White Paper

Slow-Drain Device Detection, Troubleshooting, and Automatic Recovery

Contents

What You Will Learn

Modern day data centers are observing unprecedented data growth. The amount of storage is increasing. Numbers of applications and servers are increasing. Storage area networks (SAN) provide the connectivity between servers and storage, are being pushed to their limits. Full capacity is expected 24 hours a day, 365 days a year. However, congestion in a SAN cripples application performance instantly. It is imperative for SAN administrators to build robust and self-healing networks.

Through this document you will learn:

- The concept of congestion in SAN, especially slow drain, which is the most strenuous type of congestion. Understanding the basic concepts helps you effectively solve the problem.
- The architectural benefits of Cisco® MDS 9000 Family switches. These are the only Fibre Channel switches in the industry that provide consistent and predictable performance and prevent microcongestion problems, such as head-of-line blocking. Cisco MDS 9000 Family switches build robust Fibre Channel networks.
- Cisco MDS 9000 Family switches are the only Fibre Channel switches in the industry that provide automatic recovery from slow drain even in large environments. You learn the holistic approach taken by Cisco to detect, troubleshoot, and automatically recover from slow drain.
- Troubleshooting methodology developed by Cisco while working on large SAN environment over more than a decade.
- Enhancements to Cisco Data Center Network Manager (DCNM) for fabricwide detection and troubleshooting of slow drain to help you find and fix problems within minutes using an intuitive web interface.

Scope

This document covers all 16-Gbps Fibre Channel products under the Cisco MDS 9000 Family switches. Advanced 8 Gbps and 8-Gbps line cards on Cisco MDS 9500 directors are also covered. Details are listed in Table 1. Appendix C contains various commands that can used on these platforms. Feature support matrix across platforms and Cisco NX-OS Software releases are available under Appendix E. At the time of writing, the document recommends NX-OS Release 6.2(13), or later, and DCNM release 7.2(1), or later.

Table 1. Platforms Discussed and Supported in This Document Under Cisco MDS 9000 Family Switches

Platforms under Cisco MDS 9000 Family	Model
16-Gbps Platforms	Cisco MDS 9700 Series Multilayer Directors with DS-X9448-768K9 line card
	MDS 9396S
	MDS 9148S
	MDS 9250i
Advanced 8-Gbps Platforms	Cisco MDS 9500 Series Multilayer Directors with DS-X92xx-256K9 line cards
8-Gbps Platforms	Cisco MDS 9500 Series Multilayer Directors with DS-X9248-48K9 and DS-X92xx-96K9 line cards

Introduction

We are in the era of digitalization, mobility, social networking, and Internet of Everything (IoE). More and more applications are being developed to support businesses. Many of the newer generation organizations are functioning only through applications. These applications must perform at highest capacity, all the time. Data (processing, reading, and writing) is the most important attribute of an application. Applications are hosted on servers. Data is stored on storage arrays. Connectivity between servers and storage arrays is provided by SANs. Fibre Channel (FC) is the most commonly deployed technology to build SANs. FC SANs must be free of congestion so that application performance is at peak. If not, businesses are prone to huge revenue risk due to stalled or poor-performing applications.

Congestion in FC SANs has always been the highest concern for SAN administrators. The concern has become even more severe due to the following reasons:

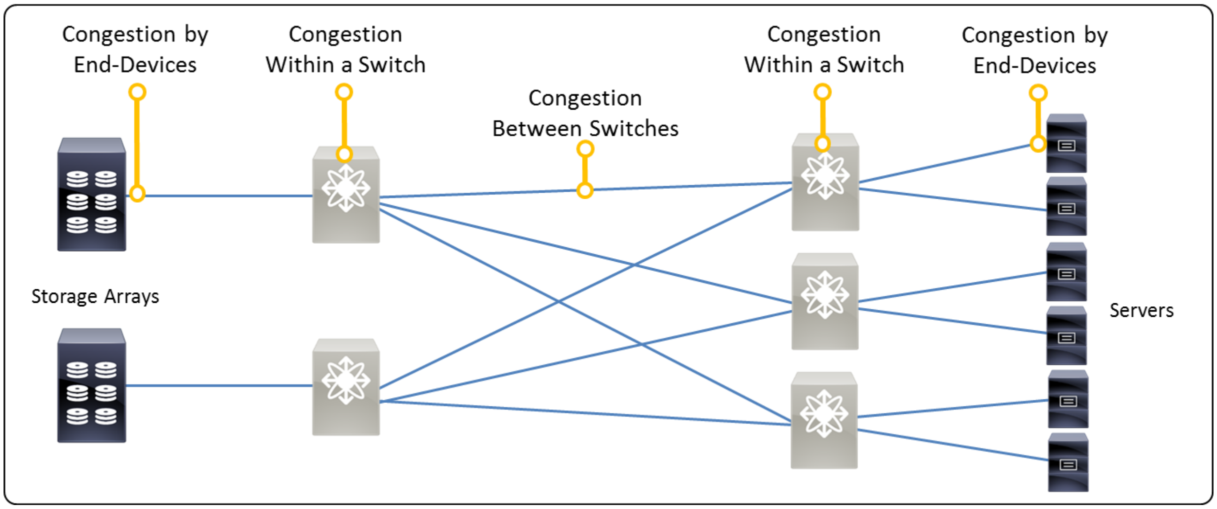
- **Adoption of 16-Gbps FC leading to heterogeneous speeds:** The last few years have seen increased adoption of 16-Gbps FC. While newer devices at 16 Gbps are connected, older devices at 1-, 2-, 4-, or 8-Gbps FC still remain as part of the same fabric. Servers and storage ports at different speeds sending data to each other tend to congest network links.
- **Data explosion leading to scaled out architectures:** Application and data explosion is resulting in more servers and storage ports. FC SANs are being scaled out. Collapsed core architectures are being scaled to edge-core architectures. Edge-core architectures are being scaled to edge-core-edge architectures. Larger networks have more applications that are impacted due to SAN congestion.
- **Legacy application and infrastructure:** While newer high-performing applications and servers are being deployed, the older and slower servers running legacy applications are still being used. This results in a common network being shared by fast and slow applications. SAN performance acceptable by a slower application may be completely unacceptable by a faster application.
- **Increased pressure on operating expenses (OpEx):** Businesses are trying to find ways to increase their bottom lines. The pressure on OpEx has never been more. Stress is increasing to fully use the existing infrastructure. SANs must be free of congestion to keep applications' performance at peak.
- **Adoption of flash storage:** More and more businesses are deploying flash storage for better application performance. Flash storage is several times faster than a spinning disk. It is pushing SANs to their limits. The existing links may not be capable enough of sustaining the bandwidth.

Cisco MDS 9000 Family switches have purposefully designed architecture, hardware, and software to keep FC SANs free of congestion. High performance is delivered by integrating the features directly with the switch port's application specific integrated circuit (ASIC). Operational simplicity is achieved by software enhancements made on Cisco NX-OS Software. Problems are being solved within minutes by the web-based, fabricwide and single-pane-of-glass visibility by Cisco Data Center Network Manager (DCNM). Overall, Cisco has taken a holistic approach to build robust and self-healing FC SAN. The details are provided in the following sections.

Understanding Storage Area Network Congestion

Storage area networks (SANs) are built of end devices, switches, and connecting links. Any of these devices can be the source of congestion (Figure 1).

Figure 1. Fibre Channel Network and Source of Congestion


Congestion by End-Devices
Congestion Within a Switch
Congestion Between Switches
Congestion Within a Switch
Congestion by End-Devices
Storage Arrays
Servers

Congestion from End Devices

FC SANs are lossless networks. All frames are acknowledged by the receiver. The sender stops sending frames if acknowledgments are not received. The inability of a receiver to receive frames at the expected rate results in congestions. Slow drain is a typical type of SAN congestion, mostly caused by misbehaving end devices. Further sections go into the details of slow drain. Information is provided to detect, troubleshoot, and automatically recover from the situation.

Congestion Between Switches

Inter-Switch Link (ISL) build the core of a network. Unlike the links that are connected to a single end device, ISLs carry traffic between multiple end devices. The traffic pattern defines the oversubscription ratio. It is assumed that not all the end devices transmit data at the same time at peak rate. For example, a Cisco MDS 9710 Multilayer Director can have 384 ports at a 16-Gbps line rate. There are 320 ports connected to servers and 64 connected to other switches towards storage arrays that provide an oversubscription ratio of 5:1. With newer applications and increased workloads, the ISLs may reach their full capacity. A change of oversubscription ratio to 4:1 or 3:1 may be desired. Numerous features on NX-OS and DCNM monitor the link utilization. Automatic alerts can be generated if link utilization exceeds configured thresholds. Performance trending and forecasting on DCNM provides early notification to SAN administrators well ahead in time so that peak application performance can be maintained.

SAN administrators should carefully analyze bandwidth of all individual links, even though multiple links grouped together (as a single port channel) can provide an acceptable oversubscription ratio. By default, the load-balancing scheme of Cisco MDS 9000 Family switches is based on source FCID (SID), destination FCID (DID) and exchange ID (OXID). All the frames of an exchange from a target to a host traverse through the same physical link of a port channel. In production networks, large number of end devices and exchanges provide uniformly distributed traffic pattern. However, in some corner cases, large-size exchanges can congest a particular link of a port channel if the links connected to end devices are of higher bandwidth than the individual bandwidth of any of the members of the port channel. To alleviate such problems, Cisco recommends that ISLs should always be of higher or similar bandwidth than that of the links connected to the end devices.

Lack of B2B Credits for the Length of the ISL

Number of buffer-to-buffer (B2B) credits should be carefully accounted on long-distance ISLs.

Note: B2B credits and Fibre Channel flow control has been described in the following section, "Flow Control in Fibre Channel."

The requirement of B2B credits on a Fibre Channel link increases with:

- Increase in distance
- Increase in speed
- Decrease in frame size

Table 2 provides numbers of B2B credits required for per-kilometer length of ISL at different speeds and frame sizes. Notice that one B2B credit is needed per frame irrespective of the frame size. Fewer numbers of B2B credits can be the reason for performance impact if the received B2B credits available on a port is close to the value as calculated by Table 2. Consider using extended B2B credits or moving long-distance ISLs to other platforms under Cisco MDS 9000 Family switches that have more B2B credits.

Table 2. Per-Kilometer B2B Credit Requirement at Different Speeds and Frame Sizes

Frame Size	1 Gbps	2 Gbps	4 Gbps	8 Gbps	10 Gbps	16 Gbps
512 bytes	2 BB/km	4 BB/km	8 BB/km	16 BB/km	24 BB/km	32 BB/km
1024 bytes	1 BB/km	2 BB/km	4 BB/km	8 BB/km	12 BB/km	16 BB/km
2112 bytes	0.5 BB/km	1 BB/km	2 BB/km	4 BB/km	6 BB/km	8 BB/km

Congestion in Switches

Fibre Channel switches available on the market today have hundreds of ports. These switches are expected to receive, process, and forward frames from any port to any port at line rate. Different vendors have different architectures. Some vendor switches have physical ports at 16 Gbps but can't switch frame at that speed on all ports and at all frame sizes. This results in severe performance degradation of applications. SAN administrators must understand the internal architecture before making a buying decision. They must ensure that the switches have been architected for:

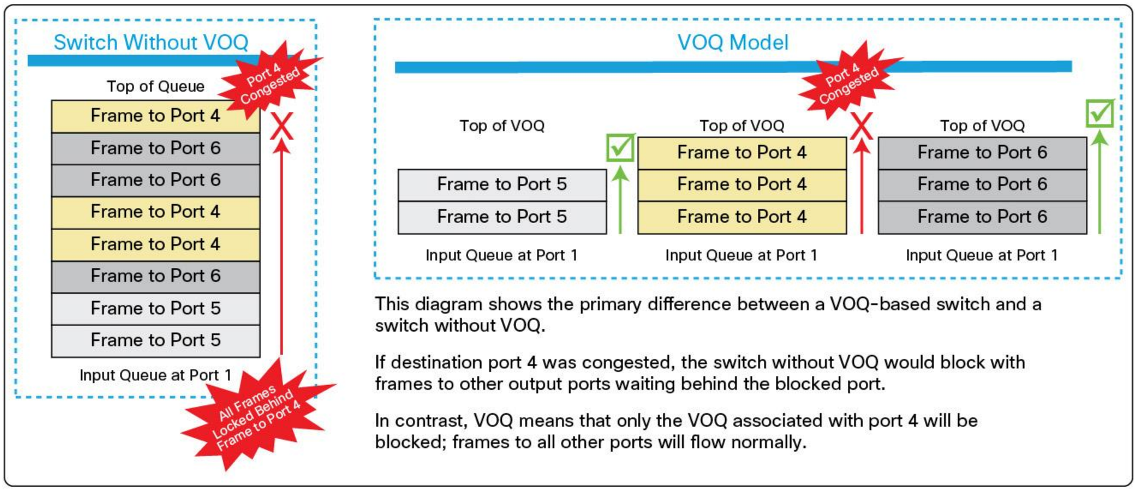
- Nonblocking line-rate performance at all ports at all frame sizes
- Consistent performance between all ports, without dependency on local-switching
- Predictable performance between all ports, irrespective of what features are enabled
- No head-of-line blocking
- Centralized coordinated forwarding between all ports, rather than each port acting on its own

If these factors are not considered well in advance, SAN administrators risk their networks to severe congestion within a switch. Such problems cannot be solved in a production network. The only solution would be the expensive approach of buying more switches or contracting professional services.

Cisco MDS 9000 Family switches have been architected to provide these benefits. Following are the unique advantages that ensure that switches are always free of congestion.

- **No limitations on per-slot bandwidth:** The Cisco MDS 9700 Series Multilayer Directors supports up to 1.5-Tbps per-slot bandwidth today. This is two times of the capacity which is required to support 48 line-rate ports at 16 Gbps. All ports on all slots are capable of sending line rate traffic to all other ports using nonblocking and non-oversubscribed design.
- **Centralized coordinated forwarding:** Cisco MDS 9000 Family switches use a centrally arbitrated crossbar architecture for frame forwarding between ports. Central arbitration ensures that frames are handed over to an egress port only when it has enough transmit buffers available. After the arbiter grants the request, a crossbar provides a dedicated data link between ingress and egress ports. There is never a situation when frames are unexpectedly dropped in the switch.
- **Consistent and predictable performance:** All frames from all ports are subject to central arbitrated crossbar forwarding. This ensures that the end applications receive consistent performance irrespective of where the server and storage ports are connected on a switch. There is no limitation of connecting the ports to the same port group on the same module to receive lower latency. Also, performance is not degraded if more features are enabled. Consistency and predictability lead to better designed and operated networks.
- **Store-and-forward architecture:** Frames are received and stored completely in the ingress port buffers before they are transmitted. This enables Cisco MDS 9000 Family switches to inspect the cyclic redundancy check (CRC) field of a Fibre Channel frame and eventually drop them if the frames are corrupted. This intrinsic behavior limits the failure domain to a port. Corrupt frames are not spread over the fabric. End devices are not bombarded with corrupt frames.
- **Virtual output queues (VOQs):** VOQ is the mechanism that prevents head-of-line blocking inside a Cisco MDS 9000 Family switch. Head-of-line occurs when the frame at the head of the queue cannot be sent because of congestion at its output port, while the frames behind this frame are blocked from being sent to their destination, even though their respective output ports are not congested. Instead of a single queue, separate VOQs are maintained at all ingress ports. Frames destined to different ports are queued to separate VOQs. Individual VOQs can be blocked, but traffic queued for different (nonblocked) destinations can continue to flow without being delayed behind frames waiting for the blocking to clear on a congested output port (Figure 2). Cisco MDS 9000 Family switches support up to 4096 VOQs per port, allowing to address up to 1024 destination ports per chassis, with 4 QoS levels.

Figure 2. Cisco MDS 9000 Family Virtual Output Queue



This diagram shows the primary difference between a VOQ-based switch and a switch without VOQ.

If destination port 4 was congested, the switch without VOQ would block with frames to other output ports waiting behind the blocked port.

In contrast, VOQ means that only the VOQ associated with port 4 will be blocked; frames to all other ports will flow normally.

All these attributes are unique only to the architecture of Cisco MDS 9000 Family switches.

The architecture of Cisco MDS 9000 Family switches has been explained in details in a white paper available on cisco.com: "Cisco MDS 9000 Family Switch Architecture." This document is focused on congestion from end devices, especially slow drain.

Introduction to Slow Drain

A slow drain device is a device that does not accept frames at a rate generated by the source. In the presence of slow-drain devices, Fibre Channel networks are likely to lack frame buffers, resulting in switch port buffer starvation and potentially choking ISLs. The impact of choked ISLs is observed on other devices that are not slow-drain devices but share the same switches and ISLs. As the size of the fabric grows, more and more ports are impacted by a slow-drain situation. Because the impact is seen across a large number of ports, it becomes extremely important to detect, troubleshoot, and immediately recover from the situation. Traffic flow is severely impacted due to which applications face latency issues or stop responding completely until recovery is made or if the slow-drain device is disconnected from the fabric.

Following are reasons for slow drain on edge devices and ISLs.

Edge Devices

An edge device can be slow to respond for a variety of reasons:

- Server performance problems: applications or the OS.
- Host bus adapter (HBA) problems: driver or physical failure.
- Speed mismatches: one fast device and one slow device.
- Nongraceful virtual machines exit on a virtualized server, resulting in frames held in HBA buffers.
- Storage subsystem performance problems, including overload.
- Poorly performing tape drives.

ISLs

- Lack of B2B credits for the distance the ISL is traversing
- The existence of slow-drain edge devices

Any device exhibiting such behavior is called a *slow-drain device*.

Cisco Solution

Cisco has taken a holistic approach by providing features to detect, troubleshoot, and automatically recover from slow drain situations. Detecting a slow-drain device is the first step, followed by troubleshooting, which enables SAN administrators to take manual action of disconnecting an offending device. However, manual actions are cumbersome and involve delay. To alleviate this limitation, Cisco MDS 9000 Family switches have intelligence to constantly monitor the network for symptoms of slow drain and send alerts or take automatic recovery actions. These actions include:

- Drop all frames queued to a slow drain device.
- Drop new frames destined to slow drain device at line rate.
- Perform link reset on the affected port.
- Flap the affected port.
- Error disable the port.

All the 16-Gbps MDS platforms (MDS 9700, MDS 9396S, MDS 9148S and MDS 9250i) provide hardware enhanced slow-drain features. These enhanced features are a direct benefit of advanced capabilities of port ASIC. Following is the summary of advantages of hardware enhanced slow drain features:

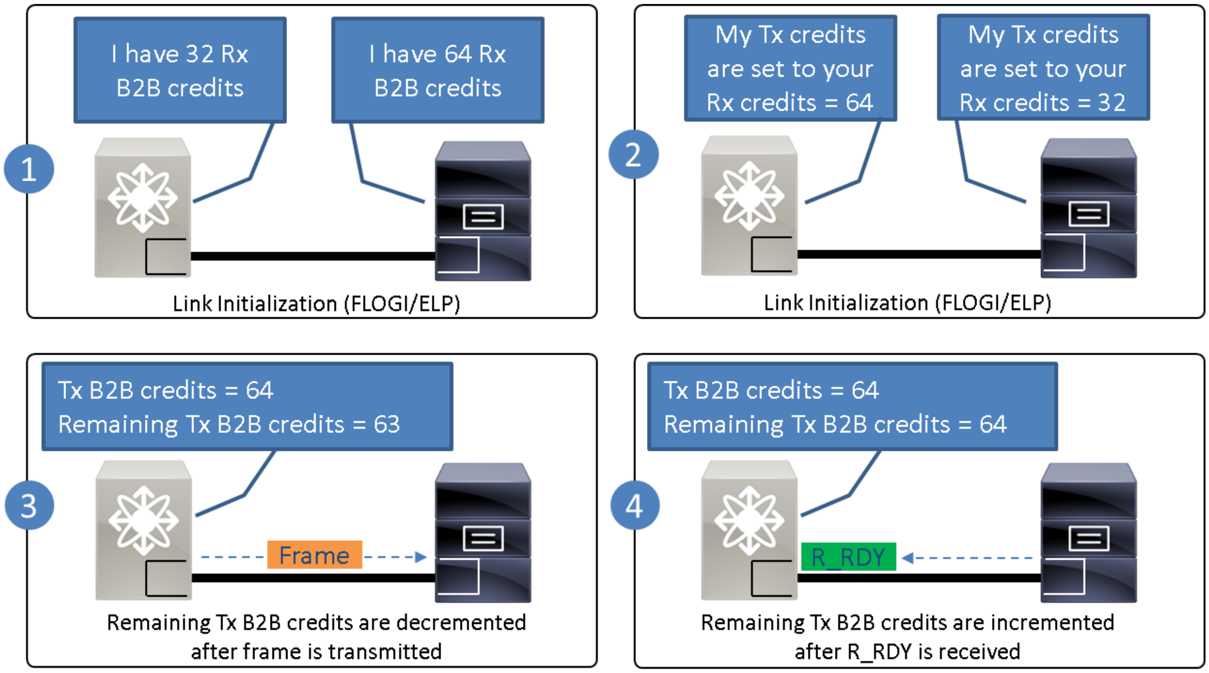
- Detection granularity in microseconds (µs) using port ASIC hardware counters
- New feature called *slowport-monitor*, which maintains history of transmit credit unavailability duration on all the ports at as low as 1 millisecond (ms)
- Graphical display of credit unavailability duration on all the ports on a switch over last 60 seconds, 60 minutes, and 72 hours
- Immediate automatic recovery from a slow-drain situation without any software delay

In addition to the hardware-enhanced slow-drain features on Cisco MDS 9000 Family switches, Cisco DCNM provides slow-drain diagnostics from Release 7.1(1) and later. DCNM automates the monitoring of thousands of ports in a fabric in a single pane of glass and provides visual representation in form of graphs showing fluctuation in counters. This feature leads to faster detection of slow-drain devices, reduced false positives, and reduced troubleshooting time from weeks to minutes.

