown in Table 1.

Table 1. Second-generation Cisco MDS 9000 Family Linecard Modules

Linecard	Front-Panel Ports per Port Group	Oversubscription if All Ports Operating at			
		1 Gbps	2 Gbps	4 Gbps	10 Gbps
DS-X9112 12-Port 1/2/4-Gbps Linecard	3	None	None	None	–
DS-X9124 24-Port 1/2/4-Gbps Linecard	6	None	None	under 2:1	–
DS-X9148 48-Port 1/2/4-Gbps Linecard	12	None	under 2:1	under 4:1	–
DS-X9704 4-Port 10-Gbps Linecard	1	–	–	–	None

Table 1 shows worst-case oversubscription, when all ports in the same port group are contending for port group bandwidth. More realistically, the bandwidth requirements per port will vary and fluctuate over time, and it is unlikely that worst-case oversubscription will actually occur.

Although Table 1 shows that a 12-port linecard is not over-subscribed if all 12 ports are pushing line rate 4 Gbps, the 12-port card is functionally identical to a 24-port linecard where 3 ports on each 6-port group are operating in full-rate mode. The 48-port linecard is also functionally identical to the 12-port linecard if 3 ports in each 12-port group are operating in full-rate mode. This is made possible by configuring individual ports to operate in full-rate mode, with bandwidth dedicated to those ports. The alternative is for ports to operate in shared mode, where bandwidth is shared between multiple ports. When a port is operating in full-rate mode, it has bandwidth guaranteed to it, at the expense of reducing available bandwidth for other ports in the same port group.

This flexibility makes it possible to deploy a chassis full of 48-port linecards. Ports used for storage, ISLs, and high-performance hosts can be configured to operate in full-rate mode, with all remaining ports configured to operate in shared mode.

All second-generation linecards use a single forwarding/queuing complex per linecard. This supports forwarding rates of up to 116 million frames per second. All announced second-generation linecards utilize crossbar switch fabric channels with 1:1 redundancy—that is, two active channels and two redundant channels. System switching performance is identical whether the system is operating with one crossbar switch fabric or two crossbar switch fabrics installed.

Control-plane functions on the linecard are performed locally on a linecard local control-plane processor.
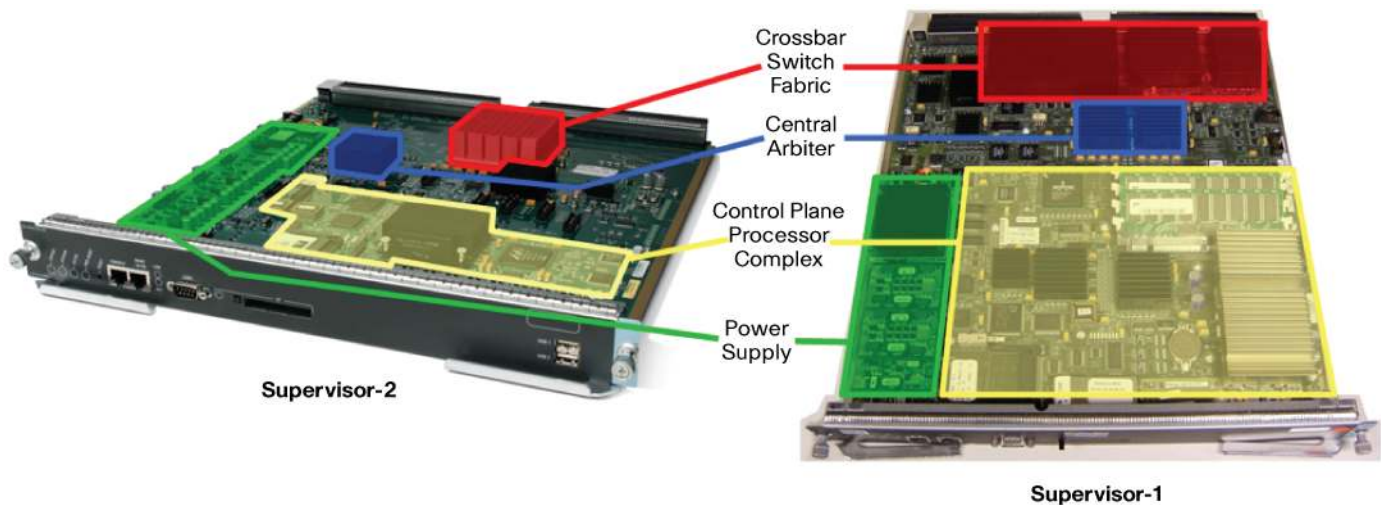
Supervisor and Crossbar Switch Fabric Modules

Cisco MDS 9000 Family supervisor modules provide two essential switch functions:

- They house the control-plane processors that are used to manage the switch and keep the switch operational. Management connectivity is provided through an out-of-band Ethernet (interface mgmt0), an RJ45 serial console port, and optionally also a DB9 auxiliary console port suitable for modem connectivity.
- They house the crossbar switch fabrics and central arbiters used to provide data-plane connectivity between all the linecards in the chassis.

