White Paper

Cisco MDS 9000 Family Diagnostics, Error Recovery, Troubleshooting, and Serviceability Features

July 2016

Author

Fausto Vaninetti
Consultant System Engineer, EBG Data Center

Key Contributors

Ed Mazurek
Technical Leader, Services

Paresh Gupta
Technical Marketing Engineer, Enterprise Switching

Mike Blair
Technical Leader, Engineering

Ram Natarajan
Manager, Software Development Engineering

Contents

What You Will Learn

This document describes the Cisco® MDS 9000 Family of 16-Gbps Fibre Channel–capable storage networking products from a very specific and often ignored perspective. The features and capabilities that help and support IT administrators in their daily jobs are explained and their benefits discussed. Day-0 activities are related to the initial preparation and setup for a new solution, and day-1 activities indicate the normal configuration steps to make the solution fully operational and fit for its purpose. This document is about day-2 activities that contribute to making the solution fit for use: in particular, monitoring, diagnostics, and troubleshooting activities, including examples and use cases. The internal architecture of Cisco MDS 9000 Family of switches is also described and its benefits explained. The buffer-to-buffer credits mechanism and the features in support to reliable frame delivery are also covered.

Enterprise-Class Storage Networks

With the introduction of Fibre Channel (FC) protocol and SANs two decades ago, the requirement to have servers and disks directly connected disappeared, providing a whole new level of flexibility and higher storage resource utilization. However, the upper-layer SCSI protocol that controls the server-to-disk communication has not changed. An I/O process in Fibre Channel represents a SCSI task, and it is still identified by an I_T_L_Q nexus, in which one **initiator** port talks to one **target** port, addressing one **LUN** to run one task (identified by **Q**). For the transport layer, SCSI expectations also have not changed. SCSI is assumed to be extremely solid and reliable. As a result, IT architects began building storage area networks with a clear understanding of the importance of protecting this critical infrastructure from any possible misbehavior. Dual fabrics, In-service Software Upgrades (ISSU) of device firmware images, highly resilient switches, and logical bundling of physical links have all been key design criteria since the first deployments of storage area networks. High-availability, no single point of failure, and reliable operation have become mantras for storage architects and achievements they are proud of.

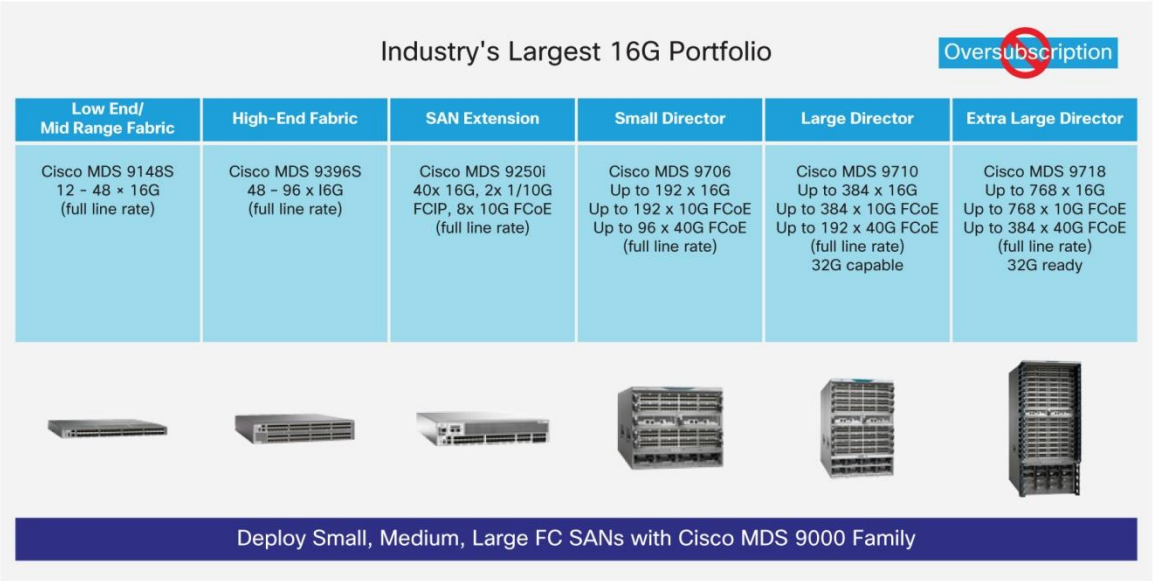
Despite the emergence of other protocols to build storage networks and the more recent trend toward bringing storage closer to computing, as represented by computational storage and hyperconverged solutions, Fibre Channel is still predominant for enterprise-class storage solutions. This is partly due to the field-proven reliability of those implementations and years of experience in designing and operating them. Over time, new capabilities have been added to networking devices to better provide operational ease and reliable transport of data within large, complex, and even multisite architectures.

The Cisco Multilayer Data Switch (MDS) 9000 Family and the Cisco Nexus® storage network operating system (Cisco NX-OS Software) have been the first to provide a broad set of diagnostic, error recovery, troubleshooting, and serviceability features that simplify the process of building, expanding, and maintaining a storage area network. These features combined help increase availability of SANs by essentially eliminating disruptions during maintenance windows and reducing the time needed to recover from possible problems.

Cisco entered the Fibre Channel networking business at the end of 2002 and since then enjoyed a large commercial success due to technical innovation, top notch performance, hardware reliability combined with software stability, multiprotocol support and investment protection. Starting from 2013, the entire Cisco MDS 9000 Family product line has been refreshed and become 16-Gbps Fibre Channel–capable. The new devices have been designed with the same original success factors in mind. For example, the modular directors can host 16G Fibre Channel linecards but will be able to accommodate 32G Fibre Channel linecards without requiring a fork-lift upgrade.

In addition, they can already support 10G and 40G Fibre Channel over Ethernet (FCoE) linecards and will soon allow insertion of 10G/40G Fibre Channel over IP (FCIP) linecards as well. As of today, the Cisco MDS 9000 Family (Figure 1) of 16-Gbps Fibre Channel–capable storage networking products comprises 3 switches and 3 directors, with an offering starting from 12 ports and up to 768 ports in a single device. In all cases, there is no oversubscription nor bandwidth sharing inside the products. Every port can run full speed with no compromise.

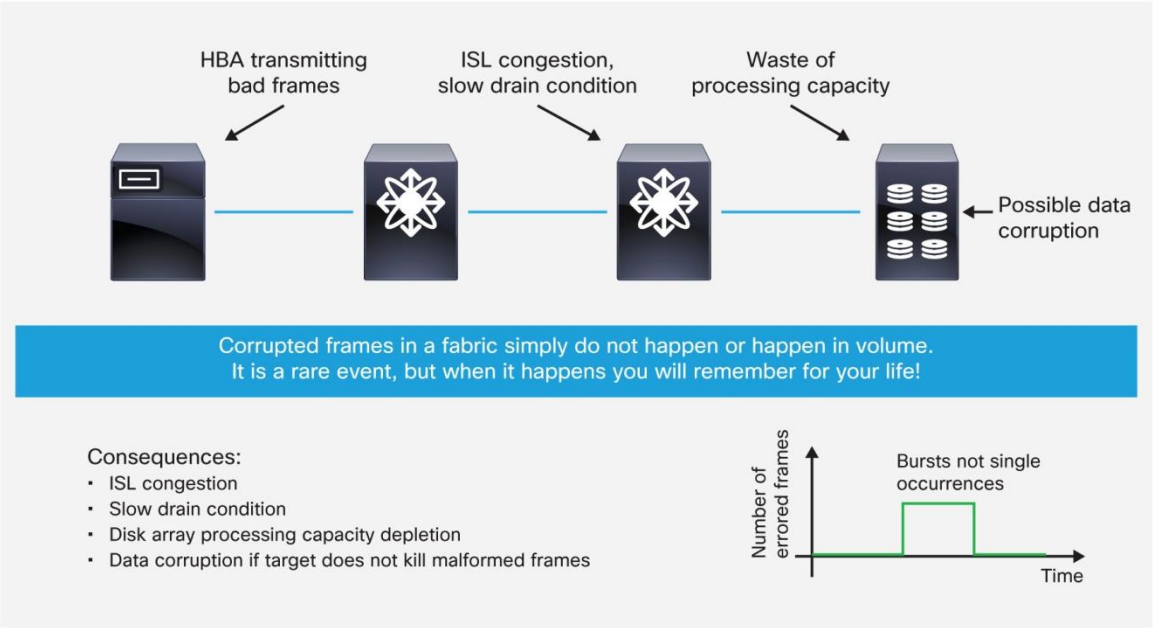
Figure 1. Cisco MDS 9000 Family of 16-Gbps Fibre Channel–Capable Devices



Store and Forward Architecture and Dropping of Corrupted Frames

MDS 9000 Family storage switches have always excelled in delivering FC frames in the most reliable way thanks to their store and forward architecture and the resulting unique capability to identify corrupted frames and drop them immediately. This prevents other devices further down the path from generating error indications and confusing the location of the actual problem. Many years of field experience have demonstrated that corrupted frames are a reality within storage networks and typically happen in bulk, often a consequence of a misbehaving or about-to-fail HBA device. MDS 9000 Family switches provide the intrinsic capability to identify those frames as soon as they enter the fabric and immediately drop them to avoid possible negative outcomes (Figure 2). This capability is tied to the advanced ASIC design and overall internal architecture that was introduced in 2002 and maintained across subsequent generations of line cards and switches. In a nutshell, MDS 9000 Family switches inspect incoming frames, recalculate their parity, and compare the value with what is stored in the cyclic redundancy check (CRC) bytes at the end of Fibre Channel frames. In the event of a mismatch, the frame is declared corrupted and consequently dropped, and appropriate counters are incremented to support the administrator in identifying trends and to perform root-cause analysis. On MDS 9000 Family switches, the port attached to the source of the corrupted frames can be automatically shut down when a specified error threshold is exceeded. This can be achieved by combining MDS 9000 Family port monitor and port guard features, to prevent any impact on applications.