Background: Flow Control in Fibre Channel

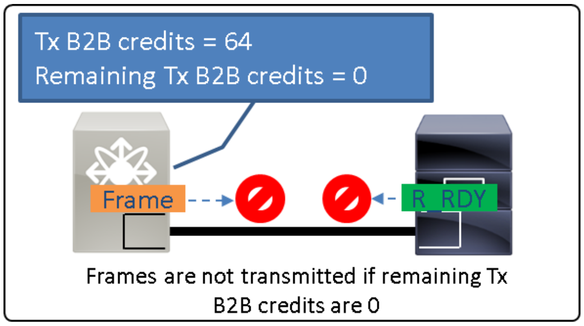
Fibre Channel is designed to build a loss-less network. To achieve this, Fibre Channel implements a credit-based flow-control mechanism (Figure 3).

Figure 3. Flow Control in Fibre Channel


1
I have 32 Rx B2B credits
I have 64 Rx B2B credits
Link Initialization (FLOGI/ELP)
2
My Tx credits are set to your Rx credits = 64
My Tx credits are set to your Rx credits = 32
Link Initialization (FLOGI/ELP)
3
Tx B2B credits = 64
Remaining Tx B2B credits = 63
Frame
Remaining Tx B2B credits are decremented after frame is transmitted
4
Tx B2B credits = 64
Remaining Tx B2B credits = 64
R_RDY
Remaining Tx B2B credits are incremented after R_RDY is received

When a link is established between two Fibre Channel devices, both neighbors inform each other about the available number of receive buffers they have. N_Port connected to F_Port exchange B2B credit information through Fabric Login (FLOGI). E_Port connected to another E_port exchange B2B credit information through Exchange Link Parameter (ELP). Before transmission of data frames, the transmitter sets the transmit (Tx) B2B credits equal to the receive (Rx) B2B credits informed by the neighbor. This mechanism ensures that the transmitter never overruns the receive buffers of the receiver. For every transmitted frame, remaining Tx B2B credits decrement by one. The receiver, after receiving the frame, is expected to return an explicit B2B credit in the form of R_RDY (Receiver_Ready, Fibre Channel Primitive) to the transmitter. A receiver typically does that after it has processed the frame, and the receive buffer is now available for reuse. The transmitter increments the remaining Tx B2B credits by one after receiving R_RDY. The transmitter does not increment the remaining Tx B2B credit if R_RDY is not received. This can be due to the receiver not sending R_RDY or R_RDY being lost on the link. Multiple occurrences of such event eventually lead to a situation where the remaining Tx B2B credit on a port reaches zero. As a result, no further frames can be transmitted. Tx port resumes sending an additional frame only after receiving a R_RDY. This strategy prevents frames from getting lost when the Rx port runs out of buffers (Rx B2B credits) and ensures that the receiver is always in control (Figure 4).

Figure 4. Frames Not Transmitted in Fibre Channel if receiver does not have enough buffers

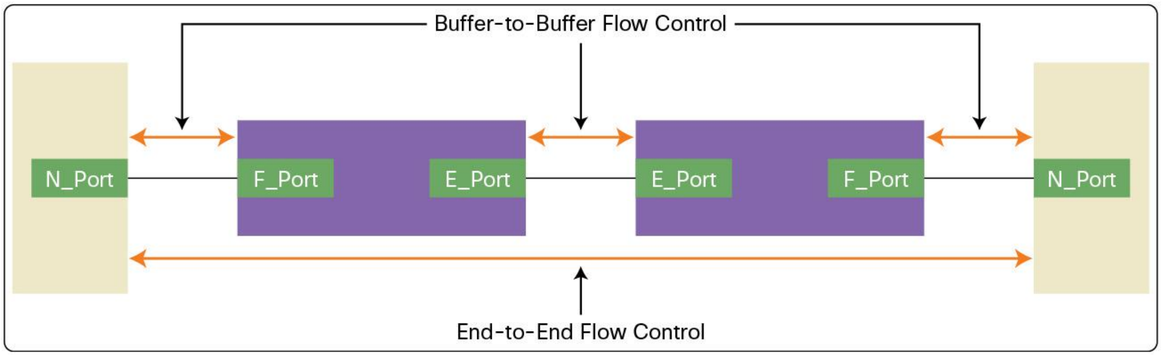

Tx B2B credits = 64
Remaining Tx B2B credits = 0
Frame
R_RDY
Frames are not transmitted if remaining Tx B2B credits are 0

Note: The terms *credit unavailability and zero remaining Tx/Rx B2B credits* signify the same situation. This is also represented by the term *delay on a port* (which means delay in receiving R_RDY to a port or delay in forwarding frames out of a port). these terms have been used interchangeably in the document to convey the same meaning.

Fibre Channel defines two types of flow control (Figure 5):

- Buffer-to-buffer (port to port)
- End-to-end (source to destination)

Figure 5. Types of Flow Control in Fibre Channel

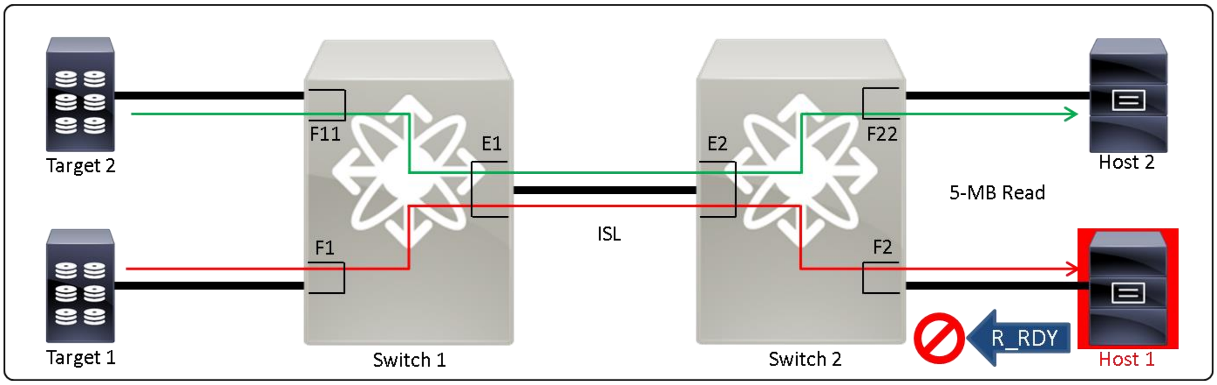


End-to-end flow control was never widely implemented. Buffer-to-buffer (B2B) flow control between every pair of neighbor ensures end-to-end lossless fabric.

Example: Slow Drain

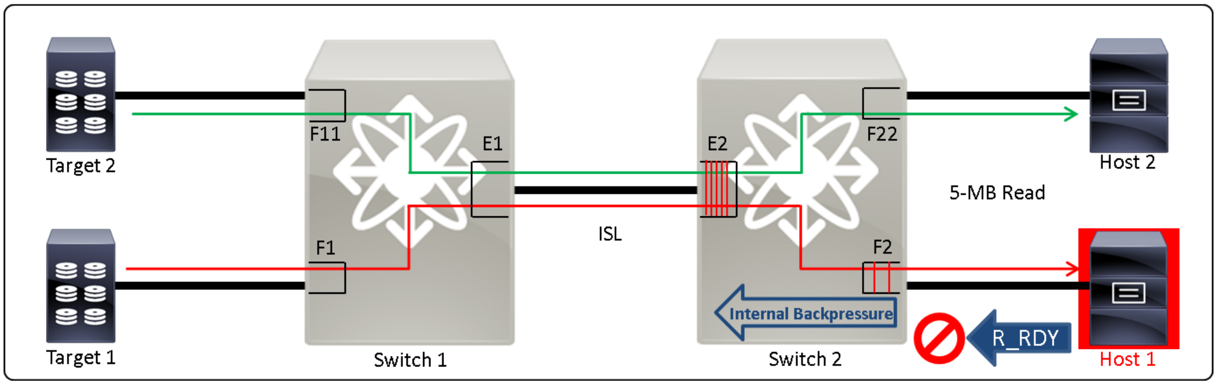
Consider the topology in Figure 6. Host 1 sends a large 5-MB read request to Target 1. A Fibre Channel frame is 2148 bytes. One frame can transport up to 2048 bytes of data. Therefore, the response from the target is approximately 2500 data frames. If Host 1 cannot process all the data frames fast enough, it delays sending back R_RDY to port F2 on Switch 2.

Figure 6. Fibre Channel Flow Control



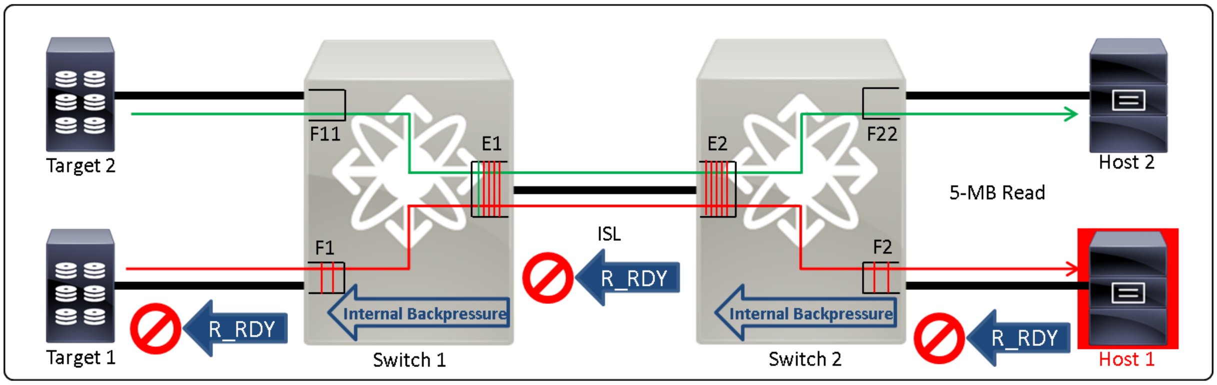
The remaining Tx B2B credits eventually fall to zero. As a result, Switch 2 does not send any further frames out of Port F2. The data frames occupy all the egress buffers on Port F2. This generates an internal backpressure towards Port E2 on switch 2. The data frames occupy all the ingress buffers on Port E2, and soon there are no remaining Rx B2B credits available with Port E2 on switch 2 (Figure 7).

Figure 7. Fibre Channel Flow Control (continued)



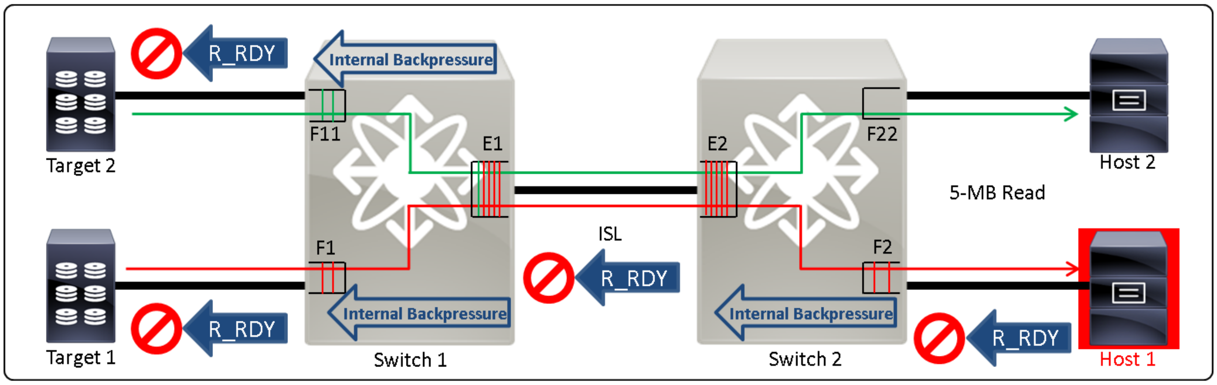
Port E2 does not send R_RDY to Port E1 on switch 1. Data frames start occupying the egress buffers on Port E1, which generates an internal backpressure towards Port F1 on switch 1. Data frames consume all the ingress buffers on Port F1 leading to zero remaining Rx B2B credits. Port F1 stops sending R_RDY to Target 1 (Figure 8).

Figure 8. Fibre Channel Flow Control (continued)



Overall, the R_RDY and internal backpressure of switches have signaled Target 1 to slow down. This is desirable behavior in a Fibre Channel network to make it lossless. However, it brings a side effect. In this example, when the remaining Tx B2B credits fall to zero on Port E1 on switch 1, it generates backpressure to Port F1 as well as to Port F11. Hence, not only the Target 1 – Host 1 flow slows down, Target 2 – Host 2 flow also slows down (Figure 9).

Figure 9. Slow-Drain Situation



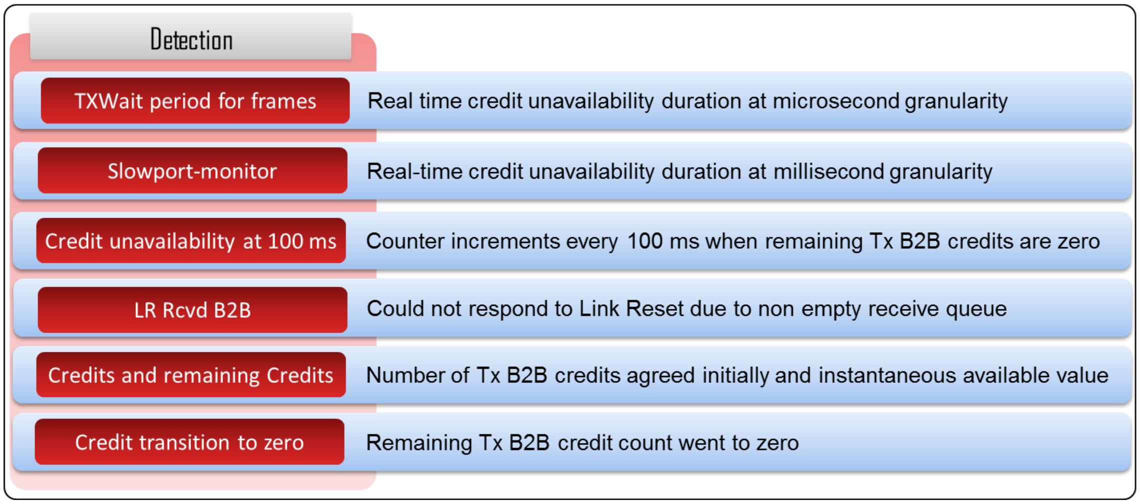
As a final situation, just because one end device in the fabric became slow, it impacted all the flows that were sharing the same switches and ISLs. This situation is known as slow drain. Host 1 in the shown topology is called a slow-drain device.

Slow-drain situations can be compared to a traffic jams on a freeway due to an internal jam in an adjacent cities. Consider a freeway that connects multiple cities. If one of the adjacent cities observes an internal traffic jam that is not resolved fast enough, soon the traffic creates congestion on the freeway, consuming all the available lanes. The obvious effect of this jam is seen on the traffic going to and coming from the congested city. However, because the freeway is jammed, the effect is seen on the traffic that is going to and coming from all other cities using the same freeway but may not be internally congested.

Slow-Drain Detection

Cisco MDS 9000 Family switches provide numerous features to detect slow-drain devices. This section explains these features along with Cisco slow-drain terminology. Figure 10 provides summary of slow-drain detection capabilities on Cisco MDS 9000 Family switches.

Figure 10. Detecting Slow Drain on Cisco MDS 9000 Family Switches



Detection	
TXWait period for frames	Real time credit unavailability duration at microsecond granularity
Slowport-monitor	Real-time credit unavailability duration at millisecond granularity
Credit unavailability at 100 ms	Counter increments every 100 ms when remaining Tx B2B credits are zero
LR Rcvd B2B	Could not respond to Link Reset due to non empty receive queue
Credits and remaining Credits	Number of Tx B2B credits agreed initially and instantaneous available value
Credit transition to zero	Remaining Tx B2B credit count went to zero

Credit Unavailability at Microseconds—TxWait Period for Frames

Cisco MDS 9000 Family switches monitor Tx B2B credit unavailability duration at nanosecond (ns) granularity by incrementing internal counters. The cumulative information of these internal counters is reported by TxWait, which increments if Tx B2B credits are unavailable on a port for 2.5 microseconds (µs) and there are frames waiting to be transmitted. TxWait is reported in multiple formats for easy understanding.

Following are reported for all FC ports:

- Absolute count since last time counter was cleared.
- Percentage Tx B2B credits unavailability over last 1 second, 1 minute, 1 hour, and 72 hours.
- TxWait history graph for last 60 seconds, 1 hour, and 72 hours.
- History of TxWait is maintained for longer duration at On Board Failure Logging (OBFL) with time stamps.

Credit Unavailability at Milliseconds—slowport-monitor

Slowport-monitor displays live continuous duration for which Tx B2B credits were unavailable on a port. It can be enabled to monitor all ports on a Cisco MDS 9000 Family switch at as low as 1 ms without any performance impact. The event is logged along with the time stamp if Tx B2B credits are unavailable for continuous duration longer than a configured duration.

Slowport-monitor is implemented directly on Port ASIC. It never misses a transient condition due to minimum monitoring interval of 1 ms with granularity of 1 ms.

Note: Slowport-monitor and TxWait are new hardware-assisted features that are available on Cisco MDS 9000 Family switches. Both features are extremely powerful and should be preferred over other detection features.

Credit Unavailability at 100 ms

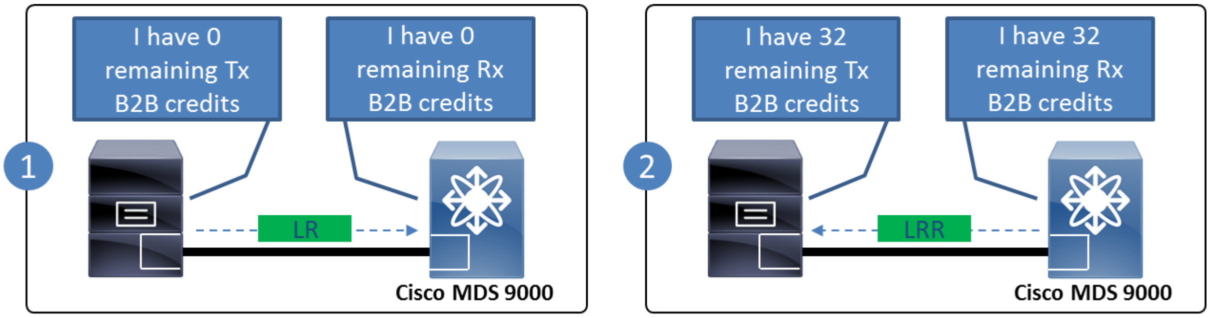
Cisco MDS 9000 Family switches increment <*component_id*>_CNTR_TX_WT_AVG_B2B_ZERO by 1 if the Tx B2B credits are at zero for 100 ms, or longer. Similar to the counter in Tx direction, <*component_id*>_CNTR_RX_WT_AVG_B2B_ZERO is incremented by 1 if remaining Rx B2B credits are at zero for 100 ms. The granularity of these counters is 100 ms (while TxWait has granularity of 2.5 µs and slowport-monitor has granularity of 1 ms). A special process in Cisco NX-OS (called *credit monitor* or *creditmon*) polls all the ports every 100 ms. Counters are incremented if B2B credits are unavailable for the whole duration between two polling intervals. The software polling mechanism may not be accurate under corner cases when control plane is heavily loaded with other high-priority tasks. Hence, slowport-monitor and TxWait should be preferred over <*component_id*>_CNTR_TX_WT_AVG_B2B_ZERO counter in the transmit direction. In the receive direction, <*component_id*>_CNTR_RX_WT_AVG_B2B_ZERO continues to be the preferred feature.