Figure 14 shows Supervisor-1 and Supervisor-2 modules.

Figure 14. Supervisor Modules

Besides providing crossbar switch fabric capacity, the supervisor modules do not handle any frame forwarding or frame processing. All frame forwarding and processing is handled within the distributed forwarding ASICs on the linecards themselves. Although each Supervisor-2 module contains a crossbar switch fabric, this is only used on Cisco MDS 9509 and MDS 9506 chassis. On the Cisco MDS 9513 chassis, the crossbar switch fabrics are on fabric cards (Figure 15) inserted into the rear of the chassis, and the crossbar switch fabrics housed on the supervisors are disabled.

Figure 15. Cisco MDS 9513 Crossbar Switch Fabric Module

Control-plane functionality on the Supervisor-1 is handled by a Pentium-3 processor operating at 1.3 GHz with 512 MB RAM. 256 MB internal Flash memory is used as a boot device. A Compact Flash slot is available for additional image and log storage.

Control-plane functionality on the Supervisor-2 is handled by a G4 PowerPC operating at 1.4 GHz with 1 GB RAM. 512 MB internal Flash memory is used as a boot device. Both a Compact Flash slot and dual USB ports are available for additional image and log storage.

SUMMARY OF LINECARDS

Table 2 summarizes the various linecard options available and the forwarding characteristics of each module.

Table 2. Linecard Summary

Linecard			Number of Ports		Fibre Channel		Targeted Function
Cisco Part Number	Description	Gen	FC	GE	Sustained Performance if all Ports Pushing 100% Line Rate	Buffer Credits per Port	
DS-X9016	16-port 1/2-Gbps Fibre Channel module		16	–	Line-rate 1/2-Gbps Fibre Channel	255	High-performance host, ISL, storage
DS-X9032	32-port 1/2-Gbps Fibre Channel module		32	–	• 2-Gbps 3.25:1 over-subscribed	12	Optimized for host connectivity
DS-X9032-SSM	32-port 1/2-Gbps Fibre Channel storage services module		32	–	• 1-Gbps 1.62:1 over-subscribed per quad		Host connectivity with storage services
DS-X9302-14K9	2-port 1-Gigabit Ethernet IPS, 14-port 1/2-Gbps Fibre Channel module	1st	14	2	Line-rate 1/2-Gbps Fibre Channel	Up to 3200	Very long-distance Fibre Channel SAN ISL extension, IP SAN extension, iSCSI
DS-X9304-SMIP	4-port IP storage services module		–	4	–	–	IP SAN extension, iSCSI
DS-X9308-SMIP	8-port IP storage services module		–	8	–	–	IP SAN extension, iSCSI

Linecard			Number of Ports		Fibre Channel		Targeted Function
Cisco Part Number	Description	Gen	FC	GE	Sustained Performance if all Ports Pushing 100% Line Rate	Buffer Credits per Port	
DS-X9112	12-port 1/2/4-Gbps Fibre Channel module		12	–	Line-rate 1/2/4-Gbps Fibre Channel	Up to 4095	High-performance 1/2/4-Gbps host/storage/ISL including very long-distance Fibre Channel SAN ISLs
DS-X9124	24-port 1/2/4-Gbps Fibre Channel module		24	–	• Line-rate 1/2-Gbps Fibre Channel • 4-Gbps 2:1 over-subscribed	Up to 4095	
DS-X9148	48-port 1/2/4-Gbps Fibre Channel module	2nd	48	–	• Line-rate 1-Gbps Fibre Channel • 2-Gbps 2:1 over-subscribed • 4-Gbps 4:1 over-subscribed	Up to 4095	
DS-X9704	4-port 10-Gbps Fibre Channel module		4	–	Line-rate 10-Gbps Fibre Channel	Up to 4095	High-performance 10G ISL (up to 660 km)

COMMON QUESTIONS AND ANSWERS

Q. If a supervisor or crossbar switch fabric is removed from the chassis, does this impact switch forwarding performance?