Figure 2. The Risk from Corrupted Frames



Buffer-to-Buffer Credit Loss Detection and Recovery

One notable feature in the category of diagnostics and error recovery technologies is buffer-to-buffer credit (BB credit) loss detection and recovery. Buffer credits form the basis of how Fibre Channel networks work and operate. Buffer-to-buffer credits are essentially a flow-control mechanism to help ensure that Fibre Channel switches do not drop frames. They provide a mechanism whereby frames get delivered to destinations without the risk of facing congestion and spending excessive time within device buffers. In simple terms, the receiving device acknowledges the reception of a frame by sending a receiver-ready (R_RDY) message back to the transmitter and that will eventually increment by 1 the buffer-to-buffer credit counter. The amount of BB credits required on a port is strongly tied to the distance the frames are expected to travel before the next hop is reached. This is the reason that MDS 9000 Family switches are equipped with a generous quantity of them. For example, the 48 x 16-Gbps Fibre Channel line card has approximately 50,000 BB credits available for its ports: three times the number in competing solutions. The quantity of usable BB credits is set on a per-hop basis, and so Fibre Channel devices always communicate the buffers that are free on their side to the peer device. Many issues can arise in the SAN in the event of buffer credit starvation or buffer credit loss.

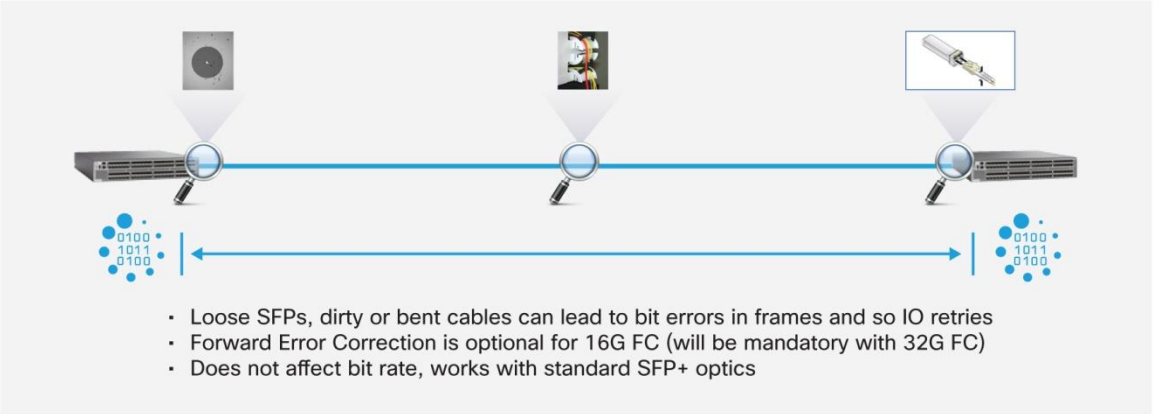
Although the Fibre Channel standards require low bit error rates, bit errors do occur from time to time. The corruption of receiver-ready (R_RDY) primitives would deteriorate communication between peer devices and would lead to a loss of credits, which could eventually cause a link to stop transmitting in one direction. The Fibre Channel standards provide a feature for two attached ports to detect and correct this situation. This feature is called buffer-to-buffer credit loss detection and recovery. It functions as follows: the sender and the receiver agree to send checkpoint primitives to each other, starting from the time that the link comes up. The sender sends a checkpoint every time it has sent the specified number of frames, and the receiver sends a checkpoint every time it has sent the specified number of R_RDY primitives. If the receiver detects lost credits, it can retransmit them and restore the credit count on the sender.

MDS 9000 Family devices fully support the buffer-to-buffer credit loss detection and recovery mechanism and can restore the buffer credits without interrupting the data flow on the Inter-Switch Links (ISLs). This feature is enabled by default on all ISLs.

Forward Error Correction

More recently, on the refreshed MDS 9000 Family portfolio with 16-Gbps Fibre Channel speeds, the robustness of Fibre Channel networks was further strengthened by introducing a new feature called forward error correction (FEC). The intent is not limited to identifying corrupted frames, but includes correcting them in real time. Lossy media such as loose transceivers (SFPs) or dirty cables may result in corrupted packets on ISLs. FEC helps reduce or avoid data stream errors that result in dropped frames and that can lead to application performance degradation (Figure 3). This feature can be optionally enabled on MDS 9000 Family switches, and its adoption is on the rise, specifically in IBM System z environments where it can be supported end to end, from host to target, and not just across ISLs. For these FICON environments, transmitter training is required in combination with FEC. As expected, the FEC capability can be enabled and monitored through the command-line interface (CLI) or the Cisco Prime™ Data Center Network Manager (DCNM) graphical user interface. When enabled, FEC allows recovery of up to 11 error bits in every 2112-bit transmission, thus enhancing the reliability of data transmissions.

Figure 3. Forward Error Correction



The use of FEC enables additional statistics on the relevant ports. When everything is operating normally, the FEC corrected errors counter should stay at zero, but it would show some value in the event that any bit was corrected. This way, even if links are working properly, the SAN administrator can easily see when FEC is correcting bits, and that is an indication that some action is needed at the next available maintenance window. In other words, FEC on MDS 9000 Family switches allows the administrator to identify possible issues on the link even if applications are not yet negatively affected. This is an example of preventive maintenance.

Because this FEC capability is fully implemented in hardware and uses an in-band approach, there is no performance impact on throughput and no bit-rate change. All MDS 9000 Family switches can detect and drop corrupted frames at the switch input, but FEC adds another layer of resiliency on 16-Gbps-capable devices to help correct errors wherever feasible and reduce the number of packet drops. A similar capability will be implemented on future 32-Gbps Fibre Channel line cards. When enabled, this feature contributes to more resilient end-to-end frame delivery.

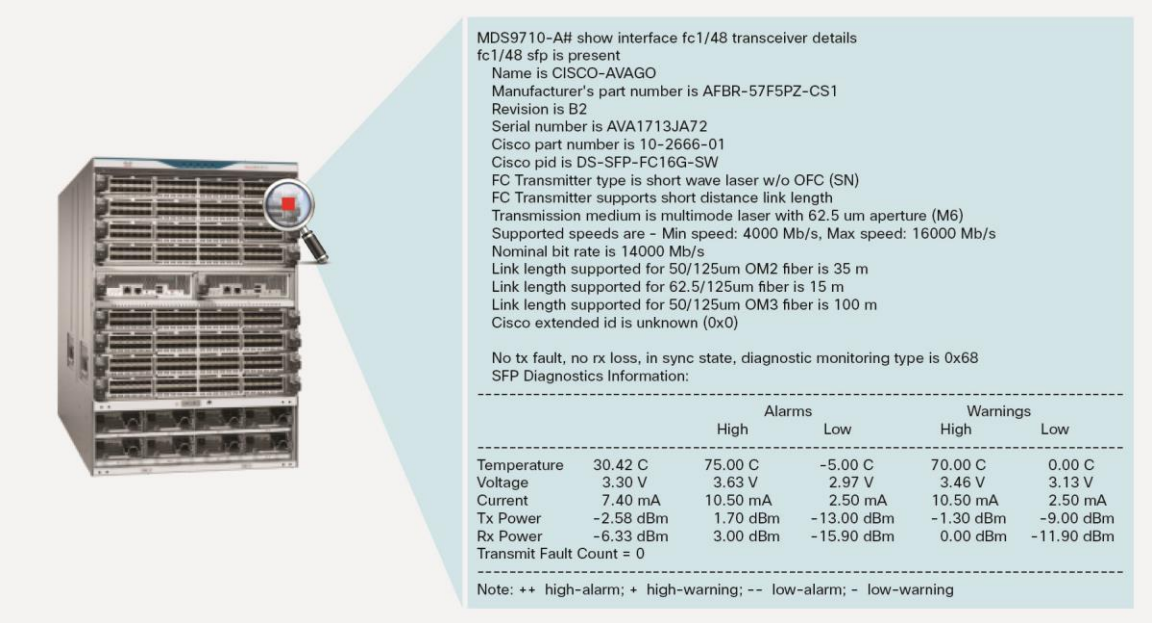
Digital Optical Monitoring

The MDS 9000 Family makes use of top-quality pluggable transceivers to help ensure proper operation over the longest possible period of time. They are designed with performance and reliability as top criteria and go through a deep testing activity so that defects are contained to fewer than 10 parts per million and the mean-time-between-failure (MTBF) value is well in excess of dozens of years. In addition, those optical transceivers support the digital optical monitoring (DOM) capability, whereby the transmitted and received power levels can be displayed. This way, defective SFPs can be quickly identified, and dirty or loosely plugged cables remediated. With DOM, the user can perform in-service transceiver monitoring and troubleshooting operations (Figure 4). The MDS 9000 device will issue syslog messages and generate SNMP traps when warning and alarm thresholds are crossed. The currently monitored values as well as the warning and alarm thresholds are clearly seen in the output of the appropriate CLI command.