Note: TxWait, slowport-monitor, and credit unavailability at 100 ms are complementary features. All of them should be used together for best results. Detailed comparison of these features is available in Appendix B.

Link Event LR Rcvd B2B on Fibre Channel Ports

If a port stays at zero Rx B2B credits for a long duration, a link reset can be initiated by the adjacent Fibre Channel device (because it presumably has zero Tx B2B credits; see Figure 11) When this reset occurs, the Cisco MDS 9000 Family switch port receives a link reset (LR) primitive. The port checks its ingress buffers and determines whether at least one frame is still queued. If no frames are queued (that is, if all received frames have been delivered to their respective destination egress ports), then a link reset response (LRR) primitive is returned. Both the adjacent Fibre Channel device and the Cisco MDS 9000 Family switch port are now back at their full complement of B2B buffers. The link resumes its function.

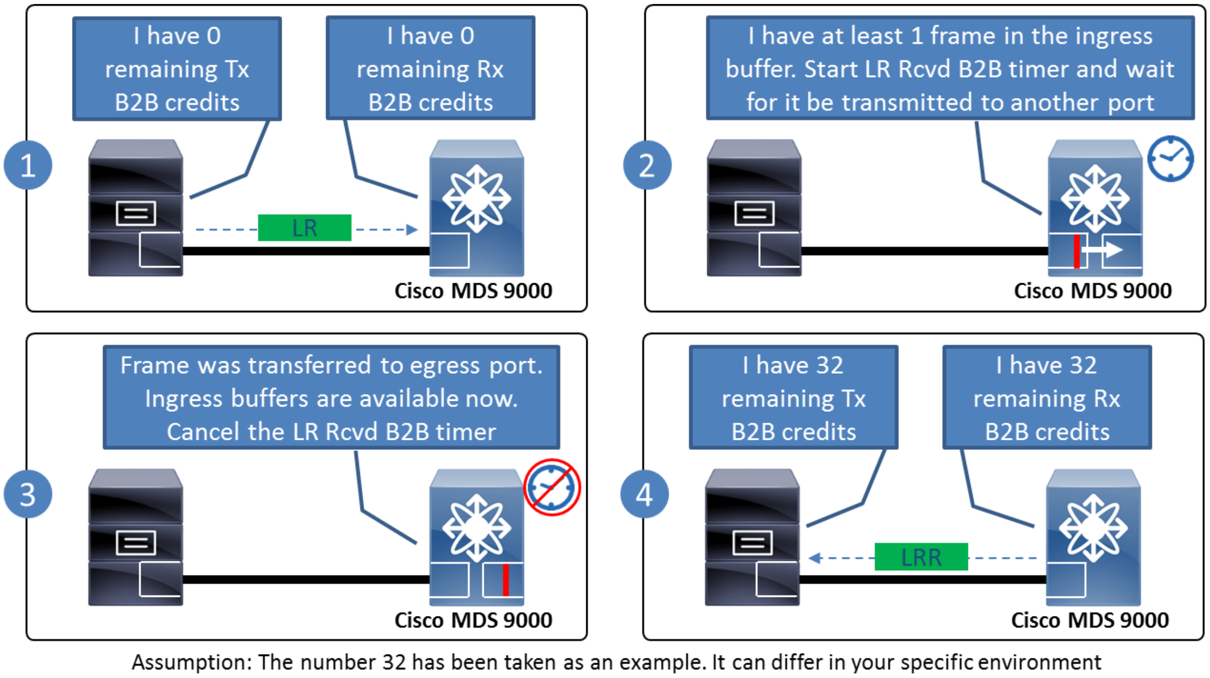
Figure 11. Exchange of LR and LRR Primitives


1
I have 0 remaining Tx B2B credits
I have 0 remaining Rx B2B credits
LR
Cisco MDS 9000
2
I have 32 remaining Tx B2B credits
I have 32 remaining Rx B2B credits
LRR
Cisco MDS 9000

Assumption: The number 32 has been taken as an example. It can differ in your specific environment

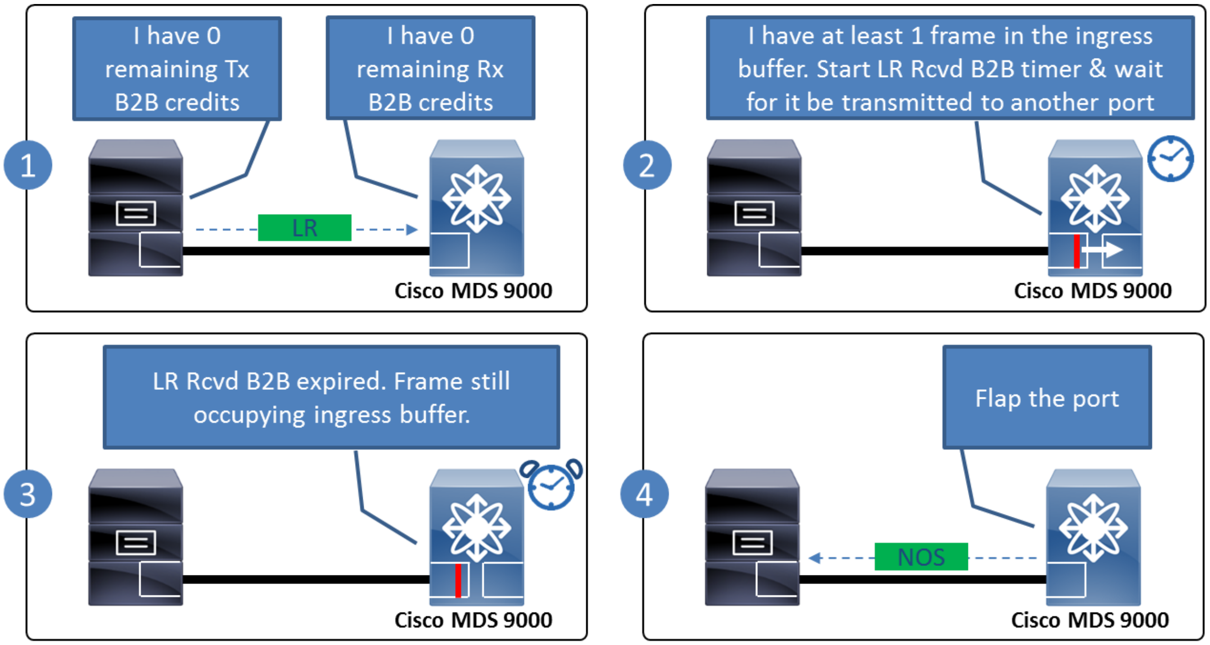
However, if at least one frame is still queued (Figure 12), the Cisco MDS 9000 Family switch starts a 90 ms LR Rcvd B2B timer. If the Fibre Channel frames can be transmitted to the egress port, then the LR Rcvd B2B timer is canceled and an LRR message is sent back to the adjacent Fibre Channel device.

Figure 12. Exchange of LR and LRR Primitives When at Least One Frame Is Still Occupying the Ingress Buffer


1
I have 0 remaining Tx B2B credits
I have 0 remaining Rx B2B credits
LR
Cisco MDS 9000
2
I have at least 1 frame in the ingress buffer. Start LR Rcvd B2B timer and wait for it be transmitted to another port
Cisco MDS 9000
3
Frame was transferred to egress port. Ingress buffers are available now. Cancel the LR Rcvd B2B timer
Cisco MDS 9000
4
I have 32 remaining Tx B2B credits
I have 32 remaining Rx B2B credits
LRR
Cisco MDS 9000
Assumption: The number 32 has been taken as an example. It can differ in your specific environment

However, if the egress port remains congested and Fibre Channel frames are still queued at the ingress port, the LR Rcvd B2B timer expires (Figure 13). No LRR is transmitted back to the adjacent Fibre Channel device, and both the ingress port and the adjacent Fibre Channel device initiate a link failure by transmitting a Not Operational Sequence (NOS) (a type of primitive sequence in Fibre Channel).

Figure 13. Indication of Upstream Congestion—Expiration of LR Rcvd B2B Timer Results in Port Flap.


1
I have 0 remaining Tx B2B credits
I have 0 remaining Rx B2B credits
LR
Cisco MDS 9000
2
I have at least 1 frame in the ingress buffer. Start LR Rcvd B2B timer & wait for it be transmitted to another port
Cisco MDS 9000
3
LR Rcvd B2B expired. Frame still occupying ingress buffer.
Cisco MDS 9000
4
Flap the port
NOS
Cisco MDS 9000

This event is logged as *LR Rcvd B2B* or *Link Reset failed nonempty recv queue*. This event indicates severe slow-drain congestion, but the cause is not the port that failed. The potential problem lies with the port to which this port is switching the frames out on the same switch. If multiple ports on a MDS switch display this behavior, then most likely all of them are switching frames to the same egress port that is facing server congestion. It can be an F_port or an E_port in a multiswitch environment.

Credits and Remaining Credits

The Rx B2B credits on a port is the number of receive buffers of the port. The Tx B2B credits on a port is the number of Rx B2B credits of the port on the other end of the link. Both of these numbers are static and do not change after the link comes up and the device logs in to the Fibre Channel fabric.

The remaining Tx and Rx B2B credits are instantaneous values. They represent the count of frames that can still be transmitted without receiving R_RDY. On a healthy port, the remaining B2B credits matches the credit count. A lower value of remaining Tx B2B credit means the connected device is not returning R_RDY fast enough or R_RDY may be lost on the return path. A lower value of Rx B2B credits means that the port is not able to return R_RDY fast enough to the port on the other end of the link. This may happen if frames cannot be switched to another port on the same switch fast enough.

The remaining credit counter provides an instantaneous value on Cisco NX-OS. The value should be monitored multiple times.

Credit Transition to Zero

Cisco MDS switches increments a hardware counter whenever remaining Tx or Rx B2B credits fall to zero. Tx B2B credit transition to zero indicates that the port on the other end of the link is not returning R_RDY. Rx B2B credit transition to zero indicates that the port is not able to return R_RDY to the device on the other end of the link. This may happen if frames cannot be switched to another port on the same switch fast enough.

Increments in this counter may be normal behavior, depending on fabric health. Very soon this counter can reach a large value. However, if the counter increments faster than usual, it may be an indication of slow drain. Finding what is faster than usual requires monitoring and benchmarking the port for a long duration. Also, the counter does not show how long the port remains at zero B2B credits. Hence, TxWait, slowport-monitor, and credit unavailability at 100-ms features should be preferred over this counter.

Defining Slow Port

A slow port refers to a port that is congested but still transmits. Though, the traffic could not be transmitted at line rate, it receives Tx B2B credits at a slow rate, that is, the receiver of the Fibre Channel frame does not immediately return an R_RDY to the sender. This could cause the Tx B2B credits to drop to zero for a short period of time. If Cisco MDS switch has frames to send, but Tx B2B credits are unavailable, then those frames would need to wait till the switch recovers some B2B credits, and thereby slowing down the rate of traffic transmission.

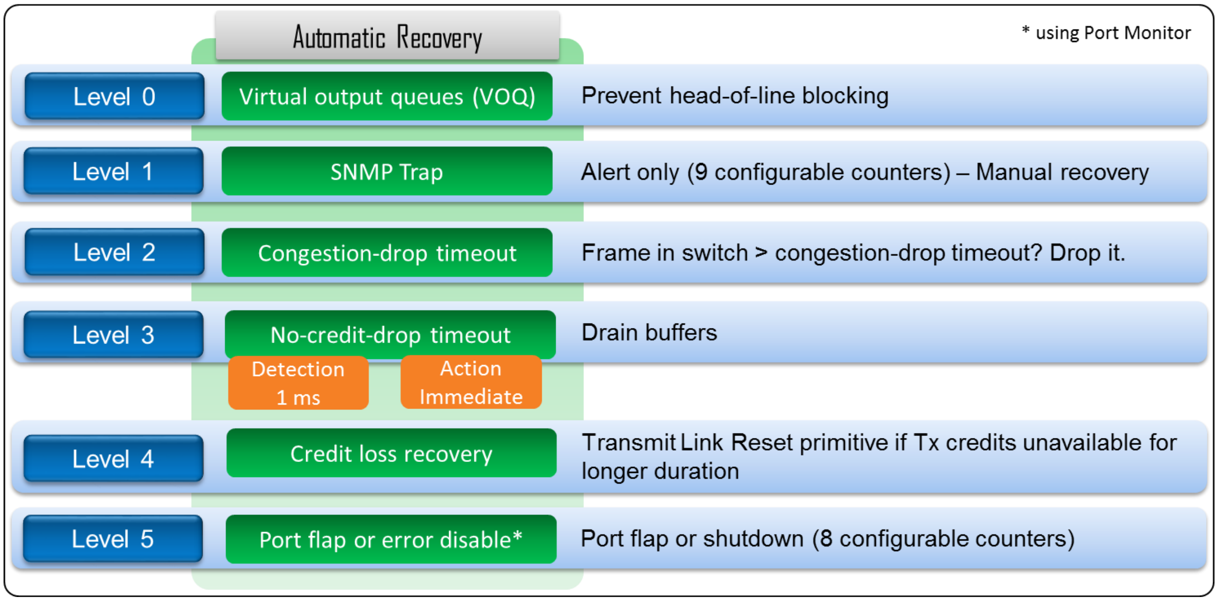
Defining Stuck Port

A stuck port refers to a port that is severely congested and unable to transmit traffic. A stuck port is subjected to prolonged B2B credit starvation. Here, a Tx B2B credit count stays at zero for a long period, blocking traffic completely and severely impacting applications.

Automatic Recovery from Slow Drain

Cisco MDS 9000 Family switches provide multiple levels of automatic recovery from slow-drain situation. Figure 14 provides summary of all available features.

Figure 14. Cisco MDS 9000 Family Switches Slow-Drain Automatic Recovery features


Automatic Recovery
* using Port Monitor
Level 0
Virtual output queues (VOQ)
Prevent head-of-line blocking
Level 1
SNMP Trap
Alert only (9 configurable counters) – Manual recovery
Level 2
Congestion-drop timeout
Frame in switch > congestion-drop timeout? Drop it.
Level 3
No-credit-drop timeout
Detection 1 ms
Action Immediate
Drain buffers
Level 4
Credit loss recovery
Transmit Link Reset primitive if Tx credits unavailable for longer duration
Level 5
Port flap or error disable*
Port flap or shutdown (8 configurable counters)

Virtual Output Queues

VOQs (see Virtual output queue bullet) prevent head-of-line blocking that recovers from microcongestion. VOQs are an intrinsic part of Cisco MDS 9000 Family switches architecture and not a feature by itself. Hence, it is considered a level 0 recovery mechanism.

SNMP Trap

Cisco MDS 9000 Family switches provide a Port Monitor feature, which monitors multiple counters at low granularity. An SNMP trap is generated if any of these counters exceeds configured thresholds over a specified duration. A Simple Network Management Protocol (SNMP) trap can trigger an external network management system (NMS) to take an action or inform a SAN administrator for manual recovery. For more details about Port Monitor, see Appendix A.

congestion-drop Timeout

A congested Fibre Channel fabric cannot deliver frames to the destination in a timely fashion. In this situation the time spent by a frame in a Cisco MDS 9000 Family switch is much longer than the usual switching latency. However, frames do not remain in a switch forever. Cisco MDS 9000 Family switches drop frames that have not been delivered to their egress ports within a congestion-drop timeout. By default, congestion-drop timeout is enabled and the value is set to 500 ms. Changing the congestion-drop timeout to a lower value can help drop frames that have been stuck in the system more quickly. This action frees up the buffers faster in the presence of a slow-drain device.

This value can be set at the switch level for port type E and F as described here:

```
MDS9700(config)# system timeout congestion-drop <value> mode (F) / (E)
MDS9700(config)# system timeout congestion-drop default mode (F) / (E)
```

Congestion-drop timeout is a switchwide recovery feature with the following attributes:

- There is no differentiation between the frames that are destined to slow devices and the frames that are destined to healthy devices but are impacted due to the congestion. Both of these frames may not be delivered to the destination timely and are subjected to congestion-drop timeout. The next level of recovery (provided by no-credit-drop timeout) drops the frames destined only to a slow-drain device.
- Dropping frames at a congestion-drop timeout is a reactive approach. The frames must be in the switch for duration longer than the configured congestion-drop timeout value. The next level of recovery (provided by no-credit-drop timeout) makes a frame-drop decision based on Tx B2B credit unavailability (which in turn leads to the frame residing in a switch for longer duration) on a port instead of waiting for a timeout.

no-credit-drop Timeout

No-credit-drop timeout is a proactive mechanism available on Cisco MDS 9000 Family switches to automatically recover from slow drain. If Tx B2B credits are continuously unavailable on a port for duration longer than the configured no-credit-drop timeout value, all frames consuming the egress buffers of the port are dropped immediately **and** all frames queued at ingress ports that are destined for the port are dropped immediately **and** while the port remains at zero Tx B2B credits, any new frames received by other ports on the switch to be transmitted out of this port are dropped.

These three actions free buffer resources more quickly than in the normal congestion-drop timeout scenarios and alleviate the problem on an ISL in the presence of a slow-drain device. Transmission of data frames resumes on the port when Tx B2B credits are available.

The efficiency of automatic recovery due to no-credit-drop timeout depends on the following factors:

- How early can it be detected that the Tx B2B credits are unavailable on a port for duration longer than no-credit-drop timeout? Only after detection, can action be taken. In other words, how soon is the action (dropping of frames) triggered after detection?
- What can be the minimum timeout value that can be detected? At higher Fibre Channel speeds, even 100 ms is a long duration. In other words, how soon can the Tx B2B credit unavailability be detected?
- What is the granularity of detection?
- How soon can it be detected that the Tx B2B credits are available on a port after a period of unavailability? This determines how soon the data traffic can be resumed on the port.

Table 3 shows details of these factors on different platforms on Cisco MDS 9000 Family switches.

Table 3. no-credit-drop Timeout Advantages on Cisco MDS 16-Gbps Platforms

Level	MDS 9500	16 Gbps platforms
How early (dropping of frames) is action triggered after detection?	Up to 99 ms	Immediate
What can be the minimum timeout value that can be detected?	100 ms	1 ms
What is the granularity of detection?	100 ms	1 ms
How early can it be detected that the Tx B2B credits are available on a port after a period of unavailability?	Up to 99 ms	Immediate

On MDS 9700, MDS 9396S, MDS 9148S, and MDS 9250i (16-Gbps platforms), no-credit-drop timeout functionality has been enhanced by special hardware capabilities on port ASICs. No-credit-drop timeout can be configured at as low as 1 ms. The timeout value can be increased up to 500 ms with granularity of 1 ms. If no-credit-drop timeout is configured, the drop action is taken immediately by port ASIC without any software delay. These advanced hardware assisted capabilities on Cisco MDS 16-Gbps platforms fully recover from slow-drain situations by pinpointing and limiting the effect only to the flows that are destined to slow-drain devices.

By default, no-credit-drop timeout is off. It can be configured at the switch level for all F_ports.

```
   switch(config)# system timeout no-credit-drop <value> mode F
   switch(config)# system timeout no-credit-drop default mode F
```

Recommended Timeout Values for congestion-drop and no-credit-drop

It is important to understand that there is no single value that is best for all situations and fabrics. It is also important to understand that a slow device can affect the fabric even when withholding B2B credits for a few milliseconds if it occurs repeatedly in the presence of large amounts of traffic.

Cisco recommends reducing the congestion-drop timeout on F_ports to 200 ms. This times out frames earlier and speeds up the freeing of buffers in the switch. Different timeout values can be set on F and E_ports. Cisco recommends to use default congestion-drop timeout value on E_ports.

Cisco recommends configuring the no-credit-drop timeout along with the congestion-drop timeout. The no-credit-drop timeout should always be lower than the congestion-drop timeout. On Cisco MDS 9500 Series Multilayer Directors (8-Gbps and advanced 8-Gbps platforms), no-credit-drop timeout of 200 ms can be configured safely. All the 16-Gbps platforms in the Cisco MDS 9000 Family (with enhanced hardware assisted functionality), no-credit-drop timeout of less than 100 ms can be configured on healthy and high-performance fabrics. Special consideration is needed to configure a balanced no-credit-drop timeout. It should not be so high that the recovery mechanism does not trigger soon enough. On the other hand, it should not be so low that frames that are legitimately delivered are dropped in the presence of some expected delay. The expected delay is the duration for which Tx B2B credits are unavailable on a port without causing a fabricwide slow drain.