A. No. There is sufficient crossbar capacity in a single crossbar switch fabric to maintain the same level of switch performance whether one or two crossbar switch fabrics are present in the system.

Q. Is it disruptive to the system to remove a supervisor, crossbar switch fabric, or linecard?

A. No. Supervisors, fabric cards, and linecards can be hot-swapped with no disruption to the switch or frame processing. In the case of physically removing supervisor or crossbar switch fabric modules, the system will ensure that there is no traffic flowing through the crossbar well before the card has lost physical contact with the backplane connector.

Q. There is only one fan tray for linecards. What happens if a fan fails?

A. Although there is only a single side-mounted fan tray for linecards, there are multiple redundant fans on the fan tray itself. These fans are powered through two independent power rails, and all fans are fitted with RPM sensors. Failure of any fan causes all other fans to speed up to compensate. In the unlikely event of multiple simultaneous fan failures, the fan tray still provides sufficient cooling to keep the switch in operation. If a fan fails, the entire fan tray is replaced. Replacement of a fan tray is nondisruptive provided the fan tray is replaced within the allotted time.

Q. Do any frames ever need to be sent to the supervisor modules for centralized forwarding?

A. No. All forwarding is always performed in hardware on the distributed forwarding ASICs on the linecards themselves. All advanced forwarding features are implemented within the standard forwarding pipeline without any impact to performance or latency. While the crossbar switch fabric resides on the Supervisor modules within Cisco MDS 9506 and MDS 9509 chassis, this is purely for linecard to linecard connectivity, and not for centralized forwarding.

Q. Why did Cisco choose to build a director switch using centralized crossbars and not build a switch using a two-tier Clos mesh with SoC technology? Would that not have enabled Cisco to release a switch sooner?

A. When Cisco first released the Cisco MDS 9000 Family of high-density multiprotocol intelligent storage switches in December 2002, Cisco set the benchmark for intelligent features and functionality in port densities never seen in the Fibre Channel marketplace. Cisco could have built a director using a two-tier Clos architecture using SoC technology, but this would have negatively impacted the intelligent features and functionality as well as compromised options for scalability and investment protection. From a switch architecture standpoint, Cisco believes SoC is suitable for relatively small port-count fixed-configuration Fibre Channel fabric switches, but is unsuitable for future modular Fibre Channel director switches.

Q. Why does the Cisco MDS always forward all frames across the crossbar even if the destination port is on the same linecard?

A. This was a conscious design decision, enabling predictable performance, latency, and jitter while enabling features such as frame fairness, QoS, PortChannels, and in-order frame delivery guarantee. There is no reason not to always forward frames through the crossbar, as the crossbar it not a bottleneck limiting overall system performance.

Q. What is the MTBF on Cisco director switches?

A. All the individual hardware components (chassis, supervisors, linecards, crossbar switch fabrics, power supplies, and fan trays) and Cisco MDS 9000 SAN-OS Software are designed to provide in excess of 99.999 percent uptime—or less than 5 minutes of downtime per year. The system as a whole is designed to exceed this availability number. Cisco has been shipping Cisco MDS director switches since December 2002. In that time, real-world MTBF has significantly exceeded 99.999 percent uptime.

Q. Some competitors claim that the Cisco backplane contains active components. Can you clarify?

A. The backplane on Cisco MDS 9506 and MDS 9509 chassis house dual redundant clock modules used for crossbar synchronization. Although it is technically accurate that the clock modules are active components, they consist of little more than an oscillator and have a real-world demonstrated MTBF in excess of 300 years. These are the same oscillators that have been used on Cisco Catalyst 6506 and 6509 switches since 1999.

- On the Cisco MDS 9513 chassis, the clock modules are no longer on the backplane but instead have been made hot-swappable and field-replaceable.
- There are other components on the backplane (capacitors, resistors) that are used for signal integrity and grounding during online insertion and removal of linecards, however all of these components are passive and are not essential for system operation.

CONCLUSION

Cisco MDS 9500 Series multilayer directors elevate the standard for director-class switches. Providing industry-leading availability, scalability, security, and management, the Cisco MDS 9500 Series enable deployment of high-performance SAN switching infrastructure with the lowest total cost of ownership. Layering a rich set of intelligent features onto a high-performance, protocol-independent switch fabric, Cisco MDS 9500 Series multilayer directors address the stringent requirements of large data center storage environments.

CISCO SYSTEMS

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

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

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

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

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

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