Despite the fact that internal sensors are calibrated and the accuracy of measured parameters is good, users should not rely on DOM to estimate fiber cable loss.

One interesting parameter that can be seen through DOM is the operating temperature of the SFP itself. Although the maximum operating temperature is 70°C, it is safer to have the SFP operating at a much lower temperature. In fact, just as for any other electronic device, the higher the operating temperature, the lower the reliability over time. Considering the number of SFPs inside a director, this clearly is a primary consideration for the overall reliability of a storage fabric. That is why Cisco MDS 9700 Series Multilayer Directors have implemented a port-side intake approach for air cooling. This way, all SFPs will be reached by cool air and not by hot air going out of the director. An example of the CLI output for a 16-Gbps Fibre Channel transceiver is shown in Figure 4.

Figure 4. Digital Optical Monitoring on a 16-Gbps Fibre Channel SFP+ Transceiver



```
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 C      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
```

Cisco Generic Online Diagnostics

The Cisco Generic Online Diagnostics (GOLD) capability is also included in the latest iteration of the MDS 9000 Family. Periodically, user-defined tests are run to verify that supervisor engines, switching modules, ASICs, communication buses, optics, and interconnections are functioning properly. Some of these advanced online diagnostic capabilities do not adversely affect normal Fibre Channel operations, allowing them to be run in production SAN environments. Other tests require that no traffic be present on interested components and can be used to perform health checks before new links are brought in production. Support for the GOLD framework started with NX-OS 6.2 on MDS 9700 Series directors as an evolution to the Cisco Online Health Management System (OHMS) diagnostic framework that was implemented starting from the previous generation of MDS 9000 Family platforms. For brevity, OHMS will not be described in detail in this document. Suffice it to say that OHMS is the predecessor of GOLD, supports similar but not identical capabilities, has different CLI commands to be configured, and is still supported on 16-Gbps top-of-rack fabric switches in the MDS 9000 Family.

GOLD is a suite of diagnostic facilities for verifying that hardware and internal data paths are operating as designed. Different module types support different GOLD tests. To see the GOLD tests that are available for a specific module type, use the following CLI command:

```
F241-15-09-9710-2# show diagnostic content module 1
Diagnostics test suite attributes:
B/C/* - Bypass bootup level test / Complete bootup level test / NA
P/*   - Per port test / NA
M/S/* - Only applicable to active / standby unit / NA
D/N/* - Disruptive test / Non-disruptive test / NA
H/O/* - Always enabled monitoring test / Conditionally enabled test / NA
F/*   - Fixed monitoring interval test / NA
X/*   - Not a health monitoring test / NA
E/*   - Sup to line card test / NA
L/*   - Exclusively run this test / NA
T/*   - Not an ondemand test / NA
A/I/* - Monitoring is active / Monitoring is inactive / NA

Module 1: 2/4/8/10/16 Gbps Advanced FC Module
                                                        Testing Interval
ID     Name                                  Attributes      (hh:mm:ss)
____   ____________________________________  ___________   ______________
 1)    ASICRegisterCheck------------->      ***N******A     00:01:00
 2)    PrimaryBootROM---------------->      ***N******A     00:30:00
 3)    SecondaryBootROM-------------->      ***N******A     00:30:00
 4)    EOBCPortLoopback-------------->      C**N**X**T*     -NA-
 5)    OBFL-------------------------->      C**N**X**T*     -NA-
 6)    PortLoopback------------------>      *P*D**XE***     -NA-
 7)    SnakeLoopback----------------->      *P*N***E**A     00:20:00
 8)    IntPortLoopback--------------->      *P*N***E**A     00:05:00
 9)    RewriteEngineLoopback--------->      *P*N***E**A     00:01:00
10)    ExtPortLoopback--------------->      *P*D**XE***     -NA-
11)    BootupPortLoopback------------>      CP*N**XE*T*     -NA-
```

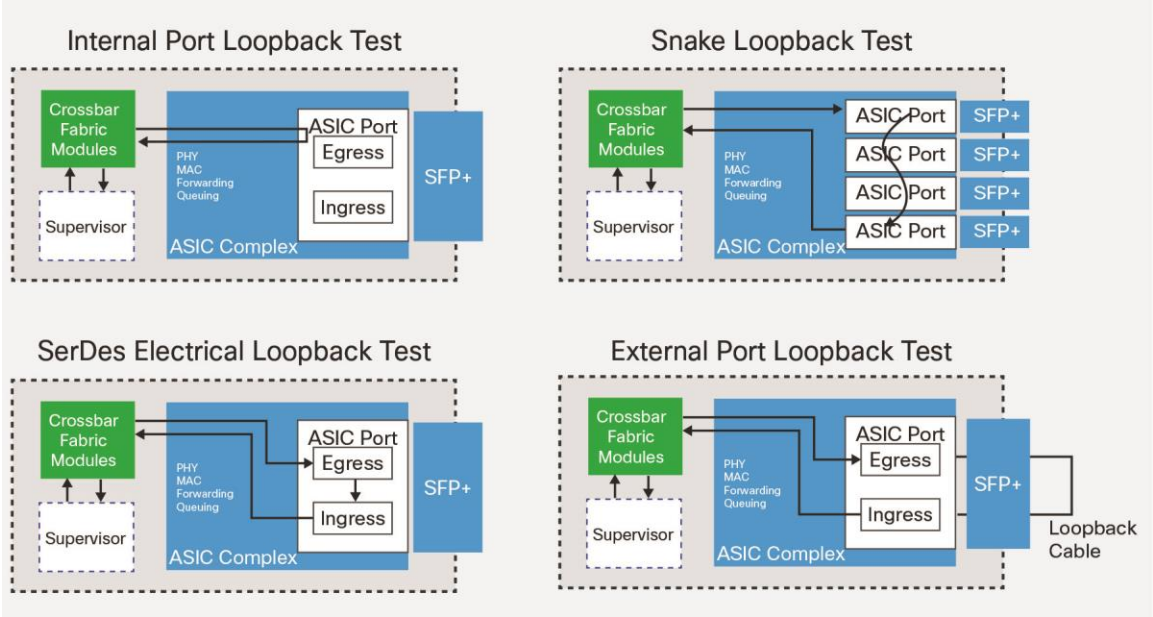
GOLD tests can be run in three modes:

- Bootup mode: Bootup diagnostics run during bootup and detect faulty hardware before an MDS 9700 Series switch brings a module online.
- Health-monitoring mode (also called runtime mode): Health-monitoring mode diagnostics are enabled by default to verify the health of a live system at periodic intervals.
- On-demand mode: All the health-monitoring tests can also be invoked on demand. On-demand diagnostics run only when invoked by the user. On-demand tests can be run in two ways:
  - **diagnostic start**: This command initiates the test and does not block the CLI.
    To see the results, a **show diagnostic result module slot test [test-id | name]** command must be issued.
  - **diagnostic run**: This command initiates the test and blocks the current CLI session until the completion of the test. After the completion of the test, the CLI session is unblocked, and the result is displayed.

GOLD is an industry-leading diagnostics subsystem that allows rapid fault isolation and continuous system monitoring: critical features in today's continuously operating environments. It can be used in conjunction with Cisco Embedded Event Manager (EEM), a policy framework for defining the actions to be taken when a configurable event or condition occurs.

As part of GOLD capabilities, the MDS 9000 Family also provides multiple loopback testing options to check port capabilities (Figure 5).

Figure 5. Cisco GOLD Tests



Some tests can be scheduled, and some tests are intended to be used before ports are put in production or on demand in specific situations. The test frame is generated by the switch supervisor. You can run manual loopback tests to identify hardware errors in the data path in the switching modules and in the control path in the supervisor modules. A variety of tests can be considered, as listed here:

- Internal port loopback (IntPortLookback) test: This is a nondisruptive test, and as such it can be scheduled to run periodically (default is every 5 minutes) or triggered on demand. However, it is limited in scope because it can test only the internal path from the supervisor to the Fibre Channel ASIC, as represented here: Sup<->central xbar<->local xbar<->Fibre Channel ASIC (internally facing side only). As a result, this test could be successful, but the port could actually be nonoperational due to problems lower in the ASIC. The IntPortLoopback test is supported beginning from MDS NX-OS 6.2(7) and can be run from the CLI and the device manager as well. Use the EXEC-level **diagnostic run module** command to explicitly run this test on demand (when requested by the user) on the specified port and immediately display results when the test is completed:

```
MDS9710# diagnostic run module 4 test IntPortLoopback port 8
```

- Snake loopback (SnakeLoopback) test: This test runs on all ports of a line card. It forms an internal (no external links) snake-like topology of front-panel ports and verifies connectivity from the supervisor to all the ports in the line card. It runs on all the ports regardless of their state, and it is a nondisruptive test with a default interval of 20 minutes. Use the EXEC-level **diagnostic run module** command to explicitly run this test on demand (when requested by the user) on the specified module and immediately display results when the test is completed:

```
MDS9710# diagnostic run module 3 test SnakeLoopback
```