On all 16-Gbps platforms in the Cisco MDS 9000 Family, slowport-monitor can be used to find an optimized balanced value for a no-credit-drop timeout. Slowport-monitor provides operational delay (abbreviated as oper delay under Cisco NX-OS show command; see Appendix C in the Slowport-monitor section for sample output) on all ports. Operational delay is the continuous duration for which a port remains at zero Tx B2B credits. A close watch on operational delay value over a period of few weeks or months can help to benchmark the Tx B2B credit unavailability duration. Benchmarking can be used to find average values along with standard deviation. Any rise in value greater than the sum of average and standard deviation is not expected and becomes a candidate for no-credit-drop timeout. For example, benchmarking of a healthy fabric over the last 3 months provides a slowport-monitor operational delay of less than 20 ms across all ports on a Cisco MDS 9700 Multilayer Director. Sometimes, a few of the ports display a delay value of 30 ms. Serious performance degradation is observed when any port has zero remaining Tx B2B credits continuously for 50 ms or more. For this particular Cisco MDS 9700 switch, no-credit-drop timeout of 50 ms can be used. Notice that the values are used as illustration to explain the process of using slowport-monitor to find a no-credit-drop timeout. The exact timeout values can differ on different fabrics.

Credit Loss Recovery

The configurable values for congestion-drop and no-credit-drop can be up to 500 ms. Fibre Channel fabrics are impacted severely if credits are unavailable for longer duration. If Tx B2B credits are not available continuously for 1 second on F_port and 1.5 second on E_port, the credit loss recovery mechanism is invoked. The credit loss recovery mechanism transmits a link reset (LR, a Fibre Channel primitive). The adjacent Fibre Channel device, after receiving a LR, is expected to respond back a link reset response (LRR, a Fibre Channel primitive). Both the adjacent Fibre Channel device and the Cisco MDS 9000 Family switches ports can now be back at their full complement of B2B buffers. Despite the name *link reset*, the link reset is really a link credit reset and does not affect or reset the port itself when successful. Notice that the exchange of a LR and an LRR on a Fibre Channel port is different from port flap. Port flap is equivalent to shutting down the port and bringing it up again.

Note: Credit loss recovery can fail if LRR is not received within 100 ms of transmitting the link reset. This leads to port flap. See the "Link Event "LR Rcvd B2B" on Fibre Channel" Ports section on page 16 for more details.

Note: Credit loss recovery is automatic and does not require any configuration by the user.

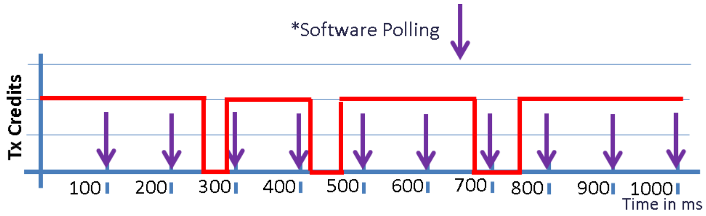
Port Flap or Error-Disable

Cisco MDS 9000 Family switches provide the Port Monitor feature, which monitors multiple counters at low granularity. Ports can be flapped if any of these counters exceeds configured thresholds over a specified duration. It is expected that flapping a port recovers the connected end device to normal condition. However, if a device or an HBA has malfunctioned permanently, it is better to disconnect it from the fabric. This can be achieved by error-disabling the switch port (same as shutting down the port) if any of these counters exceed configured thresholds over a specified duration. Error-disabled ports can be recovered by using the **shut** and **no shut** NX-OS commands after a healthy device or an HBA is connected. For more details about Port Monitor, see Appendix A.

Slow-Drain Detection and Automatic Recovery Advantage of 16-Gbps Platforms

Cisco MDS 9500 switches implement software-based slow-drain detection and recovery algorithm. The hardware port ASICS are continually polled every 100 ms to determine B2B credit unavailability. Figure 15 shows the details. The red line shows credit availability on a port plotted against time. Purple arrows show the continual software polling.

Figure 15. B2B Credit Sampling on Cisco MDS 9500 Switches

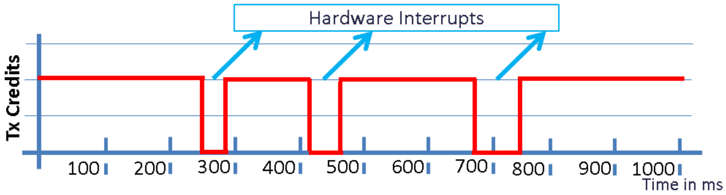


This approach gives a good snapshot of what the system is currently experiencing. This approach uses additional system resources and imposes a limit on the frequency of problem detection:

- The supervisor needs to constantly dedicate processor cycles to poll hardware.
- The supervisor needs to constantly make a decision about whether to trigger an action or recovery on the basis of a predefined policy.
- Because this feature is a snapshot mechanism, the software can miss some of the transient conditions under corner cases. In rare situations when the control plane CPU is busy, the software poll can be delayed. Due to the delay, the problematic condition may not be detected at the exact time interval or the associated action might get delayed.

The 16-Gbps platforms in the Cisco MDS 9000 Family (MDS 9700, MDS 9396S, MDS 9148S, MDS 9250i) use a hardware-based slow-drain detection and automatic recovery algorithm. In this approach, slow-drain detection and recovery is built in to the port ASIC, and instead of relying on the software for polling, the hardware can automatically detect every time a credit is not available and take appropriate action without any intervention from the supervisor (Figure 16).

Figure 16. B2B Credit Sampling on Cisco MDS 9000 16-Gbps Platforms



Here are some of the benefits of hardware-based slow-drain detection and automatic recovery algorithm:

- Detection granularity of 2.5 µs (TxWait) using port ASIC hardware counters
- New feature called Slowport-monitor, which maintains history of Tx credit unavailability duration on all ports at as low as 1 ms
- Immediate automatic recovery from slow-drain situation by port ASIC without any software delay
- Reduced load on supervisor

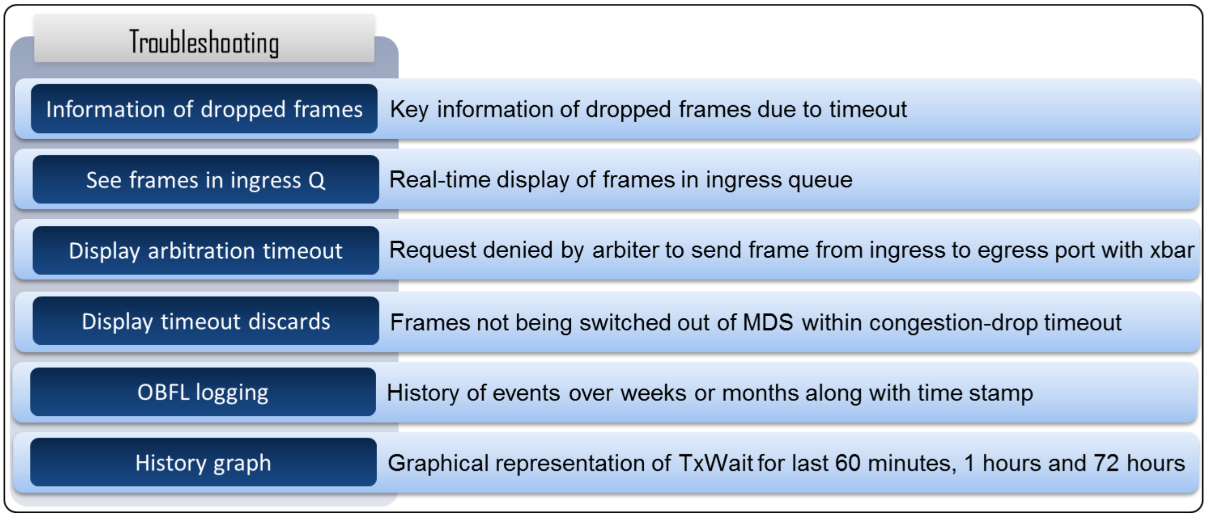
Note: Few of the enhanced features (such as Slowport-monitor) have been enabled on Cisco MDS 9500 (8 Gbps and advanced 8-Gbps platforms) from NX-OS Release 6.2(13), and later. However, the functionality is limited by the hardware capability of the MDS 9500.

Note: The software process (Creditmon) responsible for polling ports on the MDS 9500 still exists on 16-Gbps platforms. However, the process has been optimized by offloading most of the functionality to port ASICs.

Troubleshooting Slow Drain

Cisco MDS 9000 Family switches provide multiple features to troubleshoot slow-drain situations, as summarized in Figure 17.

Figure 17. Cisco MDS 9000 Family Switches Slow-Drain Troubleshooting Features



Troubleshooting	
Information of dropped frames	Key information of dropped frames due to timeout
See frames in ingress Q	Real-time display of frames in ingress queue
Display arbitration timeout	Request denied by arbiter to send frame from ingress to egress port with xbar
Display timeout discards	Frames not being switched out of MDS within congestion-drop timeout
OBFL logging	History of events over weeks or months along with time stamp
History graph	Graphical representation of TxWait for last 60 minutes, 1 hours and 72 hours

Information About Dropped Frames

A congested Fibre Channel fabric cannot deliver frames to destinations in a timely fashion. In this situation the time spent by a frame in a Cisco MDS switch is longer than the usual switching latency. However, frames do not remain within a MDS switch forever. A frame is dropped if it remains in Cisco MDS switch longer than 500 ms (default value; can be reduced further). Cisco MDS switches increment counters, as well as display key information (source FCID, destination FCID, ingress interface, egress interface, etc.) of the dropped frames. Knowing the source and destination of a frame is an extremely useful feature in slow-drain situations. Notice that a dropped frame cannot be part of a culprit flow or going to a slow-drain device. It may just be a victim. Common information between multiple dropped frames should be analyzed to indicate a slow-drain destination.

Display Frame Queued on Ingress Ports

When frames cannot be transmitted out of a port due to unavailable Tx B2B credits, they consume all the egress buffers on the port. This generates back pressure towards the ingress port on the same switch. The frames accumulate in the ingress queues of the ingress ports. These frames occupying the ingress queue of a port can be displayed in real time on Cisco MDS 9000 Family switches. All possible ingress sources must be checked to build a complete picture of traffic flows to a given destination port. A destination port index that appears occasionally in the command output likely indicates a normal device. Port indexes that appear regularly are likely to indicate slow-drain devices.

Display Arbitration Timeouts

When an ingress port needs to send a frame to an egress port, the frame is first put in a VOQ for the egress port. When the frame arrives at the head of the VOQ, a request is sent to the central arbiter to transmit the frame across the cross bar to the egress port. The frame in the VOQ is not switched to the destination port until the central arbiter grants a credit. If the egress port does not have any transmit buffers free, the arbiter does not send a grant to the ingress port. This robust mechanism ensures that frames are not subjected to congestion in a switch. The ingress port considers the request timed out after few milliseconds (the exact value depends upon the platform). The number of such request timeouts are displayed on Cisco MDS 9000 Family switches.

The egress port should be investigated because this behavior is an indication of transmission delays. These arbitration-request timeout events can be viewed on a per–egress port basis along with the ingress port numbers and a time stamp of the earliest and latest events.

Note: Arbitration timeouts are not frame drops. Arbitration requests are retried and, if the request is granted, the frame can be transmitted to the egress port successfully. If a frame never receives a grant, it will eventually be dropped at the congestion-drop timeout and be counted as a timeout discards.

Display Timeout Discards

Any frame dropped within Cisco MDS 9000 Family switch due to congestion-drop timeout or no-credit-drop timeout is accounted as timeout discard. Increment in timeout discard indicates congestion in transmit direction.

Onboard Failure Logging

Cisco MDS 9000 Family switches have an Onboard Failure Logging (OBFL) buffer that stores critical events for longer duration with time stamps. This enables deep analysis even after a particular situation gets resolved so that it can be prevented from happening again. The logs are persistent across supervisor failure and switch reboot.

Following are key OBFL sections pertaining to slow drain:

- Error-stats: contains information on timeout discards, credit loss recovery, link failures, and other errors.
- Slowport-monitor: contains slowport-monitor events.
- TxWait: contains history of TxWait.

TxWait History Graph

TxWait counter increments if Tx B2B credits are unavailable on a port for 2.5 µs. **The show interface counter** command displays the value of TxWait. To make this counter more valuable and easy to use, TxWait is provided in form of a graph for all ports on a MDS 9000 switch. This graph is displayed for the last 60 seconds, 1 hour, and 72 hours. The TxWait history graph works as a health report card of a port. If the graph is showing lower values and straight lines, the port can be considered healthy. However, higher values or spikes on the graph point to an unhealthy port state. Refer to the “TxWait” section in Appendix C for sample output.

Slow-Drain Troubleshooting Methodology

Previous sections elaborated multiple features to detect and troubleshoot slow drain. This section provides Cisco recommended methodology on the use of these tools. Without a proper methodology, it takes a very long time to pinpoint a slow-drain device, especially in large fabrics with thousands of ports.

Levels of Performance Degradation

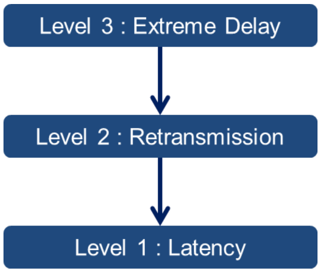
Problems can essentially be divided into three levels of service degradation (Table 4). Problems at a specific level includes all symptoms of the previous level. For example, under an extreme delay condition, the fabric might already be facing high latency and SCSI retransmission.

Table 4. Levels of Performance Degradation

Level	Host Symptoms	Default Switch Behavior
1	Latency	Frame queuing
2	SCSI retransmission	Frame dropping
3	Extreme delay	Link reset

Cisco recommended order of troubleshooting is Extreme Delay (Level 3) followed by retransmission (Level 2) followed by Latency (Level 1) as shown in Figure 18. Only after the entire extreme-delay situation is resolved, should troubleshooting be focused on retransmission. Troubleshooting a high-latency situation should be the final step.

Figure 18. Cisco Recommended Troubleshooting Order



Level 3: Extreme Delay

If a port stays at zero Tx B2B credits for a long duration (1 second for F_port and 1.5 second for E_port), credit loss recovery attempts to replenish the credits on the port by sending a link reset. If credit loss recovery is unsuccessful, the link may even flap. Both credit loss recovery and link flap introduce extreme delay in fabrics.

Level 2: Retransmission

Any frame that cannot be delivered to its destination port is held for a maximum of 500 ms (default) in an MDS switch. If that value is reached, the frame is dropped (Tx timeout discards). Frames can be dropped earlier if congestion-drop or no-credit-drop timeout are configured. Any dropped frame leads to aborting the complete Small Computer System Interface (SCSI) exchange. These aborts result in retransmissions and are listed in end-device logs.

Level 1: Latency

High latency in the fabric means SCSI exchanges are taking longer than normal. There may not be any errors or retransmission involved. High-latency situations is subtle and more difficult to detect.

Table 5 provides a quick reference of features along with NX-OS show commands that can be used to troubleshoot different levels of performance degradation. The details of the features have been explained in previous sections, and NX-OS show command details are in Appendix C. See Appendix E for more about the supported platforms and Cisco NX-OS Software versions on which these commands are available.

Table 5. Troubleshooting-Level Mapping to Slow-Drain Detection and Troubleshooting Features and Commands

Level	NX OS Features and Commands
Level 3: Extreme delay	Check for credit loss recovery with the following commands: • `show process creditmon credit-loss-events` • `show logging logfile`. The message ”Link Reset failed” is displayed.
	Check for LR Rcvd B2B with the following commands: • `slot <slot_number> show port-config internal link-events` • `show logging logfile`. The message “Link Reset failed nonempty recv queue” is displayed.
Level 2: SCSI retransmission	Check for timeout discard with the following commands: • `show interface counters` • `show logging onboard error-stats`
	Display dropped-frame information with the following commands: • `show system internal fcfwd idxmap port-to-interface` • `attach module <slot_number>` followed by `show hardware internal fcmac inst <instance_number> tmm_timeout_stat_buffer`
Level 1: Latency	Check for TxWait with the following commands: • `show interface counters` • `show process creditmon txwait-history` • `show logging onboard txwait`
	For slowport-monitor use the following commands: • `show process creditmon slowport-monitor-events` • `show logging onboard slowport-monitor-events`
	Check for credit unavailability at 100 ms with the following commands: • `show system internal snmp credit-not-available` • `show logging onboard error-stats:` Watch for <component_id>_CNTR_TX_WT_AVG_B2B_ZERO in Tx direction and <component_id>_CNTR_RX_WT_AVG_B2B_ZERO in Rx direction.
	Check for credit transition to zero with the following commands: • `show interface counters`
	Check for low remaining B2B credits with the following commands: • `show interface bbcredit`
	Display frames in ingress queues with the following commands: • `attach module <slot_number>` followed by `show hardware internal f16_que inst <instance_number> table iqm-statusmem0`
	Check for arbitration timeouts with the following commands: • `show logging onboard flow-control request-timeout`

Finding Congestion Sources

Generally, the root cause of slow-drain situations lies at one of the edge devices. The effect is spread to all other edge devices that are sharing the same switches and ISLs. In Figure 19 the host connected to MDS-2 is a slow-drain device. This results in Tx B2B credit exhaustion on the connected switch port. Cisco MDS 9000 Family switches have multiple features to detect Tx B2B credit unavailability on ports. However, in production fabric symptoms of slow drain may be visible on ports that are not directly connected to a slow-drain device. For example, the following message indicates Rx congestion on an interface. (For more details see the previous section Link Event LR Rcvd B2B on Fibre Channel Ports.)

```
%PORT-5-IF_DOWN_LINK_FAILURE: %$VSAN <>$ Interface <> is down (Link failure Link
Reset failed nonempty recv queue)
```

Cisco recommends finding the slow drain edge device which is the source of congestion for any slow drain symptom across the fabric. Figure 19 shows a troubleshooting flow chart to find a congestion source using the example of the shown topology. Port on MDS-1 is not able to receive frames if Rx B2B credits are unavailable (Rx congestion). This means that frames have already consumed the receive buffers of the port. These received frames must be sent to another port on the same switch, which might be congested already (Tx congestion), and hence, frames cannot be sent to it. The next step is to find the transmitting port. If the transmitting port is F_port, then the attached device is the slow-drain device. If the transmitting port is E_port, then continue troubleshooting the adjacent switch (MDS-2), which might be facing Rx congestion. The goal is to find an F_port facing Tx congestion.

For example, if a failed link displays a “LR Rcvd B2B” or “Link failure Link Reset failed nonempty Recv queue” message, then the port that fails is not the cause of the slow drain but is only a port that was affected. To identify the port that caused the link failure, use following steps:

1. Determine whether more than one link is failing.
2. Check the VSAN zoning database to see with which devices the adjacent Fibre Channel device is zoned. Map these to egress E_ports or F_ports. To map to egress E_ports, use the `show fspf internal route vsan <vsan> domain <dom>` command. To map to local F_ports, use the `show flogi database vsan <vsan>` command. If more than one link is failing and displays “LR Rcvd B2B,” then combine the egress E_ports or F_ports that are found and check for overlap. Overlapping ports are likely the port that caused the link failure.
3. Check the ports found in step 2 for indications of Tx B2B credit unavailability. Examples are credit loss recovery, (<*component_id*>_CNTR_CREDIT_LOSS), 100 ms Tx B2B zero (<*component_id*>_CNTR_TX_WT_AVG_B2B_ZERO), TxWait, slowport-monitor, and timeout discard (<*component_id*>_TIMEOUT_DROP_CNT).
4. If the failure port is determined to be an E_port, then continue slow-drain troubleshooting on the adjacent switch indicated by the Fabric Shortest Path First (FSPF) next-hop interface.
5. If the port is determined to be a Fibre Channel over IP (FCIP) link, then check the FCIP links for signs of TCP retransmission or other problems, such as link failures. The command `show ips stats all` can be used to check for problems.

Figure 19 provides a process flow chart to follow congestion to source with slow-drain situations.