- Port loopback (PortLoopback) test: The serializer and deserializer (SerDes) port loopback test (also called the electrical port loopback test) can be run on demand on all the ports regardless of the port state, and it checks the data path up to the PHY component. This test is disruptive to live traffic because it brings down the port for the purpose of diagnostic operations. It error disables the port on failure. In other words, the SerDes port loopback test is a hardware loopback inside the SerDes unit, just before the SFP transceiver is reached, and verifies the proper functioning of the electrical data path. The SerDes port loopback test checks the hardware for a port. This test is available for Fibre Channel interfaces only. Use the EXEC-level **diagnostic run module** command to explicitly run this test on demand (when requested by the user) on the specified port and immediately display results when the test is completed:

```
MDS9710# diagnostic run module 2 test PortLoopback port 5
```

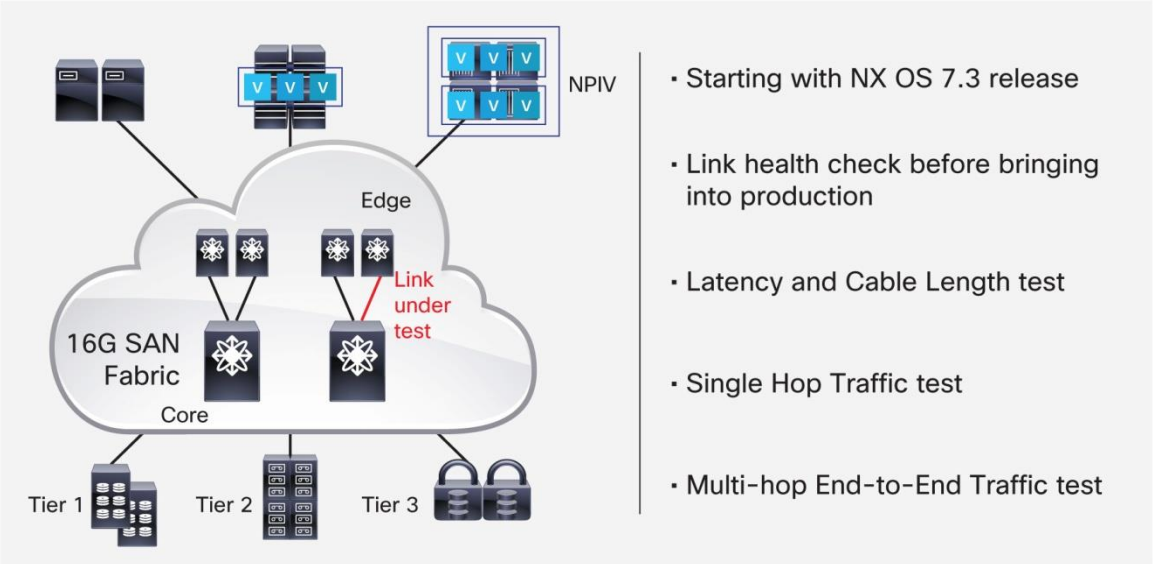
- External port loopback (ExtPortLoopback) test: This is a disruptive test that was added to GOLD capabilities with NX-OS 6.2(11c). It is also supported by the device manager from NX-OS 7.1(1). The beauty of this test is that it can verify the entire electrical and optical data path on the switch, including the SFP itself. It requires a loopback cable to be installed. The external loopback test sends and receives an FC2 frame to and from the same port. You need to connect a cable (or a plug) to loop the Rx port to the Tx port before running the test. This test is available only for Fibre Channel interfaces. The external loopback test capability is not dependent on the device that will eventually be connected to the port under test. This makes this feature compatible with HBAs from all vendors when enabled on F-ports. Use the EXEC-level **diagnostic run module** command to explicitly run this test on demand (when requested by the user) on the specified port and immediately display results when the test is completed:

```
MDS9710# diagnostic run module 3 test ExtPortLoopback port 2
```

ISL Diagnostics

Starting with Cisco MDS 9000 NX-OS Software Release 7.3, an ISL diagnostics capability is available to help check the health and performance of ISLs before activating the links for production traffic. The ISL diagnostics feature helps validate the health of ISLs between MDS 9000 Family switches in a network. These tests are expected to measure traffic loss rate, link latency, and cable length, among other parameters (Figure 6).

Figure 6. Cisco MDS 9000 Family ISL Diagnostics

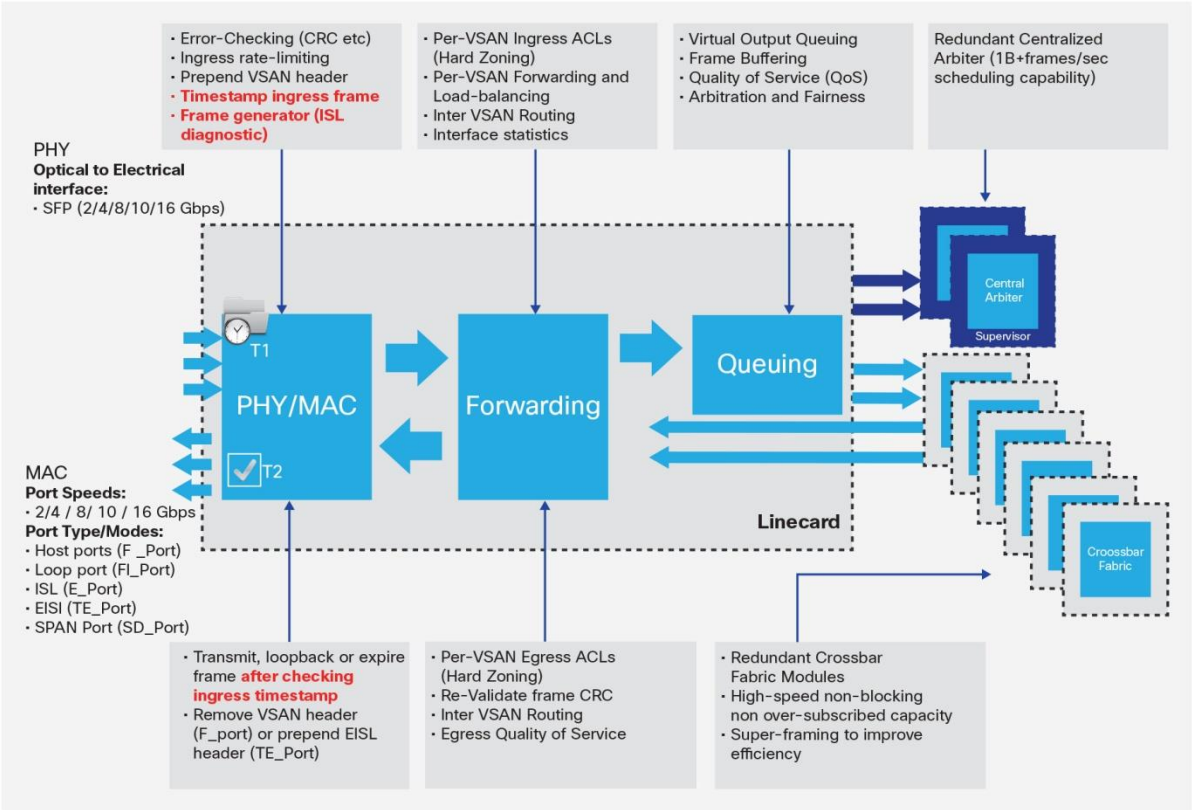


In doing so, they also verify the integrity of cables, SFPs, and electronic components along the frame path. The use case for this new ISL diagnostics feature is clear. It offers the opportunity to measure distances, diagnose optics and cables, and verify multihop data paths before ISLs are put into production, making sure the links are well characterized and fully operational. Because it is a capability of ASICs inside MDS 9000 Family line cards, it works similarly on shortwave, longwave, extended longwave, and WDM optics. The accuracy of the link measurement is within +/– 2 meters.

The ISL diagnostics feature is available on all of MDS 9700 Series directors (as well as on Cisco MDS 9396S 16G Multilayer Fabric Switches in a future release), and it has been made available also on the previous generation of Cisco MDS 9500 Series Multilayer Directors.

At the foundation of this new feature are three individual technologies: frame generators, port reflectors, and time stamps (Figure 7).

Figure 7. Logical Packet Walk Inside Cisco MDS 9700 Series Mission-Critical Directors



The following tests can be performed using ISL diagnostics:

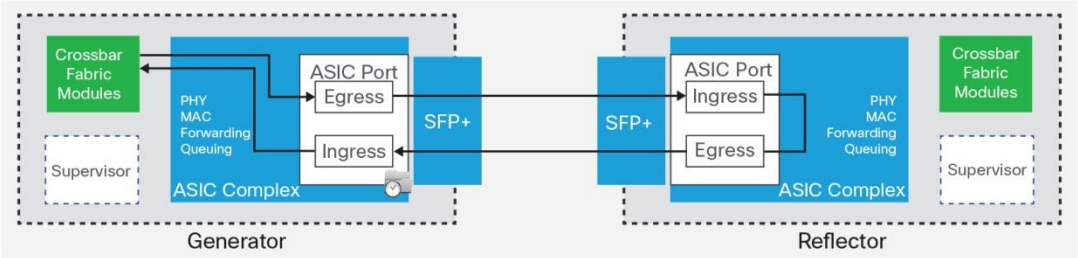
- Latency and cable length test
- Single-hop traffic test
- Multihop end-to-end traffic test

Because these tests are intended for validating a link before it is inserted into production, no traffic should be running on the links to be tested. These tests have the clear benefit of speeding up fabric deployment and reducing diagnostic time. These tests help ensure predictable application performance over Fibre Channel links. In particular, the latency and cable length test provides granular distance and latency measurements for accurate buffer credit assignment. Moreover, the single-hop and multihop traffic tests can simulate application-level I/O profiles.