Figure 19. Flow Chart to Follow Congestion to Source of Slow Drain

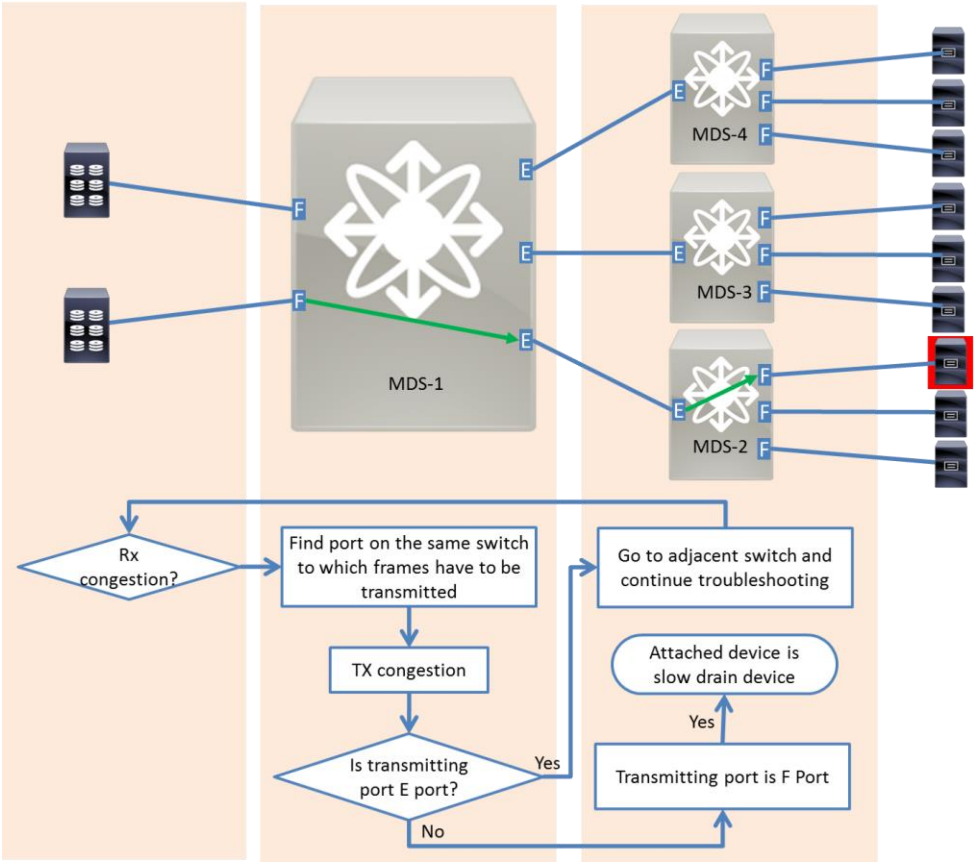

MDS-4
MDS-3
MDS-2
MDS-1
F
E
Rx congestion?
Find port on the same switch to which frames have to be transmitted
Go to adjacent switch and continue troubleshooting
TX congestion
Attached device is slow drain device
Yes
Is transmitting port E port?
Yes
Transmitting port is F Port
No

Table 6 provides a list of features that can be used to troubleshoot RX congestion and TX Congestion, and to find the Tx port from Rx port on a Cisco MDS 9000 Family switch.

Table 6. Cisco MDS 9000 Family Switches Slow-Drain Features to Troubleshoot RX or Tx Congestion

Troubleshooting Rx Congestion	Troubleshooting Tx Congestion	Linking Rx to Tx ports
LR Rcvd B2B	TxWait period for frames	VSAN zoning database
0 receive B2B credit remaining	Slowport-monitor	Information about dropped frames
Receive B2B credit transitions to zero	Credit unavailability at 100 ms	Information about frames in ingress queue
Excessive received link reset	0 transmit B2B credit remaining	Arbitration timeout
	Transmit B2B credit transitions to zero	
	Timeout discards	
	Credit loss recovery	
	Excessive transmitted link reset	

Generic Guidelines

In identifying a slow-drain device, be aware of the following:

- Logs are detailed and can roll over an active port. Though events are stored at OBFL, troubleshooting should begin quickly when slow-drain problems are detected.

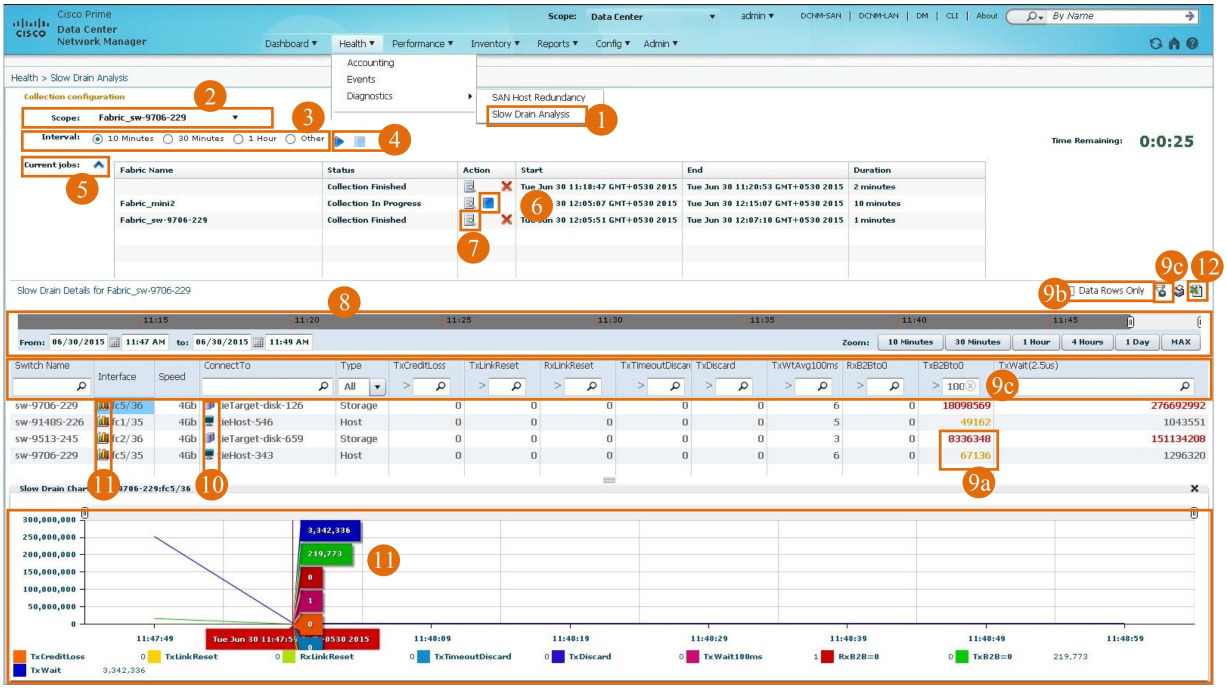
- If credit loss recovery and/or transmit frame drops occur on an ISL, then traffic destined to any egress port on the switch can be affected, and so multiple edge devices may report errors. If either condition is seen on ISLs, then the investigation should continue on the ISL peer switch. If an edge switch shows signs of frame drops, then each edge port on the edge switch should be checked.

Detecting and Troubleshooting Slow Drain with Cisco Data Center Network Manager

Cisco Data Center Network Manager (DCNM) provides slow drain diagnostics in release 7.1(1), and later. DCNM provides a fabricwide view of all switches, ISLs and end devices. DCNM can watch thousands of ports in a fabric in a slow-drain situation. The health of these ports is displayed in single pane of glass to easily identify top offenders. DCNM slow-drain diagnostics reduces troubleshooting time from days to minutes by pinpointing the problem. Often, SAN administrators struggle to find a starting point to locate a slow-drain device in large and complex topologies. Cisco recommends using DCNM slow-drain diagnostics as the very first troubleshooting step when a slow- drain device is suspected in a fabric.

Following is a description of using slow-drain diagnostics on Cisco DCNM, as shown in Figure 20.

Figure 20. Slow Drain Diagnostics on Cisco DCNM



Scope: Fabric_sw-9706-229

Interval: ● 10 Minutes ○ 30 Minutes ○ 1 Hour ○ Other

Current jobs:

Fabric Name	Status	Action	Start	End	Duration
	Collection Finished		Tue Jun 30 11:18:47 GMT+0530 2015	Tue Jun 30 11:20:53 GMT+0530 2015	2 minutes
Fabric_mini2	Collection In Progress		30 12:05:07 GMT+0530 2015	Tue Jun 30 12:15:07 GMT+0530 2015	10 minutes
Fabric_sw-9706-229	Collection Finished		Tu un 30 12:05:51 GMT+0530 2015	Tue Jun 30 12:07:10 GMT+0530 2015	1 minutes

Slow Drain Details for Fabric_sw-9706-229

From: 06/30/2015 11:47 AM to: 06/30/2015 11:49 AM

Switch Name	Interface	Speed	ConnectTo	Type	TxCreditLoss	TxLinkReset	RxLinkReset	TxTimeoutDiscard	TxDiscard	TxWtAvg100ms	RxB2Bto0	TxB2Bto0	TxWait(2.5us)
				All	>	>	>	>	>	>	>	> 100	
sw-9706-229	fc5/36	4Gb	ieTarget-disk-126	Storage	0	0	0	0	0	6	0	18098569	276692992
sw-9148S-226	fc1/35	4Gb	ieHost-546	Host	0	0	0	0	0	5	0	49162	1043551
sw-9513-245	fc2/36	4Gb	ieTarget-disk-659	Storage	0	0	0	0	0	3	0	8336348	151134208
sw-9706-229	fc5/35	4Gb	ieHost-343	Host	0	0	0	0	0	6	0	67136	1296320

🟧 TxCreditLoss	0	🟨 TxLinkReset	0	🟩 RxLinkReset	0	🟦 TxTimeoutDiscard	0	🟦 TxDiscard	0	🟪 TxWait100ms	1	🟥 RxB2B=0	0	🟩 TxB2B=0	219,773	
🟦 TxWait	3,342,336															

4. A history of previous analysis or the status of currently running analysis can be watched by choosing from the Current Jobs drop-down menu. Analysis can be run in parallel for different fabrics at the same time by using two browser sessions. Counters are polled every 10 seconds for the duration specified. The analysis can be stopped manually by clicking the stop button.

5. To display the output of the analysis while it is running, click the Show Results icon under Current jobs drop-down menu.

   The change of multiple values is displayed for the complete fabric. For detailed analysis, counters can be zoomed to 10 minutes, 30 minutes, 1 hour, 4 hours, 1 day, or the maximum duration.

6. To display core granularities, use the time slider, or select From and To time stamps.
7. Use multiple options to pinpoint slow-drain devices from thousands of ports in a large and complex fabric in minutes:

   a. The counters are color coded. Counters in red indicate drastic change, counters in orange indicate optimum change, and so on. Troubleshooting should be started on ports showing counters in red.

   b. Ports with only nonzero counter values can be filtered by selecting the Data Rows Only radio button.

   c. You can filter ports further with the filtering options available with all fields. For example, if one of the switches is suspect, it can be inspected by filtering under the Switch Name column. If one of the end-devices is suspected, it can be inspected by filtering under the ConnectTo column. Switch ports can be filtered based on their connectivity to end devices (F-port) or ISLs (E-port). F_ports can further be filtered by connection to Host or Storage. Specific counters can be filtered by providing a minimum threshold value in the text box. Only ports that have with counter values higher than the provided value are displayed.

8. To open the end-to-end topology showing the end device in the fabric, click the icon of the connected device just before the device name under the ConnectTo column.
9. To display graphical display of counters over the analyzed period, click the graph icon just before the interface name. To display the counter at that particular time stamp, position the cursor over the graph.

   Graphical representation of the counter is an extremely powerful feature that enables locating an abnormal condition, reducing false positives. Large values of many counters (such as RxB2Bto0, TxB2Bto0, or TxWait) may be acceptable in a fabric. However, unexpected sudden change in these counters can indicate a problem: for example, if a stable fabric rises 1000 in TxWait every 5 minutes (numbers are just for illustration). This change can be treated as a typical expected value. However, a problem may exist if TxWait increments in millions over the next 5-minute interval. Locating such sudden spikes becomes extremely intuitive and fast using graph.

Output of a job on a fabric can be exported to Microsoft Excel format for sharing, deep inspecting, and archiving.

As of DCNM release 7.2(1), counters listed in Table 7 can be monitored.

Table 7. Counters Monitors by DCNM Slow-Drain Diagnostics

DCNM counter name	Description	Reference section
TxCreditLoss	Number of times when remaining Tx B2B credits were zero for 1 second on F_port and 1.5 seconds on E_port. This results in credit loss recovery by transmitting link reset (Fibre Channel primitive).	Credit Loss Recovery and LR Rcvd B2B
TxLinkReset	Number of link resets transmitted on a port.	Credit Loss Recovery and LR Rcvd B2B
RxLinkReset	Number of link resets received on a port.	Credit Loss Recovery and LR Rcvd B2B
TxTimeoutDiscard	Number of frames dropped due to congestion-drop timeout and no-credit-drop timeout.	Timeout Discards
TxDiscard	Number of frames dropped in transmit direction. This includes TxTimeoutDiscard.	Timeout Discards
TxWtAvg100ms	This counter increments every 100 ms of Tx B2B credit unavailability.	Credit unavailability at 100ms
RxB2Bto0	Number of times when remaining Rx B2B credits fall to zero, even for an instant.	Credit transition to zero
TxB2Bto0	Number of times when remaining Tx B2B credits fall to zero, even for an instant.	Credit transition to zero
TxWait (2.5us)	This counter increments every 2.5 µs when remaining Tx B2B credits are zero.	TxWait period for frames

Note: DCNM slow-drain diagnostics uses SNMP object identifiers (OIDs) for analysis. The actual counters must be supported by the managed switch. See the support matrix in Appendix E for supported features across various platforms and NX-OS releases.

Note: For DCNM release 7.2(1), and later, the displayed value of a counter is the collective change value across all previous jobs on a given fabric. Counters for particular job can be seen by zooming to a particular time window.

Using DCNM to Detect Slow-Drain Devices

DCNM slow-drain diagnostics can be used to locate a slow-drain device after a situation is known, called the *reactive approach*. Slow-drain diagnostics can also be used to profile a complete fabric by benchmarking slow-drain counters on all ports. This is called the *proactive approach*.

Reactive Approach

Cisco recommends using DCNM slow-drain diagnostics as the very first troubleshooting step when a slow-drain device is suspected in a fabric. The following workflow can be used to locate a slow drain device:

- Smaller jobs at 10 minutes or 30 minutes duration should be run. If a longer task is under progress, live counters should be monitored.
- Before starting a job, consider deleting previous collections from DCNM. This deletion helps to reduce the time slider by removing old data. Exporting data in Microsoft Excel format is a good practice before deleting any data. If deleting previous collections is not desired then use the zoom functionality or time slider to monitor only the latest collected data.
- Select the Data Rows Only radio button as a very first filtering step.
- Look for counters in the same order as the levels of congestion they represent: TxCreditLoss ➔ TxLinkReset, RxLinkReset ➔ TxDiscard, TxTimeoutDiscard ➔ TxWtAvg100ms ➔ TxWait ➔ TxB2Bto0, RxB2Bto0.
- Look for counters in red. Click the Show Filter icon and enter a large number to enable filtering. A large number should be chosen so you can display only a handful of devices. If not enough devices are shown

after applying the filter or displayed ports are not suspect, consider reducing the filter value so that more ports can be displayed.

- Filter ports based on their connectivity to Host, Storage, or Switch. Ports connected to a host should be analyzed before a port is connected to storage and switches. Troubleshooting ISL ports (ports connected to switch) should be the last step.
- Display the graph for filtered ports. Watch for ports with high values and spikes.
- Display the topology and pay special attention to degraded data rate on the suspected ports.
- A port displaying low-transmit data rate and increments in slow-drain counters at high rate is a suspect port that might be connected to a slow-drain device.

Proactive Approach

Slow-drain diagnostics enables prevention of slow drain, which is a complementary functionality on top of detection, troubleshooting, and automatic recovery features described throughout this document. A well-performing fabric can face a slow-drain situation even if one end device malfunctions. The malfunctioning slow-drain end device can display patterns or interim intervals of poor performance before it starts causing severe performance degradation to an application.

To understand it better, let’s take example of an HBA with an average TxWait of 1 second over an interval of 5 minutes. This means in a window of 5 minutes, frames had to wait for 1 second, because Tx B2B credits were unavailable. The TxWait of 1 second is a cumulative value spread across 5 minutes with quanta of 2.5 µs. If the application on the host with this HBA has been performing very well over the last 3 months (or any other long duration), it is safe to assume that TxWait of 1 second in a 5-minute window is a typical expected value for the F port. This is called *slow-drain port profiling* or *benchmarking*.

If the HBA malfunctions and becomes a slow-drain device, TxWait can increase to a value where the application performance is severely impacted. The impact is seen on other hosts sharing the same switches and ISLs. Let’s assume that this increased TxWait value is 20 seconds in a window of 5 minutes. Any such spike in the counters in DCNM needs attention and should be carefully analyzed. This should be done even before the application end users complain about performance degradation. In other situations, the delay value may not rise to 20 seconds and stay there permanently. There may be an interim 5-minute window when the delay value gets to 1 second while in other windows the delay value might be higher. Any such random spike in the delay value might be a peek into the future that the HBA is about to malfunction. Fabricwide benchmarking on all F-ports enable a SAN administrator to maintain a history of acceptable delay values. Ports with random spikes in delay values above the acceptable value should be kept under a watch list. If the number of spikes in TxWait on a port is increasing, then probably the connected end device is about to malfunction completely.

This proactive approach by fabric profiling or benchmarking can be done natively on Cisco MDS 9000 Family switches using slowport-monitoring and TxWait. The centralized fabricwide visibly of DCNM makes this exercise simpler and faster.

Summary

Congestion in SANs cripples application performance instantly. Problems such as slow drain originate from one misbehaving end device but can impact all other end devices that are sharing the same pair of switches and ISLs. Following is the summary of recommendations to detect, troubleshoot, and automatically recover from slow drain on Cisco MDS 9000 16-Gbps platforms.

Always do the following:

- Configure slowport-monitor at 10–25 ms for both E and F_ports. The value can be further reduced without any side effects.
- Configure congestion-drop timeout on F_ports at 200 ms.
- Configure no-credit-drop timeout on F_port at 50 ms. If the SAN administrator find this value to be aggressive, 100 ms is a good start. Use the output of slowport-monitor to refine the value of no-credit-drop timeout.
- Configure port-monitor policies as per the thresholds specified in Appendix A.

To detect and troubleshoot:

1. Use DCNM slow-drain diagnostics.
2. Use the TxWait counter effectively by monitoring the port health by graphical display and percent of congestion available as output of Cisco NX-OS show commands.
3. Use the output of the slowport-monitor feature.
4. Follow the Cisco recommended methodology of moving towards the source of congestion.
5. Use the other features described in this document.

Last but not the least, it is highly recommended to benchmark the credit unavailability duration on all fabric ports to forecast the events.

Conclusion

Cisco MDS 9000 Family switches have been architected to build robust and self-healing SANs. All ports function at full line rate with predictable and consistent performance. The holistic approach taken by Cisco to detect, troubleshoot and automatically recover from slow drain keeps the performance of your SAN at peak. Slow drain diagnostics on Cisco DCNM reduced the troubleshooting time to minutes from weeks. Overall, Cisco MDS 9000 Family switches are the best choice for building SANs to support most critical business applications.

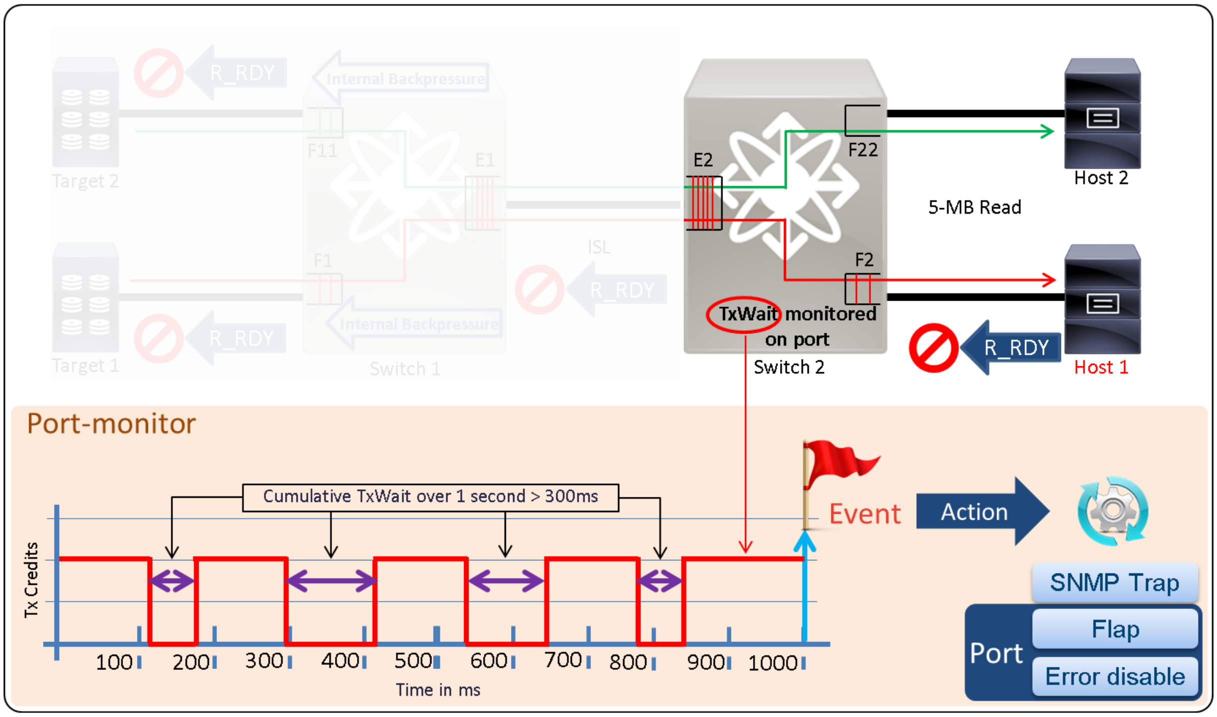
Appendix A: Slow-Drain Detection and Automatic Recovery with Port Monitor

The Port Monitor feature automates detection and recovery from slow drain. Port Monitor tracks various counters and events that are flagged if the value of the counters exceeds configured thresholds over a specified duration. In response to these events, Port Monitor triggers automated actions. Generating a SNMP trap is the default action. Port Monitor uses port guard (optional) functionality to flap or error-disable a port, as described in the following:

- **Port flap:** Leads to link down followed by link up, similar to using the NX-OS **shutdown** command followed by **no shutdown** on an interface.
- **Error-disable:** Leads to link down until a manual user intervention brings the link back up, similar to using the **shutdown** NX-OS command on an interface. User must manually execute the **no shutdown** command on the interface to bring the link back up.

Consider a port with TxWait of 100 ms over a poll interval of 1 second. This means the Tx B2B credits on the port were unavailable for 100 ms in the monitored duration of 1 second. An administrator may want to receive an automated alert or even shutdown the port if the TxWait exceeds more than 300 ms over the poll interval of 1 second. The administrator can configure a Port Monitor policy to achieve this (Figure 21).

Figure 21. Port Monitor Functionality Using TxWait on Cisco MDS 9000 Family Switches



An advantage of Port Monitor is its unique ability to monitor hardware-based counters at extremely low granular time intervals. For example, an SNMP trap can be generated at as low as 1 ms of credit unavailability duration in a span of 1 second using slowport-monitor counters under Port Monitor. The Port Monitor feature provides more than 19 different counters. However, the scope of this document is limited only to the counters that are specific to slow drain. Table 8 lists counters that apply to the slow-drain solution.

Table 8. Port Monitor Counters Applicable for Slow-Drain Detection and Automatic Recovery

Port Monitor Counter Name	Description	Sections with More Details
credit-loss-reco	Number of times when remaining Tx B2B credits were zero for 1 second on F_port and 1.5 seconds on E_port, resulting in credit loss recovery by transmitting link reset (Fibre Channel primitive)	Credit Loss Recovery and LR Rcvd B2B
lr-tx	Number of link resets transmitted on a port	Credit Loss Recovery and LR Rcvd B2B
lr-rx	Number of link resets received on a port	Credit Loss Recovery and LR Rcvd B2B
timeout-discards	Number of frames dropped due to congestion-drop timeout and no-credit-drop timeout	Timeout Discards
tx-discard	Number of frames dropped in transmit direction, including TxTimeoutDiscard	Timeout Discards
Tx-credit-not-available	Counter that increments every 100 ms when remaining Tx B2B credits are zero	Credit unavailability at 100ms
tx-wait	Counter that increments every 2.5 µs when remaining Tx B2B credits are zero.	TxWait period for frames
tx-slowport-oper-delay	Duration for which Tx B2B credits were unavailable on a port	Slowport-monitor
tx-slowport-count	Number of times for which Tx B2B credits were unavailable on a port for a duration longer than the configured admin-delay value in slowport-monitor	Slowport-monitor

These counters can also be monitored using SNMP OIDs. See Appendix D for details.

Note: Port Guard can be used only with threshold type delta.

A slow-drain Port Monitor policy can be created for access ports, trunk ports, or all ports. Only one policy can be active for each port type at a time. If the port type is all ports, then there can be only one active policy. A brief introduction about the Port Monitor configuration is provided by taking the example of tx-credit-not-available (credit unavailability at 100 ms). The configuration of other counters is similar.

Configuration Example

A Port Monitor policy can be configured at the switch level or port level (F_port, E_port, or all ports):

```
switch(config)# port-monitor name Cisco
switch(config-port-monitor)# port-type <access/all/trunk>
  access-port  Configure port-monitoring for access ports
  all          Configure port-monitoring for all ports
  trunks       Configure port-monitoring for trunk ports
switch(config-port-monitor)# counter tx-credit-not-available poll-interval <1>
<threshold type> rising-threshold <10> event <4> falling-threshold <0> event <4>
portguard errordisable
```

- **port-type:** Allows users to customize the specific policy to access or trunk ports or all ports.
- **counter:** One of the counters listed in Table 8.
- **poll-interval:** Indicates the polling interval in which slow-drain statistics are collected; measured in seconds (configured to 1 second in this example).
- **Threshold type:** Determines the method for comparing the counter with the threshold. If the type is set to **absolute**, the value at the end of the interval is compared to the threshold. If the type is set to **delta**, the change in value of the counter during the polling interval is compared to the threshold. For tx-credit-not-available, **delta** should be used.
- **rising-threshold:** Generates an alert if the counter value is lower than the **rising-threshold** value in the last polling interval and is greater than or equal to this threshold at this interval. Another alert is not generated until the counter is less than or equal to a falling threshold at the end of another polling interval
- **event:** Indicates the event number to be included when the **rising-threshold** value is reached. This event can be syslog or a SNMP trap. Counters can be assigned different event numbers to indicate different severity levels.
- **falling-threshold:** Generates an alert if the counter is higher than the **rising-threshold** value prior in a last polling interval and lower than or equal to the **falling-threshold** value at this interval.
- **portguard:** Advanced option that can be set to apply error-disable or flap the affected port.

For example, in the sample command for counter tx-credit-not-available, the poll interval is 1 second and the rising threshold is set to 10 percent (which translates to 100 ms). The rising-threshold event is triggered if Tx B2B credits are unavailable for continuous duration of 100 ms in a polling interval of 1 second. This event results in a SNMP trap, and the port is put into error disable state. It remains in that state until someone manually issues a **shut** and **no shut** command on that port.

Table 9 provides a support matrix of various slow-drain–specific counters in Port Monitor. The recommended values can be configured as a starting threshold values. Monitoring over weeks or months provides more information to further refine the thresholds.

Table 9. Port Monitor Slow-Drain Quick Reference Matrix

Counter Name	Supported Platforms	Minimum NX-OS Version	Recommended Thresholds			
			Threshold Type	Interval	Rising Threshold	Falling Threshold
credit-loss-reco	ALL	5.x.x or 6.x.x	Delta	60	1	0
lr-rx	ALL	5.x.x or 6.x.x	Delta	60	5	0
lr-tx	ALL	5.x.x or 6.x.x	Delta	60	5	0
timeout-discards	ALL	5.x.x or 6.x.x	Delta	60	50	10
tx-credit-not-available	ALL	5.x.x or 6.x.x	Delta	1	10%	0%
tx-discards	ALL	5.x.x or 6.x.x	Delta	60	50	10
slowport-count[1]	MDS 9500 only with DS-X9248-48K9, DS-X9224-96K9 or DS-X9248-96K9 line cards	6.2(13)	Delta	1	5	0
slowport-oper-delay[2]	MDS 9700, MDS 9396S, MDS 9148S, MDS 9250i and MDS 9500 only with DS-X9232-256K9 or DS-X9248-256K9 line cards	6.2(13)	Absolute	1	50 ms (for 16-Gbps platforms) 80 ms (for other platforms)	0
tx-wait	MDS 9700, MDS 9396S, MDS 9148S, MDS 9250i and MDS 9500 only with DS-X9232-256K9 or DS-X9248-256K9 line cards	6.2(13)	Delta	1	20%	0

[1]: Slowport-monitor must be enabled for this counter to work and increments only if the slowport-monitor admin delay (configured value) is less than the duration for which remaining Tx B2B credits remain zero.

[2]: Slowport-monitor must be enabled for this counter to work. Threshold is exceeded only if the slowport-monitor admin delay (configured value) is less than and the reported operation delay (oper-delay) is more than the configured rising threshold.

Following is the NX-OS configuration based on the recommended thresholds from Table 9 with an alert-only action. SAN administrators can make changes according to their requirements.

```
port-monitor name Custom_SlowDrain_AllPorts
  port-type all
  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 tx-slowport-count poll-interval 1 delta rising-threshold 5 event 4 falling-threshold 0 event 4
  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
 port-monitor activate AllPorts
```

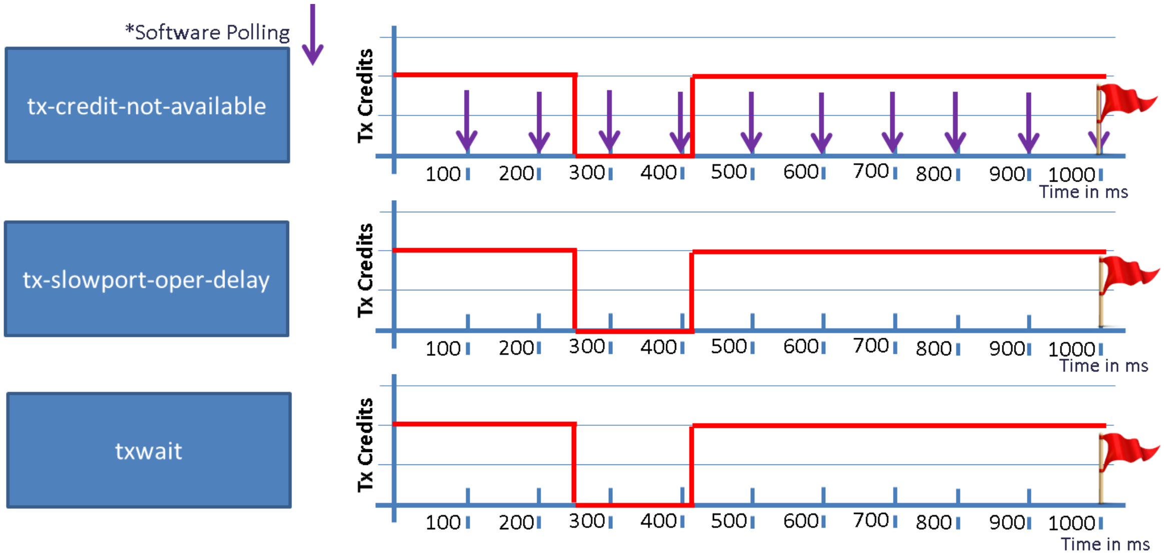
Appendix B: Difference Between TxWait, slowport-monitor, and Credit Unavailability at 100 ms

TxWait, slowport-monitor, and credit unavailability at 100 ms are complementary features. These are monitored by Port Monitor with counters tx-wait, tx-slowport-oper-delay, and tx-credit-not-available, respectively. All should be used together for best results. This appendix provides a detailed comparison.

TxWait is a hardware-based counter. It increments every 2.5 µs of Tx B2B credit unavailability on a port. Slowport-monitor provides continuous duration for which Tx B2B credits are unavailable on a port. The minimum reported duration is 1 ms. Only the durations longer than the configured admin-delay are displayed. Credit unavailability at 100 ms is a software poll–based mechanism that has been part of the Cisco MDS 9000 Family switch for many years now. Due to software poll, counters are incremented only if the credit unavailability duration on a port is more than 100 ms continuously. Also, the credits have to be zero continuously between two software polling cycles. Difference between these three counters is illustrated by taking example of 1 second poll-interval and 100 ms credit unavailability as threshold.

Consider Figure 22. The red line shows credit availability plotted against time. Purple arrows indicate software polling. Notice that software polling does not exist with tx-slowport-oper-delay and txwait. In this example, remaining Tx B2B credits fall to zero at 250 ms and do not recover until 410 ms. All three counters flag this event.

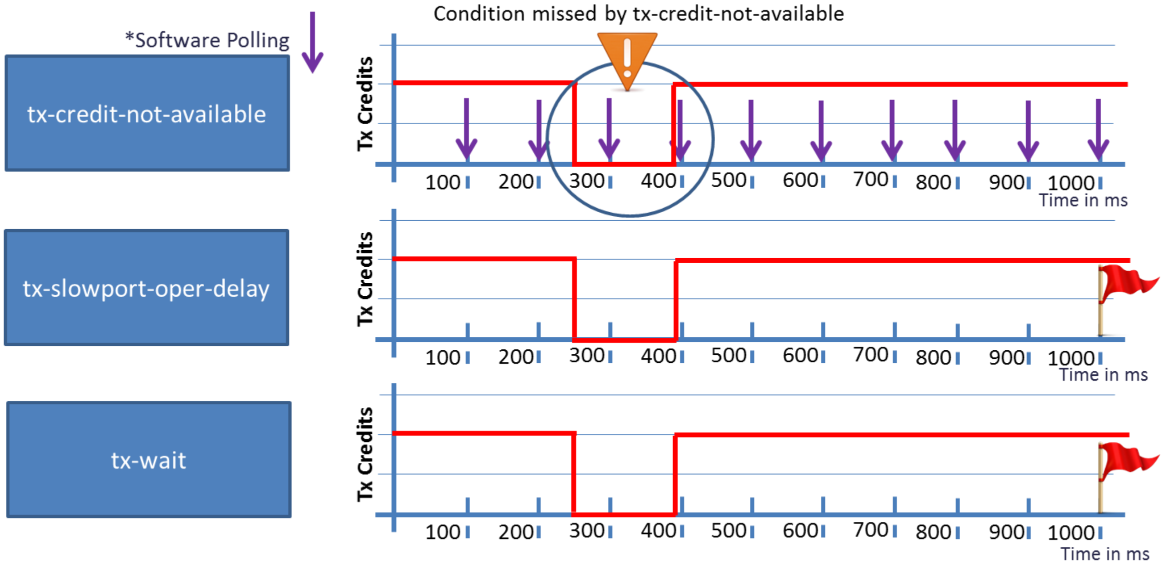
Figure 22. Comparison One of Port Monitor Counters: tx-credit-not-available, tx-slowport-oper-delay, tx-wait.



Port-monitor (Poll-interval: 1 second, threshold: 100 ms)
Event will be flagged by all three counters

Consider the credit unavailability scenario in Figure 23. Remaining Tx B2B credits fall to 0 at 250 ms and do not recover until 390ms. The overall duration is still more than 100 ms but tx-credit-not-available does not flag this event. Software polls are executed every 100 ms. There were no two consecutive polls when the remaining Tx B2B credits were zero continuously. The hardware-based implementation of tx-slowport-oper-delay and txwait helps to flag this condition.

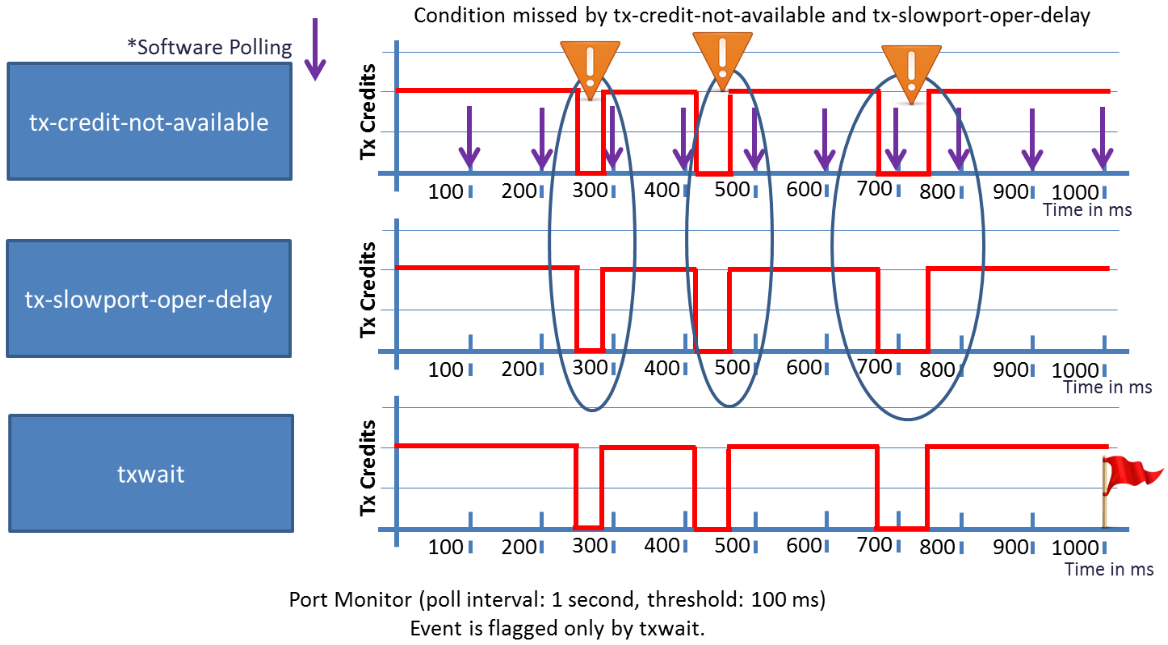
Figure 23. Comparison Two of Port Monitor Counters: tx-credit-not-available, tx-slowport-oper-delay, tx-wait.

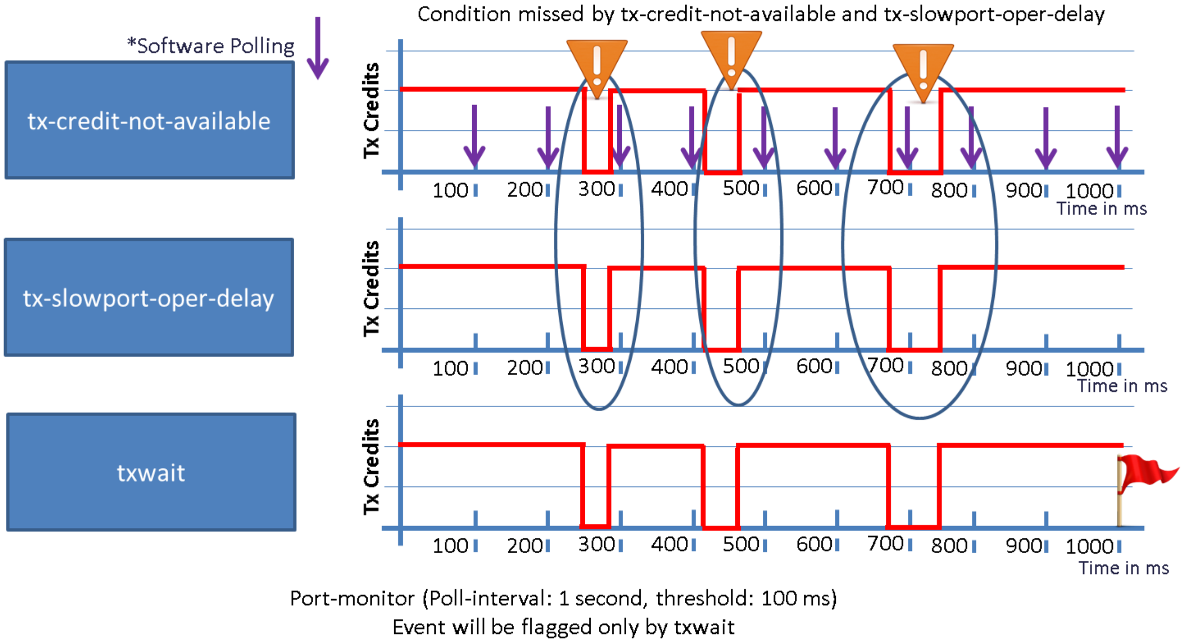


Port Monitor (poll interval: 1 second, threshold: 100 ms)
Event is flagged only by tx-slowport-oper-delay and txwait.

Consider the credit unavailability scenario in Figure 24. Remaining Tx B2B credits fall to 0 multiple times in a poll-interval of 1 second. None of the credit unavailability durations is longer than 100 ms on its own but the sum of these durations is longer than 100 ms. TxWait is the only counter that flags such condition.

Figure 24. Comparison 3 of Port Monitor counters: tx-credit-not-available, tx-slowport-oper-delay, tx-wait.