Latency and Cable Length Test

The latency test measures the latency and length of an ISL between two MDS 9000 Family switches. It is a functional test, not a stress test. The process works as follows: the supervisors sends a start-test instruction to the MAC chipset of the relevant port. A frame compliant with the IEEE 1588 standard is then generated at the generator switch and time-stamped. The frame is looped back (at the SerDes level) by the reflector switch port to the generator switch, where another time stamp is captured (Figure 8). Time stamps allow the roundtrip latency of the link to be measured, as well as latency inside the switch itself. The cable length is calculated by dividing by 2 the roundtrip link latency, after subtracting the time taken by the frame to leave the switch from the time of packet generation. This way, a good degree of accuracy can be achieved.

Figure 8. Latency and Cable Length Test



This example shows how to run the latency and cable length test on an MDS 9700 Series director and display the results appearing on the CLI screen:

```
MDS9718# diagnostic isl latency-test interface 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
```

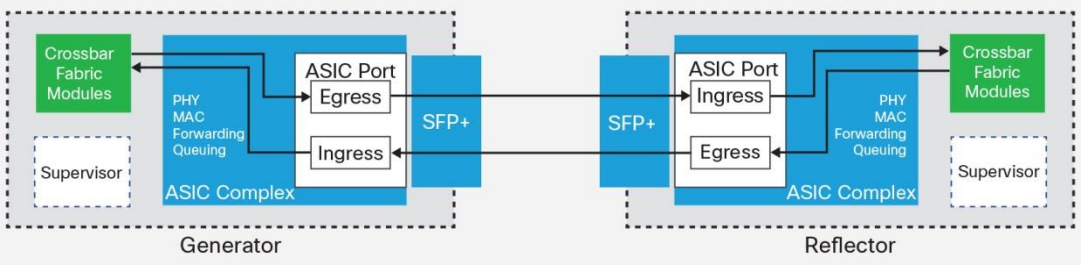
The cable length information is not required when putting multiple physical links into a single logical bundle called a port channel. In fact, MDS 9000 port channels do not suffer from link-length mismatch and do not impose a limit on member links' deskew value because the load-balancing mechanism is based on SID, DID, and OXID, and per-frame load-balancing is avoided. However, the cable length information can be of help to properly allocate BB credits and estimate the I/O performance of applications using the SAN.

Single-Hop Traffic Test

The single-hop traffic test validates the health of an ISL by checking the efficiency of the link to handle traffic at various frame rates. Fibre Channel frames are generated in the generator switch using the internal traffic generator facility available in the media access control (MAC) hardware inside the port-facing ASIC complex. These frames are transmitted from the generator switch port over the ISL under test. The reflector switch receives the frames, switches them through the normal fabric switching path, and transmits them back through the receiving port onto the ISL under test. Frames are dropped by MAC hardware when they reach the generator switch (Figure 9).

The efficiency of the ISL traffic is calculated based on the number of packets coming back to the generator switch port, which should be 100 percent under normal conditions. The traffic can be generated at different speeds and variable frame sizes during a specific time period. This test is sometime referred to as a link-saturation stress test. The generated frames are kept within the default VSAN 1.

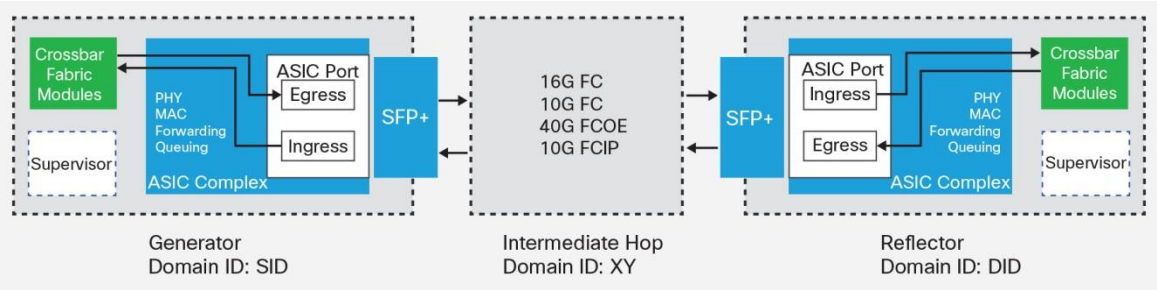
Figure 9. Single-Hop Traffic Test



Multihop End-to-End Traffic Test

The multihop test extends the health validation to the entire path between an initiator switch and a target switch in a fabric. Before connecting a host and a target in a fabric, it is now possible to test the fabric path between the host port and the target port using a multihop test. There can be multiple hops between the host switch and target switch. There is no specific configuration required on the intermediate switches; the test just works transparently for them. Fibre Channel frames are generated at the generator switch port and transmitted to the first-hop link. These frames traverse intermediate switches in agreement with standard FSPF routing algorithms until they reach the reflector switch. The reflector switch then switches the frames and returns them to the generator switch. Based on the number of packets received back on the generator switch, the efficiency of the ISL is displayed and should be 100 percent under normal conditions. The multihop traffic test is based on the domain IDs of the generator and reflector switches. For best flexibility, the generated frames are kept within the VSAN specified when enabling this feature. That VSAN is expected to be allowed all the way down from the host switch to the target switch, and the recommended approach is to keep it as a dedicated VSAN for this test activity. To know the domain IDs of interest, the **show topology** command can be of help. This test is sometime referred to as a link-saturation stress test since frames are hardware generated and can be at line rate. The intermediate ISLs can use any possible transmission technology including 10-Gbps Fibre Channel, 40-Gbps FCoE, and 10-Gbps FCIP (Figure 10). Specifically for testing FCIP links, the Cisco SAN Extension Tuner can be used as well.

Figure 10. Multihop Traffic Test

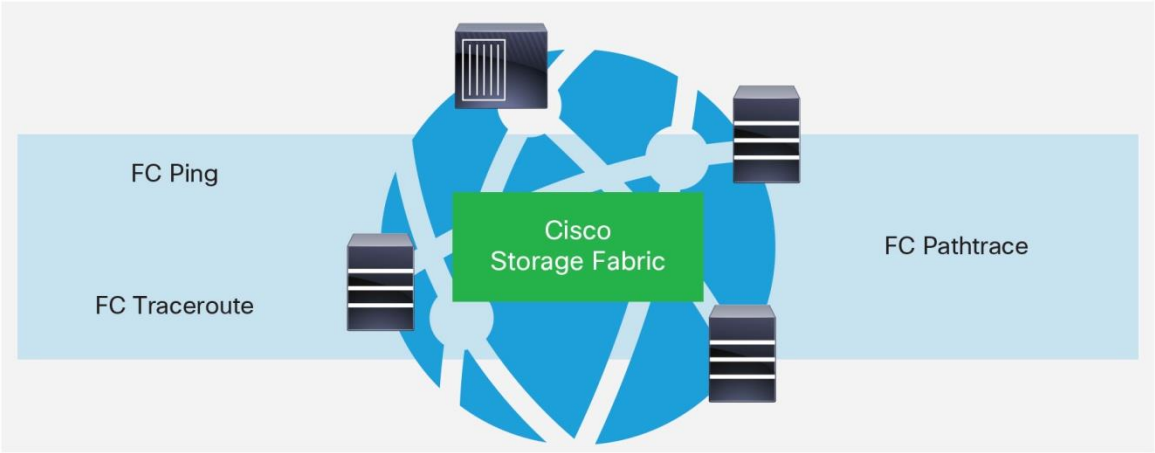


Of course, the ISL diagnostics feature within the SAN fabric can always be complemented by host-based applications to test end-to-end connectivity between a virtual machine and its LUN. This host-level test would also verify multipathing operation.

Fibre Channel Ping, Fibre Channel Traceroute, and Fibre Channel Pathtrace

The MDS 9000 Family was the first to bring Fibre Channel ping, Fibre Channel traceroute, and Fibre Channel pathtrace to storage networks (Figure 11). With Fibre Channel ping, administrators can check the connectivity of an N-port and determine its round-trip latency. With Fibre Channel traceroute, administrators can check the reachability of a switch by tracing the path followed by frames. Starting with MDS 9000 NX-OS 6.2(5), the new Fibre Channel pathtrace feature builds on the Fibre Channel traceroute capability to provide more statistics about each hop in the path, such as ingress and egress ports, number of transmitted and received frames, and errors. These are just a few examples in the ample portfolio of troubleshooting tools that MDS 9000 Family products offer.

Figure 11. Connectivity Tests Available on Cisco MDS 9000 Family Storage Fabrics



The Fibre Channel ping tool checks end-to-end connectivity by using the destination port WWN or FCID or device alias. Fibre Channel ping allows you to ping a Fibre Channel N-port or end device, sending it a series of frames. After these frames reach the target device's N-port, they are returned to the source, and a time stamp is taken. Fibre Channel ping helps verify the connectivity and latency to an N-port.