Condition missed by tx-credit-not-available and tx-slowport-oper-delay
*Software Polling
tx-credit-not-available
tx-slowport-oper-delay
txwait
Tx Credits
100
200
300
400
500
600
700
800
900
1000
Time in ms
Port Monitor (poll interval: 1 second, threshold: 100 ms)
Event is flagged only by txwait.


Condition missed by tx-credit-not-available and tx-slowport-oper-delay
*Software Polling
tx-credit-not-available
tx-slowport-oper-delay
txwait
Tx Credits
100
200
300
400
500
600
700
800
900
1000
Time in ms
Port-monitor (Poll-interval: 1 second, threshold: 100 ms)
Event will be flagged only by txwait

As shown in Figure 24, it is clear that txwait helps to find transient conditions of credit unavailability. This is an added advantage over tx-slowport-oper-delay, which finds continuous duration of credit unavailability.

Table 10 provides comparison of the three counters as by the Port Monitor feature.

Table 10. Difference Between tx-wait, tx-slowport-oper-delay and tx-credit-not-available as Monitored by Port Monitor

Attribute	Tx-wait	tx-slowport-oper-delay	tx-credit-not-available
Monitored by default	Yes	No	Yes
Supported Actions	syslog, trap, port-guard	syslog, trap	syslog, trap, port guard
Minimum supported poll interval	1 second	1 second	1 second
Threshold unit	Percentage of poll interval (40% means 400 ms in 1s)	Delay in ms	Percentage of poll interval (10% means 100ms in 1s)
Trigger if threshold delay is continuous	Yes	Yes	Yes
Trigger if threshold delay is NOT continuous but the aggregate value over poll-interval exceeds threshold	Yes	No	No
Minimum granularity	10 ms	1 ms	100 ms
Implemented in Software/Hardware	Hardware	Hardware	Software

Appendix C: Cisco MDS 9000 Family Slow-Drain detection and Troubleshooting Commands

TxWait Period for Frames

TxWait is a counter that increments if a port has zero remaining Tx B2B credit for 2.5 µs. This counter reports credit unavailability duration on a port by multiple intuitive ways.

Note: TxWait is reported only on 16-Gbps and advanced 8-Gbps platforms. The remainder of platforms in the Cisco MDS 9000 Family report zero.

Displaying the TxWait Counter

Use the show interface counter command to display TxWait counter.

```
MDS9700# show interface fc1/13 counters
<output truncated>
6252650 2.5us Txwaits due to lack of transmit credits
<output truncated>
```

TxWait can be converted to seconds using following formula:

$$\boxed{\text{TxWait value in seconds} = \frac{\text{(TxWait value in 2.5 µs ticks) x 2.5}}{1{,}000{,}000}}$$

Apply the previous formula to the TxWait value displayed in the previous output:

$$\text{TxWait value in seconds} = \frac{(6252650) \times 2.5}{1,000,000} = 15.63$$

This indicates that MDS was not able to transmit frames for more than 15 seconds since the counters were last cleared.

Low-level troubleshooting may require clearing the counters (use the **clear counter** NX-OS command) followed by displaying the TxWait value multiple times.

Displaying Percentage Tx Credits Not Available

Low-level troubleshooting has further been simplified by displaying the percentage of duration for which remaining Tx B2B credits were zero for last 1 second, 1 minute, 1 hour, and 72 hours. The percentage value is derived from TxWait counter. This information is an added advantage over the raw TxWait counter to provide a longer snapshot of the health of the port.

```
MDS9700# show interface fc1/13 counters
<output truncated>
      Percentage Tx credits not available for last 1s/1m/1h/72h: 1%/5%/3%/2%
<output truncated>
```

TxWait Logging at OBFL

History of TxWait is maintained at OBFL. To reduce the flood of information, the TxWait value is logged only if credits are unavailable for 100 ms or more during the 20-second monitoring window.

```
MDS9706-A# show logging onboard txwait module 1
Notes:
    - sampling period is 20 seconds
    - only txwait delta value >= 100 ms are logged
---------------------------------
 Module: 1 txwait count
---------------------------------
-----------------------------------------------------------------------------
| Interface | Delta TxWait Time     | Congestion | Timestamp                 |
|           | 2.5us ticks | seconds |            |                           |
-----------------------------------------------------------------------------
|   fc1/11  |  3435973    | 08      |      42%   | Sun Sep 30 05:23:05 2015 |
|   fc1/11  |  6871947    | 17      |      85%   | Sun Sep 30 05:22:25 2015 |
```

The Delta TxWait counter displays the increment in the Txwait counter over the 20-seconds window. The value is also displayed in seconds. The duration for which remaining Tx B2B credits were zero over the span of 20 seconds is displayed in percentages. For example, the first row has TxWait incremented by 3435973 over last 20 seconds on Sunday, Sep 30, 2015 at 05:23:05. 3435973 ticks at 2.5 µs translates to 8.6 seconds which is 42 percent of 20 seconds.

Note: Use the `starttime mm/dd/yy-HH:MM:SS` parameter on `show logging onboard` commands to display information pertaining only to a given time stamp.

TxWait History Graph

TxWait value on all ports is displayed in the form of a graph for better visualization of the health of a port. Graphical representation has a distinctive advantage over raw numbers due to better visualization, easier view of spikes in values, and simplified benchmarking. Use `show process creditmon txwait-history` to display the graphical representation of duration for which remaining Tx B2B credits are zero. This graph is maintained for last 60 seconds, 60 minutes, and 72 hours. For the sake of simplicity, the output of this command is similar to that of the `show processes cpu history` command in NX-OS.

```
MDS9700# show process creditmon txwait-history module 1 port 1
TxWait history for port fc1/1:
==============================
                 79998                79993            999999
                 08887                58882            9899999
     000000000000299870000000000000000029994000000000000362999500
1000              ###                  ###             ######
 900              ####                 ###             ######
 800              ####                ####             ######
 700             #####                ####             ######
 600             #####                ####             ######
 500             #####                ####             ######
 400             #####                ####             ######
 300             #####                #####            ######
 200             #####                #####            ######
 100             #####                #####            #######
    0....5....1....1....2....2....3....3....4....4....5....5....6
              0    5    0    5    0    5    0    5    0    5    0
          Credit Not Available per second (last 60 seconds)
                 # = TxWait (ms)
<output truncated>
```

This output shows the graph just for 60 seconds. A similar graph is displayed for 60 minutes and 72 hours, as well. The x-axis displays the last 60 seconds. the y-axis displays duration in milliseconds for which remaining Tx B2B credits are zero on the port. The output should be read vertically from top to bottom. The top three rows show actual duration in milliseconds. The middle 10 rows show the graph being plotted by the pound or number symbol (#). The bottom two rows show a timeline. For example, the remaining Tx B2B credits were zero for 989 ms at the 15th second and 752 ms at the 35th second (shown in bold).

slowport-monitor

Slowport-monitor can be enabled by the *system* `timeout slowport-monitor` config level command.

```
MDS9700(config)# system timeout slowport-monitor ?
  <1-500>  Configure number of milliseconds
  default  timeout value for HW slowport monitoring

MDS9700(config)# system timeout slowport-monitor default ?
  mode  Enter the port mode

MDS9700(config)# system timeout slowport-monitor default mode ?
  E  mode
  F  mode
```

After this command is enabled, events are logged whenever a port has zero remaining Tx B2B credits for duration longer than the configured value. The configured value is called *admin delay*. The default admin delay is 50 ms. The command allows admin delay values from 1 to 500 ms with 1 ms granularity. These values help to capture slow-drain devices with enhanced precision. In addition, separate admin delay values can be configured for F and E_ports. There is no extra control place CPU overheard due to slowport-monitor. Hence, the admin delay value of as low as 1 ms can be configured safely. However, this may result in a flood of information depending on the health of the fabric. A good starting admin delay value can be 10 ms. If many ports are showing delay values higher than 10 ms, then troubleshooting should be performed on those ports. If few ports show delay values higher than 10 ms, then admin delay value can be reduced further.

Note: The admin delay value in slowport-monitor must always be less than no-credit-drop timeout value. If no-credit-drop timeout is 1 ms, the slowport-monitor admin delay value should also be 1 ms. If no-credit-drop timeout is 50 ms then the slowport-monitor admin delay can be 50 ms or less.

show process creditmon slowport-monitor-events

Use `show process creditmon slowport-monitor-events` to display the last 10 events per port.

```
MDS9700# show process creditmon slowport-monitor-events

        Module: 01      Slowport Detected: YES
=========================================================================
 Interface = fc1/13
----------------------------------------------------------------
| admin  | slowport  | oper  |            Timestamp            |
| delay  | detection | delay |                                 |
| (ms)   | count     | (ms)  |                                 |
----------------------------------------------------------------
|   5    |      1300 |   20  | 1. 04/01/15 23:03:38.823        |
|   5    |      1296 |   19  | 2. 04/01/15 23:03:38.724        |
|   5    |      1291 |   19  | 3. 04/01/15 23:03:38.623        |
----------------------------------------------------------------
```

Admin delay is the value configured by the `system timeout slowport-monitor` command. Though the events are captured by the port ASIC at as low as 1 ms, but a window of 100 ms is used to display the output. The slowport detection count is the number of times when remaining Tx B2B credits are zero for a duration longer than the configured admin delay value. It is an absolute counter. Increments in the last 100 ms can be obtained by subtracting the value displayed in the row below. The operation delay (oper delay) is the actual duration for which remaining Tx B2B credits on the port are zero at the displayed time stamp. If there was more than one event in the 100 ms period, then it is the average of all events. In the previous example, at time 23:03:38.823, Tx B2B credits were unavailable continuously for 20 ms. This event occurred 4 times (1300–1296). So there was a total of 80 ms (approx. 20 ms x 4) of time in the previous 100 ms interval when credits were not available.

Difference Between 16-Gbps and Advanced 8-Gbps Platforms

Note: Table 1 provides list of 16-Gbps and advanced 8-Gbps platforms.

On 16-Gbps platforms, oper delay is displayed only if the remaining Tx B2B credits are zero for continuous span of duration that is longer than the admin delay value in the 100 ms window. On advanced 8-Gbps platforms, oper delay is the cumulative duration in the 100 ms window when remaining Tx B2B credits are zero. The difference is the continuity of the duration: for example, if the admin delay is configured to be 5 ms and remaining Tx B2B credits are zero for 4 ms two times in a window of 100 ms. 16-Gbps platforms do not display this event in the oper delay value. Advanced 8-Gbps platforms display an oper delay of 8 ms. Due to this difference in capability, the following applies to the output of `show process creditmon slowport-monitor-events` on advanced 8-Gbps platforms:

- The slowport-monitor detection count increments by only 1 in a window of 100 ms. On 16-Gbps platforms, the count can increment by more than 1.
- Txwait oper delay has been used in place of oper delay under the output of the show command. This is to signify the true meaning of the delay value, which actually is the cumulative duration.

Difference Between 16-Gbps and 8-Gbps Platforms

Note: Table 1 provides list of 16-Gbps and advanced 8-Gbps platforms.

8-Gbps platforms have basic slowport-monitor capability. In a span of 100 ms, it can only be determined if the remaining Tx B2B credits are zero for a duration longer than the configured admin delay value. The exact duration and number of times of this event cannot be determined. Following is the output of `show process creditmon slowport-monitor-events` on 8-Gbps platforms:

```
MDS9500# show process creditmon slowport-monitor-events module 2

        Module: 02      Slowport Detected: YES
=========================================================================
 Interface = fc2/1
--------------------------------------------------------
| admin  | slowport  |            Timestamp             |
| delay  | detection |                                  |
| (ms)   | count     |                                  |
--------------------------------------------------------
```

```
|  10    |       194 | 1. 04/29/15 17:19:19.345       |
|  10    |       193 | 2. 04/29/15 17:19:17.254       |
|  10    |       192 | 3. 04/29/15 17:19:16.514       |
|  10    |       191 | 4. 04/29/15 17:19:13.045       |
|  10    |       190 | 5. 04/29/15 17:19:12.945       |
|  10    |       189 | 6. 04/29/15 17:19:12.845       |
|  10    |       188 | 7. 04/29/15 17:19:11.755       |
|  10    |       187 | 8. 04/29/15 17:19:10.843       |
|  10    |       186 | 9. 04/29/15 17:19:10.550       |
|  10    |       185 |10. 04/29/15 17:19:10.145       |
--------------------------------------------------------
```

Notice that oper delay column is not displayed. Also, the slowport detection count can increment only by 1 in the 100 ms time stamp window.

OBFL slowport-monitor-events
Only the last 10 events are displayed under `show process creditmon slowport-monitor-events`. More events are captured at OBFL and can be displayed using **show logging onboard slowport-monitor-events**.

```
MDS9700# show logging onboard slowport-monitor-events
---------------------------------
 Module: 1 slowport-monitor-events
---------------------------------
--------------------------------------------------------------------------
| admin  | slowport  | oper  |            Timestamp         | Interface
| delay  | detection | delay |                              |
| (ms)   | count     | (ms)  |                              |
--------------------------------------------------------------------------
|  20    |         49 |  489  | 05/11/15 21:04:46.779        | fc1/13
|  20    |         48 |  489  | 05/11/15 21:04:46.272        | fc1/13
|  20    |         47 |  489  | 05/11/15 21:04:45.779        | fc1/13
|  20    |         46 |  489  | 05/11/15 21:04:45.272        | fc1/13
```

Credit Unavailability at 100 ms

Show System Internal SNMP credit-not-available

Use the `show system internal snmp credit-not-available` NX-OS command to display the time stamps when remaining Tx B2B credits on a port that were zero for 100 ms.

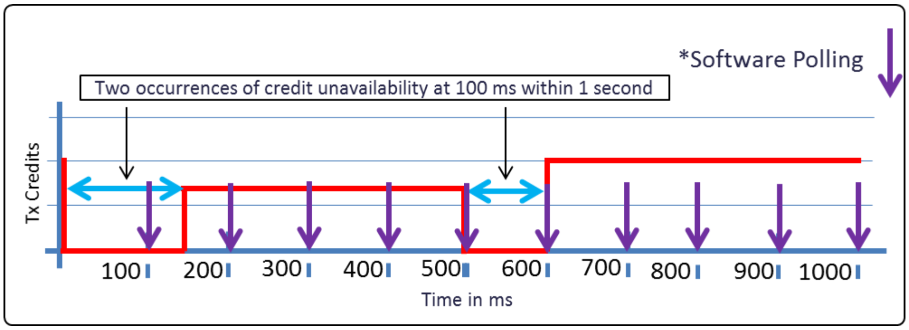
```
MDS9700# show system internal snmp credit-not-available

 Module: 1      Number of events logged: 50
-------------------------------------------------------------------------------------------
Port     Threshold         Interval(s)    Event Time                Type     Duration of
         Rising/Falling                                                      time not
                                                                             available
-------------------------------------------------------------------------------------------
fc1/1    10/0(%)            1            Tue May 19 16:12:52 2015  Falling   0%
fc1/1    10/0(%)            1            Tue May 19 16:13:19 2015  Rising    10%
fc1/1    10/0(%)            1            Tue May 19 16:13:23 2015  Falling   0%
fc1/1    10/0(%)            1            Tue May 19 16:14:06 2015  Rising    10%
fc1/1    10/0(%)            1            Tue May 19 16:14:08 2015  Falling   0%
fc1/1    10/0(%)            1            Tue May 19 16:14:33 2015  Rising    10%
fc1/1    10/0(%)            1            Tue May 19 16:14:34 2015  Falling   0%
fc1/1    10/0(%)            1            Tue May 19 16:15:12 2015  Rising    20%
fc1/1    10/0(%)            1            Tue May 19 16:15:14 2015  Falling   0%
fc1/1    10/0(%)            1            Tue May 19 16:15:42 2015  Rising    10%
```

The output is displayed in the percentage of a 1-second interval when the Tx B2B credits are not available at the time stamp. In the previous output, Tx B2B credits were unavailable for 100 ms (which is 10 percent of 1 second) on Tuesday, May 19, 2015 at 16:13:19. Similarly, Tx B2B credits were unavailable for 200 ms (which is 20 percent of 1 second) on Tuesday May 19, 2015 at 16:15:12.

The interval(s) column displays the number of times the threshold of 10 percent or 0 percent was crossed during the 1 second interval. Consider the scenario in Figure 25. The red line shows Tx B2B credit availability on a port plotted against time. The port is observing variable delay in receiving R_RDY. At the start of a second, the remaining Tx B2B credits on a port fall to zero. At the 150th millisecond, R_RDY is returned, resulting in nonzero remaining Tx B2B credits. At the 500th millisecond, the remaining Tx B2B credits, again, fall to zero. At the 600th millisecond, R_RDY is returned, resulting in nonzero remaining Tx B2B credits. This sequence of events displays 2 under the Interval(s) counter for that particular second.

Figure 25. Illustration of Credit Unavailability at 100 ms



This command is available on all Cisco MDS 9000 Family switches.

OBFL error-stats

Use `show logging onboard error-stats` to list the ports with zero remaining Tx B2B credits for 100 ms.

```
MDS9700# show logging onboard error-stats
--------------------------------------------------------------------------------
 ERROR STATISTICS INFORMATION FOR DEVICE DEVICE: FCMAC
--------------------------------------------------------------------------------
    Interface      |                                |         |    Time Stamp
      Range        |    Error Stat Counter Name     |  Count  |MM/DD/YY HH:MM:SS
                   |                                |         |
--------------------------------------------------------------------------------
fc1/13             |F16_TMM_TOLB_TIMEOUT_DROP_CNT   |1496855  |04/07/15 22:44:23
fc1/13             |FCP_SW_CNTR_TX_WT_AVG_B2B_ZERO  |217      |04/07/15 22:44:23
fc1/13             |FCP_SW_CNTR_CREDIT_LOSS         |19       |04/07/15 22:44:23
fc1/13             |F16_TMM_TOLB_TIMEOUT_DROP_CNT   |1486654  |04/07/15 22:44:03
fc1/13             |FCP_SW_CNTR_TX_WT_AVG_B2B_ZERO  |108      |04/07/15 22:44:03
fc1/13             |FCP_SW_CNTR_CREDIT_LOSS         |9        |04/07/15 22:44:03
```

Watch for counter names <*component_id*>_CNTR_TX_WT_AVG_B2B_ZERO in Tx direction and <*component_id*>_CNTR_RX_WT_AVG_B2B_ZERO in Rx direction. The count column displays an absolute value of the counter. In the previous output, the value of FCP_SW_CNTR_TX_WT_AVG_B2B_ZERO incremented from 108 to 217 between 22:44:03 and 22:44:23 on April 7, 2015. This means in the displayed 20-second interval, the remaining Tx B2B credits on fc1/13 were zero for 100 ms 109 times. This is a superior command due to display of multiple counters along with time stamps.

This command is available on all Cisco MDS 9000 Family switches, although the counter names displayed vary from platform to platform. See Appendix F for counter names and descriptions for Cisco MDS 9000 Family switches and line cards.

LR Rcvd B2B

Show Logging Log File

Display the logs generated and watch for the message, “Link Reset failed nonempty recv queue.”

```
MDS9710-1# show logging logfile
%PORT-2-IF_DOWN_LINK_FAILURE: %$VSAN 100%$ Interface fc5/32 is down (Link
failure)
%PORT-5-IF_DOWN_LINK_FAILURE: %$VSAN 100%$ Interface fc5/32 is down (Link failure
Link Reset failed nonempty recv queue)
```

This command is available on all Cisco MDS 9000 Family switches.

Command: show port-config internal link-events

The `show port-config internal link-events` command is available after attaching to a module. Watch for LR Rcvd B2B events.

```
MDS9700# attach module 1
Attaching to module 1 ...
To exit type 'exit', to abort type '$.'

Wind River Linux glibc_small (standard) 3.0

module-1# show port-config internal link-events
*************** Port Config Link Events Log ***************
----                            ------        -----  -----  ------
Time                            PortNo        Speed  Event  Reason
----                            ------        -----  -----  ------
...
Jul 28 00:46:39 2012  00670297  fc1/25        ---   DOWN   LR Rcvd B2B
```

This command is available on all Cisco MDS 9000 Family switches.

Credits and Remaining Credits

Use the `show interface bbcredit` NX-OS command to display credit and the remaining credit count. The command should be executed multiple times to get a better picture of remaining credit count.

```
MDS9710# show interface fc1/1 bbcredit
fc1/1 is up
    Transmit B2B Credit is 32
    Receive B2B Credit is 32
    Receive B2B Credit performance buffers is 0
      30 receive B2B credit remaining
      21 transmit B2B credit remaining
      21 low priority transmit B2B credit remaining
```

This command is available on all Cisco MDS 9000 Family switches.