This is how the Fibre Channel ping feature could be invoked using the CLI by providing the FCID and VSAN of the destination port:

```
sw107-9250i# fcping ?
  device-alias  Device-alias of the destination N-Port
  fcid          FC-id of the destination N-Port
  pwwn          PWWN of the destination N-Port

sw107-9250i# fcping fcid 0xc40100 vsan 1
28 bytes from  0xc40100  time = 776 usec
28 bytes from  0xc40100  time = 744 usec
28 bytes from  0xc40100  time = 736 usec
28 bytes from  0xc40100  time = 746 usec
```

```
28 bytes from  0xc40100  time = 746 usec

5 frames sent, 5 frames received, 0 timeouts
Round-trip min/avg/max = 736/749/776 usec
```

The Fibre Channel traceroute feature allows you to trace the route followed by data traffic. Fibre Channel traceroute identifies the path taken on a hop-by-hop basis and reports back the switch WWN and domain ID. You can use Fibre Channel traceroute to test the connectivity along the path between the generating switch and the switch closest to the destination. The frames are routed normally as long as they are forwarded through TE ports. After the frame reaches the edge of the fabric (the F-port connected to the destination node with the given port WWN or FCID or device alias), the frame is looped back (swapping the source ID and the destination ID) to the originator. The Fibre Channel traceroute feature works only on TE ports, so make sure that no E-ports exist in the path to the destination. An example of the use for this feature is provided here:

```
MDS9148s-1# fctrace device-alias ag104_1 vsan 237
Route present for : 10:00:00:00:00:09:00:01
20:00:00:0d:ec:24:f5:00(0xfffc9f)
20:00:00:2a:6a:b9:d1:90(0xfffc02)
20:00:54:7f:ee:ea:6f:00(0xfffc33)
```

The Fibre Channel pathtrace is a utility tool that traces the path from the switch on which the CLI is run to a destination domain or destination device referenced by its FCID. This feature works with Fibre Channel, FCoE, and FCIP ports. Information is injected into the Fibre Channel pathtrace feature by the FSPF routing protocol process and port manager process running on MDS 9000 Family devices. As a result, Fibre Channel pathtrace collects information about the available paths inside the fabric and outputs the shortest route with additional relevant details. Each line in the Fibre Channel pathtrace output displays the source interface, destination interface, cost, speed, and other statistics (when used with the **detail** option). The **reverse** option of the CLI can be used to display the reverse path information (from the destination back to the source). This output format was carefully chosen to help storage network administrators in their daily jobs. Fibre Channel pathtrace provides an end-to-end view of the shortest path without the need to connect to individual switches and see the FSPF topology hop by hop. If the destination cannot be reached, Fibre Channel pathtrace will show the device in which connectivity stops.

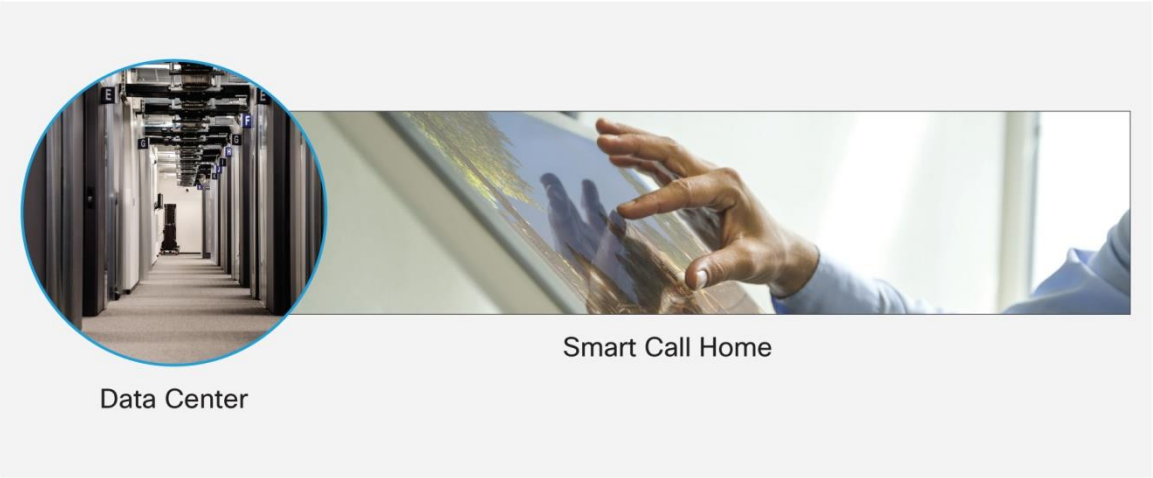
```
Switch1# pathtrace fcid 0x7e0300 vsan 3
The final destination port type is F_Port
---------------------------------------------------------------------------
Hop Domain In-Port          Out-Port         Speed Cost  Switchname
---------------------------------------------------------------------------
0   78     embedded         fc1/5            8G    125   switch1
1   126    fc1/5            fc1/8            4G    -     switch2
NOTE: The stats are displayed for the egress interface only
```

```
Switch1# pathtrace fcid 0x7e0300 vsan 3 detail
The final destination port type is F_Port
--------------------------------------------------------------------------------
Hop 0       Domain In-Port          Out-Port         Speed Cost  Switchname
            78     embedded         fc1/5            8G    125   switch1
--------------------------------------------------------------------------------
Stats for egress port: fc1/5
    TxRt(B/s): 608
    RxRt(B/s): 2368
        TxB_B: 64
        RxB_B: 64
      TxFrame: 332611
      RxFrame: 332641
       Errors: 0
      Discard: 0
          CRC: 0
--------------------------------------------------------------------------------
Hop 1       Domain In-Port          Out-Port         Speed Cost  Switchname
            126    fc1/5            fc1/8            4G    -     switch2
--------------------------------------------------------------------------------
Stats for egress port: fc1/8
    TxRt(B/s): 224
    RxRt(B/s): 320
        TxB_B: 32
        RxB_B: 64
      TxFrame: 1717
      RxFrame: 1721
       Errors: 0
      Discard: 0
          CRC: 0
NOTE: The stats are displayed for the egress interface only
```

Cisco Smart Call Home

The MDS 9000 Family offers the Cisco Smart Call Home feature for proactive fault management (Figure 12). Call Home provides a notification system triggered by software and hardware events. It forwards alarms and events, packaged with other relevant information in a standard format, to external entities. Alert grouping capabilities and customizable destination profiles offer the flexibility needed to notify specific individuals or support organizations only when necessary. These notification messages can be used to automatically open technical-assistance tickets and resolve problems before they become critical. 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, and the Cisco Technical Assistance Center (TAC).

Figure 12. Cisco Smart Call Home Feature



Smart Call Home is a component of Cisco SMARTnet Service contract that offers proactive diagnostics, real-time alerts, and personalized web-based reports on Cisco MDS 9000 switches (and other Cisco devices). Smart Call Home provides fast resolution of system problems by analyzing Call Home messages sent from your devices and providing a direct notification path to Cisco customer support. Smart Call Home offers the following features:

- Continuous device health monitoring and real-time diagnostics alerts.
- Analysis of Call Home messages from your device and where appropriate, automatic Service Request (SR) generation, routed to the appropriate TAC team, including detailed diagnostic information to speed problem resolution.
- Web-based access to Call Home messages and recommendations, inventory and configuration information for all Call Home devices, including a history summary report. Provides access to associated Field Notices, Security Advisories and End-of-Life Information. It also provides recommendations for known issues, including those for which Service Requests were raised, and best practices.

Smart Call Home now offers a low touch registration process. Customers do not need to know their device serial number or contract information. They can register devices without manual intervention from Cisco by sending a message from those devices.

Syslog and Onboard Failure Logging

The MDS 9000 Family syslog capabilities greatly enhance debugging and management. Syslog severity levels can be set individually for all MDS 9000 NX-OS functions, facilitating logging and display of messages ranging from brief summaries to very detailed information for debugging. Messages can be selectively routed to a console and to log files. Messages are logged internally, and they can be sent to external syslog servers, where further analysis can eventually take place.

Specific critical events, error conditions, and important statistics are also automatically recorded with their time stamps in nonvolatile storage (NVRAM) onboard the MDS 9000 Family switch and director line cards. This onboard failure logging (OBFL) capability provides an event data recorder for networking devices and comes extremely useful for performing root-cause analyses of slow-drain situations even after they are cleared.

Post-mortem analysis of failed cards is also possible by retrieving the stored information. OBFL is on by default on all MDS 9000 Family switches and director line cards.

The OBFL process on each line card runs separately at (typically) 20-second intervals and records any counter that has changed value in the last interval. When it detects a counter that has changed value, it records the following information:

- Interface or interface range
- Counter name
- Current counter value
- Date and time of when OBFL detected the counter's changed values

To determine the amount the counter has incremented in the OBFL interval, the previous counter value for same counter name for the same interface must be subtracted from the current counter value. There are various sections in OBFL, and they have different purposes. The main sections are:

- cpuhog: Information about processes using excessive CPU
- environmental-history: Information about temperature sensors
- error-stats: Information about errors related to performance, congestion and slow drain
- interrupt-stats: Information about various module interrupts
- slowport-monitor-events: Information about slow-port monitor
- txwait: Information about the amount of time that interfaces spend at zero Tx credits
- stack-trace: Information about process crashes

Each of these recorded events can be displayed starting at a request date and time and ending at a specific date and time. This capability allows problems that occurred even months ago to be investigated. This is often the first place to look after a problem has occurred. OBFL is a unique feature of Cisco storage networking devices and is considered extremely valuable by support specialists.

The following example shows how frame drops would be time-stamped, if they happened within the reporting interval, so that it is easier to correlate frame drops within the switch with external notification of drops. A simple counter of drops with no time stamps would not serve this purpose.

```
RTP-SAN-15-10-9148s-1#  show  logging  onboard  starttime  05/11/16-00:00:00  error-
stats
----------------------------
    Supervisor Module:
----------------------------

-------------------------------
Module: 1 error-stats
-------------------------------

----------------------------
    Module:  1
----------------------------
```

```
-------------------------------------------------------------------------------
 ERROR STATISTICS INFORMATION FOR DEVICE : FCMAC
-------------------------------------------------------------------------------
    Interface      |                                |          |    Time Stamp
      Range        |    Error Stat Counter Name     |  Count   |MM/DD/YY HH:MM:SS
                   |                                |          |
-------------------------------------------------------------------------------
fc1/1              |FCP_SW_CNTR_RX_WT_AVG_B2B_ZERO  |2         |05/18/16 01:50:08
fc1/1              |FCP_SW_CNTR_RX_WT_AVG_B2B_ZERO  |1         |05/17/16 20:50:02
fc1/10             |IP_FCMAC_CNT_STATS_ERRORS_RX_BAD|9         |05/17/16 18:40:19
fc1/9              |IP_FCMAC_CNT_STATS_ERRORS_RX_BAD|8         |05/17/16 18:40:19
fc1/8              |IP_FCMAC_CNT_STATS_ERRORS_RX_BAD|21        |05/17/16 18:40:19
fc1/1              |IP_FCMAC_CNT_STATS_ERRORS_RX_BAD|1070      |05/17/16 18:22:19
fc1/2              |IP_FCMAC_CNT_STATS_ERRORS_RX_BAD|2         |05/17/16 18:20:18
fc1/10             |IP_FCMAC_CNT_STATS_ERRORS_RX_BAD|19        |05/16/16 10:51:12
fc1/10             |IP_FCMAC_CNT_STATS_ERRORS_RX_BAD|19        |05/12/16 12:23:17
fc1/10             |IP_FCMAC_CNT_STATS_ERRORS_RX_BAD|93        |05/12/16 12:18:57
fc1/8              |IP_FCMAC_CNT_STATS_ERRORS_RX_BAD|38        |05/12/16 12:08:57
fc1/8              |IP_FCMAC_CNT_STATS_ERRORS_RX_BAD|4         |05/11/16 11:43:22
fc1/10             |IP_FCMAC_CNT_STATS_ERRORS_RX_BAD|52        |05/11/16 11:07:21
fc1/8              |IP_FCMAC_CNT_STATS_ERRORS_RX_BAD|41        |05/11/16 11:03:21
fc1/10             |IP_FCMAC_CNT_STATS_ERRORS_RX_BAD|154       |05/11/16 11:02:21
fc1/10             |IP_FCMAC_CNT_STATS_ERRORS_RX_BAD|100       |05/11/16 10:57:21
fc1/10             |IP_FCMAC_CNT_STATS_ERRORS_RX_BAD|39        |05/11/16 10:55:21
fc1/1              |VIP_TMM_TO_DROP_CNT             |13        |05/11/16 09:44:39
fc1/1              |VIP_TMM_TO_DROP_CNT             |11        |05/11/16 09:44:19
fc1/2              |VIP_TMM_TO_DROP_CNT             |111608064|05/11/16 09:43:59
* VIP_TMM_TO_DROP_CNT incremented by 2 in the last 20 seconds
```

Port Monitor and Alerting

The MDS 9000 Family port-monitor (PMON) capability is a well-known feature and is available to all MDS 9000 Family switches. PMON was normally configured on a per-switch basis, but starting with DCNM 10 there is an ability to configure port-monitor policies and to push them to all or a subset of switches. This makes the deployment of port-monitor much simpler and consistent at scale. With NX-OS 6.2(13) an alerting function can complement PMON for an enhanced user experience. Thanks to PMON, administrators can deploy an easy-to-use capability to monitor physical-port critical counters. Additionally, it is possible to automatically alert external monitoring software of any anomalies. The configured entities are monitored in hardware directly, saving computing resources on the supervisor. Whenever a monitored entity crosses a preconfigured threshold value, an automated action is triggered whereby the SNMP component of the PMON feature generates an alerting message. You can also use PMON in combination with the port guard feature to error-disable or flap the port when the set threshold is met for a configured counter.

Administrators can optionally use DCNM software to view these alerts in an intuitive way and gain fabric-wide control by identifying situations that are hampering the flow of data traffic. Specifically, alerts can be seen in the DCNM web client on the Health > Events tab.

Currently, there are 19 hardware counters that can be monitored with PMON. Counters for transmitted and received link reset (LR) frames are among them. The number of times that credit loss recovery was initiated because a port had 0 Tx credits for 1 to 1.5 seconds is another counter (most severe indication of congestion). Quite important, the Txwait counter can also be monitored, and it measures the time that a port remains at 0 Tx credits as frames are queued. Out of the 19 parameters, 9 are applicable to situations in which devices are causing congestion in the SAN and can be blended into a custom PMON policy.

To help administrators, two predefined PMON policies are built into NX OS: “default” and “slow-drain”. In accordance with customers’ requests, in recent NX-OS releases “slow-drain” policy is made active automatically. It is worth noting that only one policy can be active at a time on a particular port type, so before you apply a new policy, you must first deactivate the active one. However, it is possible to have a policy for access ports and another policy for ISL ports active at the same time. If needed, custom PMON policies can be created as well. This allows for tailoring the policy to specific customer needs and also leverage counters not included in the predefined policies. The following custom policy is provided as an example. It alerts on slow-drain conditions, and it would apply to both access ports (F-ports) and trunk ports (E-ports):

```
port-monitor name AllPorts
  port-type all
  no monitor counter link-loss
  no monitor counter sync-loss
  no monitor counter signal-loss
  no monitor counter invalid-words
  no monitor counter invalid-crc
  counter tx-discards poll-interval 60 delta rising-threshold 50 event 3 falling-
threshold 10 event 3
  counter lr-rx poll-interval 60 delta rising-threshold 5 event 2 falling-
threshold 1 event 2
  counter lr-tx poll-interval 60 delta rising-threshold 5 event 2 falling-
threshold 1 event 2
  counter timeout-discards poll-interval 60 delta rising-threshold 50 event 3
falling-threshold 10 event 3
  counter credit-loss-reco poll-interval 60 delta rising-threshold 1 event 2
falling-threshold 0 event 2
  counter tx-credit-not-available poll-interval 1 delta rising-threshold 10 event
4 falling-threshold 0 event 4
  counter rx-datarate poll-interval 10 delta rising-threshold 80 event 4 falling-
threshold 50 event 4
  counter tx-datarate poll-interval 10 delta rising-threshold 80 event 4 falling-
threshold 50 event 4
  no monitor counter err-pkt-from-port
  no monitor counter err-pkt-to-xbar
  no monitor counter err-pkt-from-xbar
  no monitor counter tx-slowport-count
  counter tx-slowport-oper-delay poll-interval 1 absolute rising-threshold 50
event 4 falling-threshold 0 event 4
  counter txwait poll-interval 1 delta rising-threshold 20 event 4 falling-
threshold 0 event 4
```

Now activate the policy named AllPorts:

```
MDS9710-1# conf t
MDS9710-1(config)# port-monitor activate AllPorts
MDS9710-1(config)# end
```

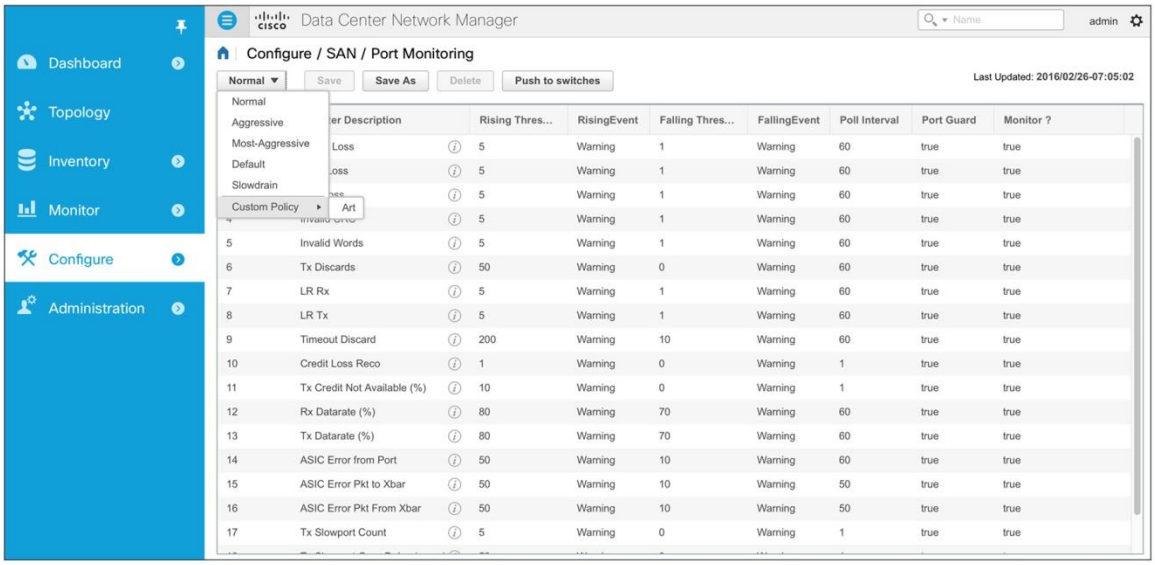
This is how you can verify the new policy after activation:

```
 MDS9710-1# show port-monitor active

Policy Name  : AllPorts
Admin status : Active
Oper status  : Active
Port type    : All Ports
--------------------------------------------------------------------------------
------------------------
Counter                       Threshold  Interval Rising Threshold event Falling
Threshold  event PMON Portguard
-------                   ---------  -------- ---------------- ----- -------------
----- ----- --------------
TX Discards                  Delta       60        50                  3      10
3     Not enabled
LR RX                         Delta       60        5                   2      1
2     Not enabled
LR TX                         Delta       60        5                   2      1
2     Not enabled
Timeout Discards            Delta       60        50                  3      10
3     Not enabled
Credit Loss Reco            Delta       60        1                   2      0
2     Not enabled
TX Credit Not Available   Delta       1          10%                 4      0%
4     Not enabled
RX Datarate                  Delta       10        80%                 4      50%
4     Not enabled
TX Datarate                  Delta       10        80%                 4      50%
4     Not enabled
TX-Slowport-Oper-Delay    Absolute    1          50ms                4      0ms
4     Not enabled
TXWait                        Delta       1         20%                 4      0%
4     Not enabled
-
```

Rather than requiring administrators to deal with individual switches one at a time, DCNM provides administrators with a fabric-wide capability to make PMON configuration changes and collect alerts. The two prebuilt or custom created policies can be used and pushed to the selected switches by port type. For a faster implementation, thresholds for monitored counters can be set according to three predefined monitoring templates: normal, aggressive, and most aggressive (Figure 13). These templates represent Cisco best practices and guidance matured through multiple years of in-field experience building high performing and reliable Fibre Channel networks.

Figure 13. Port Monitor User Interface

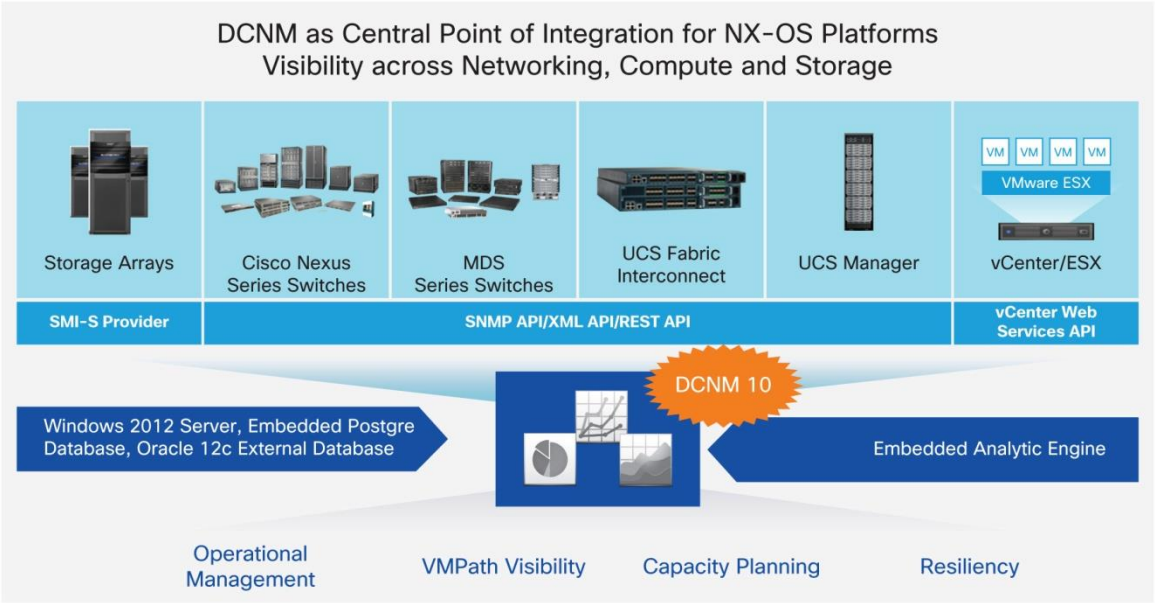


PMON allows the classification of events in the following 5 categories: Informational (5), Warning (4), Error (3), Critical (2), Fatal (1). DCNM Health -> Events displays the different classifications appropriately to allow the most important events to be easily identified.

Cisco Prime DCNM and SAN Host-Path Redundancy Analysis

DCNM can further provide ease of operation beyond what is intrinsically available from MDS 9000 Family devices and their CLI. In fact, DCNM has been designed to help address operational challenges for storage teams above and beyond what is possible with individual device-based tools. DCNM has multidomain visibility across computing, network, and storage domains (Figure 14). One of the most prominent of these features is SAN host-path redundancy analysis. When used, it provides visibility and checks connectivity to maintain Fibre Channel network best practices in mission-critical environments. For reliable production deployments, the paths must be redundant between server and storage arrays, with physical fabric separation. DCNM can run automated or on-demand host-redundancy checks to see if this best practice is maintained. DCNM checks for hardware failures and path misconfiguration and uses multidomain visibility to look beyond Fibre Channel networks and into storage arrays and virtual computing resources. In addition to preventing possible outages due to missing redundancy, the feature is used in the change management window for switch and array upgrades or during normal operations for replacement of hardware such as switch modules, SFPs and front-end array adapters. DCNM shows value even for slow-drain troubleshooting and analysis, providing a graphical and intuitive view of the relevant counters. With DCNM, it is also possible to generate the tech support file for an entire fabric with a single click, saving valuable time otherwise spent collecting individual files from multiple devices.

Figure 14. Cisco Prime Data Center Network Manager



Conclusion

The list of features that MDS 9000 Family switches provide to support enhanced serviceability, troubleshooting, and diagnostics is long. This document described some of them. In particular, this document discussed the new generation of 16-Gbps-capable MDS 9000 Family switches and their support for advanced ASICs and specialized firmware to greatly improve management at scale, diagnostics, error recovery, and troubleshooting in SANs. Some features are inherited by the previous generation of products and driven by years of experience in designing and operating SANs. Other features are newer and integrated into the latest iteration of ASICs, but they do not require additional hardware or licenses. The positive impact of these capabilities is even more apparent for large, complex, extended multisite networks. All of the mentioned capabilities help achieve secure and robust data transport and optimize bandwidth utilization, reducing operational expenditures with built-in diagnostics and SAN management tools. The MDS 9000 Family of storage networking devices paired with the DCNM management tool provides IT administrators with all the required monitoring and forecasting tools needed for safe operation of the network and a consolidated view of the data center infrastructure.

For More Information

- Landing page for MDS 9000 Family and data sheets:
  - http://www.cisco.com/go/mds
  - http://www.cisco.com/c/en/us/products/storage-networking/mds-9700-series-multilayer-directors/datasheet-listing.html
- To learn more about buffer-to-buffer credit recovery:
  - http://www.cisco.com/c/en/us/td/docs/switches/datacenter/mds9000/sw/7_3/configuration/interfaces/interfaces/buffers.html

- To learn more about forward error correction:
  - http://www.cisco.com/c/en/us/td/docs/switches/datacenter/mds9000/sw/7_3/configuration/interfaces/interfaces/gen2.html#67390
- To learn more about digital optical monitoring:
  - http://www.cisco.com/c/en/us/td/docs/switches/datacenter/mds9000/sw/command/b_cisco_mds_9000_cr_book/show_commands.html#wp3660796084
- To learn more about ISL diagnostics:
  - http://www.cisco.com/c/en/us/td/docs/switches/datacenter/mds9000/sw/7_3/configuration/sysmgmt/sysmgmt/isl_diag.html
- To learn more about Fibre Channel ping, Fibre Channel traceroute, and pathtrace:
  - http://www.cisco.com/c/en/us/td/docs/switches/datacenter/mds9000/sw/command/b_cisco_mds_9000_cr_book/f_commands.html#wp1864894660
  - http://www.cisco.com/c/en/us/td/docs/switches/datacenter/mds9000/sw/command/b_cisco_mds_9000_cr_book/t_commands.html#wp3490502110
  - http://www.cisco.com/c/en/us/td/docs/switches/datacenter/mds9000/sw/command/b_cisco_mds_9000_cr_book/p_commands.html#wp8176664670
- To learn more about smart call home:
  - http://www.cisco.com/c/en/us/td/docs/switches/datacenter/mds9000/sw/7_3/configuration/sysmgmt/sysmgmt/call.html
- To learn more about system onboard failure logging:
  - http://www.cisco.com/c/en/us/td/docs/switches/datacenter/mds9000/sw/7_3/configuration/sysmgmt/sysmgmt/sys.html
- To learn more about Cisco GOLD:
  - http://www.cisco.com/c/en/us/td/docs/switches/datacenter/mds9000/sw/7_3/configuration/sysmgmt/sysmgmt/gold.html
- To learn more about PMON:
  - http://www.cisco.com/c/dam/en/us/products/collateral/storage-networking/mds-9700-series-multilayer-directors/white-paper-c11-736963.pdf
- To learn more about Cisco Prime DCNM:
  - http://www.cisco.com/c/en/us/products/collateral/cloud-systems-management/prime-data-center-network-manager/datasheet-c78-736613.html

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 C11-737404-00 07/16