Credit Transition to Zero

Use the `show interface counters` NX-OS command to display Tx and Rx B2B credit transition to zero.

```
MDS9710# show interface fc1/1 counters
<Output truncated>
     33 Transmit B2B credit transitions to zero
     394351077 Receive B2B credit transitions to zero
<Output truncated>
```

The actual `credit transitions...` text in the `show interface <> counters` command has changed several times, but it all indicates credit transitions to zero. Few releases of NX-OS use `credit transitions from zero`. The counters actually increment when the credits drop to zero.

This command is available on all Cisco MDS 9000 Family switches.

Dropped-Frame Information

The last 32 frames dropped by MDS 9700 and MDS 9396S can be displayed by the `show hardware internal fcmac inst <#> tmm_timeout_stat_buffer` command. This is a module-level command which can be executed by attaching to that particular module with the `attach module` NX-OS exec command.

```
MDS9700# attach module 1
Attaching to module 1 ...
To exit type 'exit', to abort type '$.'

Wind River Linux glibc_small (standard) 3.0

module-1#
module-1# show hardware internal fcmac inst 0 tmm_timeout_stat_buffer

Port Group num: 0 TMM TIMEOUT BUFFERS
---------------------------------------------
TO_RD:22 TO_WR:6 NUM PKTS:32
--------------------------------------------------------------
TMM TIMEOUT Packet :0
CHIPTIME :14227(0x3793)        ZERO:0        FCTYPE:0
SID:330040        DID:170040        RCTL:0
TSTMP_VALID:1 HDRTSTMP:14176(0x3760)        HDRCTL:6144        SI:12
DI:2        AT:0        PORTNUM:1

TMM TIMEOUT Packet :1
CHIPTIME :14227(0x3793)        ZERO:0        FCTYPE:0
SID:330040        DID:170040        RCTL:0
TSTMP_VALID:1 HDRTSTMP:14176(0x3760)        HDRCTL:6144        SI:12
DI:2        AT:0        PORTNUM:1
<output truncated>
```

```
module-1# exit
rlogin: connection closed.
MDS9700# show system internal fcfwd idxmap port-to-interface
Port to Interface Table:(All values in hex)
-------------------------------------------------------------------------------
 glob|                          |VL|lcl| if |slot|port| mts | port| flags
 idx |         if_index         |  |idx|type|    |    | node| mode|
-----|--------------------------|--|---|----|----|----|-----|-----|-------------
    0| 01000000 fc1/1           | 0| 00| 01 | 00 | 00 | 0102| 08  | 00
    1| 01001000 fc1/2           | 0| 01| 01 | 00 | 01 | 0102| 00  | 00
    2| 01002000 fc1/3           | 0| 02| 01 | 00 | 02 | 0102| 00  | 00
<output truncated>
   12| 01012000 fc1/13          | 0| 12| 01 | 00 | 12 | 0102| 00  | 00
```

The first command (`show hardware internal fcmac inst <#> tmm_timeout_stat_buffer`) shows the last 32 dropped frames starting from 0. SID and DID fields show the source FCID and destination FCID, respectively, of the frame. *SI* and *DI* stand for *Source Index* and *Destination Index*, respectively. SI and DI represent the ingress and the egress port on the switch. The mapping of SI and DI can be obtained by the `show system internal fcfwd idxmap port-to-interface` NX-OS exec level command. In the previous output, SI of 12 represents fc1/13 while DI of 2 represents fc1/3.

These values help to trace the end-to-end path of a frame. The frame might be destined to a slow-drain device or it might just be a victim. The information about the dropped frames should be collected multiple times. Common DID may represent a slow-drain device.

Display Frames in Ingress Queue

The following command output shows egress queues for all ingress ports on a given port ASIC. Upstream congestion is indicated when this output indicates that ISLs are the ingress ports. If the two outputs correlate, then a delay occurs for any downstream device. Any ports that appear regularly in the output shown here should be investigated as the cause of congestion in the fabric. Notice that this information is only an instantaneous snapshot of the queue; the command should be repeated to see whether the queue is moving.

For Cisco MDS 9500 Series DS-X92xx-256K9 modules, use the following:

```
MDS9500# slot <x> show hardware internal que inst <#> memory iqm-statusmem <0/1>
+--------------------------------------------------------------------------------
| IQM: PG0 Status Memory for Tbird Que Driver
| Inst 1; port(s) 9-16
|
Note: Only non-zero entries are displayed
Each non-zero bit indicates pending frame in VOQ for that IB
+----------+--------+--------+--------+--------+
| GI (Hex) | Prio 0 | Prio 1 | Prio 2 | Prio 3 |
+----------+--------+--------+--------+--------+
|        c | 000000 | 000000 | 000000 | 000001 |
+----------+--------+--------+--------+--------+
         ^                                    ^
         |                                    |
egress port (slow)                       ingress port
```

This command-line interface (CLI) captures the mapping of the destination index to the global index (DI to GI mapping), revealing where the frames are destined if they are queued for an extended time. This mapping indicates that one, or more, frames are queued to destination index 0xC, which is associated with port fc4/13. This mapping can be determined using the command shown here:

```
MDS9500# show system internal fcfwd idxmap port-to-interface
Port to Interface Table:(All values in hex)
--------------------------------------------------------------------------------
 glob|                          |VL|lcl| if |slot|port| mts | port| flags
 idx |         if_index         |  |idx|type|    |    | node| mode|
-----|--------------------------|--|---|----|----|----|-----|-----|-------------
…
    b| 0100b000 fc1/12          | 0| 0b| 01 | 00 | 0b | 0102| 00  | 00
  194| 01613000 fc13/20         | 0| 13| 01 | 0c | 13 | 0d02| 00  | 00
    c| 0100c000 fc4/13          | 0| 0c| 01 | 00 | 0c | 0102| 00  | 00
  195| 01614000 fc13/21         | 0| 14| 01 | 0c | 14 | 0d02| 00  | 00
```

Also, this mapping can be verified with this command:

```
MDS9500# show port internal info interface fc4/13
fc4/13 - if_index: 0x0118C000, phy_port_index: 0xc
     local_index: 0xc
```

Ingress ports that have one, or more, queued frames are indicated by a 24-bit hexadecimal map (3 bytes). In this example, this is 000001 (hexadecimal). This points to the first port in ASIC instance 1. Instance 1 is for ports 9 through 16 as shown by **Inst 1; port(s) 9-16**. This is specific to this MDS 9500 Series DS-X9232-256K9 module. Other module types have different layouts. Notice that because the instance has only 8 ports, only the rightmost byte (8 bits) is applicable. Hence, 000001 is really 01, which, when broken down to bits, is 00000001. Each 1 bit represents a port that has one, or more, queued frames.

Because, in this example, the first bit is on, it indicates that the first interface is instance 1. That works out to interface fc4/9 where the frames are ingressing. In this example they are destined to fc4/13 (GI - Global Index 0xc). For a Cisco MDS 9500 Series DS-X9248-256K9 module, each ASIC instance handles 12 ports, so instance 1 starts with port 13.

Similar information can be obtained on other modules:

- Cisco MDS 9710 DS-X9448-768K9 module:
  `show hardware internal f16_que inst 1 table iqm-statusmem0`
- Cisco MDS 9500 Series DS-X92xx-96K9 and DS-X92xx-48K9 modules:
  `show hardware internal up-xbar 0 queued-packet-info`

Note: The information displayed is real-time data, not historical data. Consequently, it should be done while the slow-drain event is occurring.

Arbitration Timeouts

Use show logging onboard flow-control request-timeout to display arbitration timeouts.

```
module# show logging onboard flow-control request-timeout
----------------------------
    Module: 1
----------------------------
----------------------------------------------------------------------------------
|  Dest  |    Source    |Events|      Timestamp         |               Timestamp |
|  Intf  |     Intf     | Count|        Latest          |                Earliest |
---------------------------------------------------------------------------------
|fc3/1   |fc1/10,      |     8|Tue Jun 12 22:32:11 2012|Tue Jun 12 22:32:12 2012 |
|        |fc1/11,      |      |                        |                         |
|        |fc1/12,      |      |                        |                         |
|        |fc1/22,      |      |                        |                         |
|        |fc1/23,      |      |                        |                         |
----------------------------------------------------------------------------------
```

This command is available on all Cisco MDS 9000 Family switches.

Check for Transmit Frame Drops (Timeout Discard)

Show Interface Counter

Use the `show interface counters` command to display frames that have been discarded inside the Cisco MDS 9000 switch due to a timeout. The frames can be timed out due to congestion-drop or no-credit-drop timeout features.

```
MDS9700# show interface fc1/1 counters
<output truncated>
    162586865166 timeout discards, 0 credit loss
<output truncated>
```

Repeat the command to see if the timeout discards counter is incrementing.

This command is available on all Cisco MDS 9000 Family switches.

OBFL error-stats

Timeout discards are reported under the `show interface <> counters` command only through a counter with no indication as to when these events occurred. Use `show logging onboard error-stats` to get time stamps along with drop count.

```
MDS9700# show logging onboard error-stats
--------------------------------------------------------------------------------
 ERROR STATISTICS INFORMATION FOR DEVICE: FCMAC
--------------------------------------------------------------------------------
    Interface      |                                |         |    Time Stamp
      Range        |    Error Stat Counter Name     |  Count  |MM/DD/YY HH:MM:SS
                   |                                |         |
--------------------------------------------------------------------------------
fc1/13             |F16_TMM_TOLB_TIMEOUT_DROP_CNT   |1496855  |04/07/15 22:44:23
fc1/13             |FCP_SW_CNTR_TX_WT_AVG_B2B_ZERO  |217      |04/07/15 22:44:23
fc1/13             |FCP_SW_CNTR_CREDIT_LOSS         |19       |04/07/15 22:44:23
fc1/13             |F16_TMM_TOLB_TIMEOUT_DROP_CNT   |1486654  |04/07/15 22:44:03
fc1/13             |FCP_SW_CNTR_TX_WT_AVG_B2B_ZERO  |108      |04/07/15 22:44:03
fc1/13             |FCP_SW_CNTR_CREDIT_LOSS         |9        |04/07/15 22:44:03
```

Watch for counter <*component_id*>_TIMEOUT_DROP_CNT. In the previous output the <*component_id*>_TIMEOUT_DROP_CNT counter incremented from 1486654 to 1496855 in the 20-second window (between 22:44:03 and 22:44:23 on April 7, 2015).

This command is available on all Cisco MDS 9000 Family switches, although the counter names displayed vary from platform to platform. See Appendix F for counter names and descriptions for Cisco MDS 9000 Family switches and line cards.

Credit Loss Recovery

show process creditmon credit-loss-events

Use the `show process creditmon credit-loss-events` command to display the time stamps when remaining Tx B2B credits are zero for 1 second on F_port and 1.5 second on E_port.

```
MDS9700# show process creditmon credit-loss-events

        Module: 01      Credit Loss Events: YES

----------------------------------------------------
| Interface |  Total |           Timestamp          |
|           | Events |                              |
----------------------------------------------------
| fc1/13    |  11524 | 1. Sat Mar 29 14:21:48 2014 |
|           |        | 2. Sat Mar 29 14:21:47 2014 |
|           |        | 3. Sat Mar 29 14:21:46 2014 |
|           |        | 4. Sat Mar 29 14:21:45 2014 |
|           |        | 8. Sat Mar 29 14:21:41 2014 |
|           |        | 9. Sat Mar 29 14:21:40 2014 |
|           |        |10. Sat Mar 29 14:21:39 2014 |
----------------------------------------------------
```

This command is available on all Cisco MDS 9000 Family switches.

Link reset and link reset response (LRR) are exchanged for credit loss recovery. These counters are available under the `show interface counter detail` NX-OS exec-level command.

```
MDS9000# show interface fc1/3 counter details
<output truncated>
      0 link reset received while link is active
      0 link reset transmitted while link is active
<output truncated>
```

This command is available on all Cisco MDS 9000 Family switches.

OBFL error-stats

Use show `logging onboard error-stats` to list when remaining Tx B2B credits are zero for 1 second on F_port and 1.5 second on E_port.

```
MDS9700# show logging onboard error-stats
--------------------------------------------------------------------------------
 ERROR STATISTICS INFORMATION FOR DEVICE DEVICE: FCMAC
--------------------------------------------------------------------------------
    Interface      |                                |          |    Time Stamp
      Range        |    Error Stat Counter Name     |  Count   |MM/DD/YY HH:MM:SS
                   |                                |          |
--------------------------------------------------------------------------------
fc1/13             |F16_TMM_TOLB_TIMEOUT_DROP_CNT   |1496855   |04/07/15 22:44:23
fc1/13             |FCP_SW_CNTR_TX_WT_AVG_B2B_ZERO  |217       |04/07/15 22:44:23
fc1/13             |FCP_SW_CNTR_CREDIT_LOSS         |19        |04/07/15 22:44:23
fc1/13             |F16_TMM_TOLB_TIMEOUT_DROP_CNT   |1486654   |04/07/15 22:44:03
fc1/13             |FCP_SW_CNTR_TX_WT_AVG_B2B_ZERO  |108       |04/07/15 22:44:03
fc1/13             |FCP_SW_CNTR_CREDIT_LOSS         |9         |04/07/15 22:44:03
```

Watch for counter name <*component_id*>_CNTR_CREDIT_LOSS. In this output, the value of FCP_SW_CNTR_CREDIT_LOSS is incremented from 9 to 19 between 22:44:03 and 22:44:23 on April 7, 2015.

This command is available on all Cisco MDS 9000 Family switches, although the counter names displayed vary from platform to platform. See Appendix F for counter names and descriptions for Cisco MDS 9000 Family switches and line cards.

show tech-support slowdrain

This is a new variant of show tech-support command. It contains the output of all the slowdrain relevant commands. For best results, it is recommended to execute this command from DCNM on all the switches of the fabric at the same time.

Appendix D: Cisco MDS 9000 Family Slow-Drain–Specific SNMP MIBs

This section provides details about available SNMP Management Information Bases (MIBs) and OIDs, which can be used to collect counters remotely on an NMS. The NMS provides a centralized view of a fabric, which proves to be extremely useful for a situation such as slow drain, which is known to have a fabricwide impact.

Table 11. Cisco MDS 9000 Family Slow-Drain OIDs

SNMP Object	Description
fcIfTxWaitCount (1.3.6.1.4.1.9.9.289.1.2.1.1.15)	TxWait counter
fcHCIfBBCreditTransistionFromZero (1.3.6.1.4.1.9.9.289.1.2.1.1.40)	Tx B2B credit transition to zero
fcIfBBCreditTransistionToZero (1.3.6.1.4.1.9.9.289.1.2.1.1.41)	Rx B2B credit transition to zero
fcIfTxWtAvgBBCreditTransitionToZero (1.3.6.1.4.1.9.9.289.1.2.1.1.38)	Credit unavailability at 100 ms
fcIfCreditLoss (1.3.6.1.4.1.9.9.289.1.2.1.1.37)	Credit Loss (recovery)
fcIfTimeOutDiscards (1.3.6.1.4.1.9.9.289.1.2.1.1.35)	Timeout discards
fcIfOutDiscard (1.3.6.1.4.1.9.9.289.1.2.1.1.36)	Total number of frames discarded in egress direction, which includes timeout discards
fcIfLinkResetIns (1.3.6.1.4.1.9.9.289.1.2.1.1.9)	Number of link reset protocol errors received by the FC port from the attached FC port
fcIfLinkResetOuts (1.3.6.1.4.1.9.9.289.1.2.1.1.10)	Number of link reset protocol errors issued by the FC port to the attached FC port.
fcIfSlowportCount (1.3.6.1.4.1.9.9.289.1.2.1.1.44)	Duration for which Tx B2B credits were unavailable on a port
fcIfSlowportOperDelay (1.3.6.1.4.1.9.9.289.1.2.1.1.45)	Number of times for which Tx B2B credits were unavailable on a port for a duration longer than the configured admin-delay value in slowport-monitor

For more information about each OID, such as the object name and MIB, use the Cisco SNMP Object Navigator.

Appendix E: Cisco MDS 9000 Family Slow-Drain Feature Support Matrix

The following Tables 12–14 describe slow-drain detection, troubleshooting, and recovery features,

Table 12. Slow-Drain Detection Features

Feature Name	Details	Supported Platforms	Minimum NX-OS Release
`TxWait (2.5 µs)`		MDS 9700, MDS 9396S, MDS 9148S, MDS 9250i and MDS 9500 only with DS-X9232-256K9 & DS-X9248-256K9 line cards	6.2(13)
`Slowport-monitor (1 ms)`		MDS 9700, MDS 9148S & MDS 9250i	6.2(9)
		Extended to MDS 9396S and MDS 9500 only with DS-X9232-256K9, DS-X9248-256K9, DS-X9248-48K9, DS-X9224-96K9 & DS-X9248-96K9 line cards	6.2(13)
`Credit unavailability at 100 ms`		All	5.x.x or 6.x.x
`LR Rcvd B2B`		All	5.x.x or 6.x.x
`Credits and remaining credits`		All	5.x.x or 6.x.x
`Credit transition to zero`		All	5.x.x or 6.x.x

Table 13. Slow-Drain Troubleshooting Features

Feature Name	Details	Supported Platforms	Minimum NX-OS Release
`Information about dropped frames`		MDS 9700 and MDS 9396S	6.x.x
`Display frames in ingress Q`		MDS 9700, MDS 9396S and MDS 9500	5.x.x or 6.x.x
`Display arbitration timeout`		All	5.x.x or 6.x.x
`Timeout discards`		MDS 9700, MDS 9396S, MDS 9148S, MDS 9250i & MDS 9500 only with DS-X9232-256K9 and DS-X9248-256K9 line cards	6.2(13)
`TxWait history graph`		All	All
`OBFL Error stats`		MDS 9700 and MDS 9396S	6.x.x
`show tech-support slowdrain`		All	6.2(13)

Table 14. Slow-Drain Automatic Recovery Features

Feature Name	Details	Supported Platforms	Minimum NX-OS Release
`congestion-drop timeout`		All	5.x.x or 6.x.x
`no-credit-drop timeout`		Software only: All	5.x.x or 6.x.x
		Hardware assisted: MDS 9700, MDS 9148S, MDS 9250i	6.2(9)
		MDS 9396S	6.2(13)
`Credit loss recovery`		All	5.x.x or 6.x.x
`Port Monitor`	`credit-loss-reco`	All	5.x.x or 6.x.x
	`lr-rx`	All	5.x.x or 6.x.x
	`lr-tx`	All	5.x.x or 6.x.x
	`timeout-discards`	All	5.x.x or 6.x.x
	`tx-credit-not-available`	All	5.x.x or 6.x.x
	`tx-discards`	All	5.x.x or 6.x.x

	slowport-count	MDS 9500 only with DS-X9248-48K9, DS-X9224-96K9 or DS-X9248-96K9 line cards	6.2(13)
	slowport-oper-delay	MDS 9700, MDS 9396S, MDS 9148S, MDS 9250i and MDS 9500 only with DS-X9232-256K9 or DS-X9248-256K9 line cards	6.2(13)
	tx-wait	MDS 9700, MDS 9396S, MDS 9148S, MDS 9250i and MDS 9500 only with DS-X9232-256K9 or DS-X9248-256K9 line cards	6.2(13)

Appendix F: Cisco MDS 9000 Family Counter Names and Descriptions

Table 15 lists the counter names available under NX-OS shows command and describes their meaning.

Table 15. Counters

Counter Name	Counter Description
FCP_CNTR_RCM_CH0_LACK_OF_CREDIT[2] AK_FCP_CNTR_RCM_CH0_LACK_OF_CREDIT[3] THB_RCM_RCP0_RBBZ_CH0[4] F16_RCM_RCP0_RBBZ_CH0[5]	Total count of transitions to zero for Rx B2B credits on ch0; these transitions typically indicate that the switch is applying back pressure to the attached device because of perceived congestion, and this perceived congestion can be the result of a lack of Tx B2B credits being returned on an interface over which this device is communicating.
FCP_CNTR_LAF_TOTAL_TIMEOUT_FRAMES[2] AK_FCP_CNTR_LAF_TOTAL_TIMEOUT_FRAMES[3] THB_TMM_TOLB_TIMEOUT_DROP_CNT[4] F16_TMM_TOLB_TIMEOUT_DROP_CNT[5]	Timeout drops at egress.
FCP_CNTR_QMM_CH0_LACK_OF_TRANSMIT_CREDIT[2] AK_FCP_CNTR_QMM_CH0_LACK_OF_TRANSMIT_CREDIT[3] THB_TMM_PORT_TBBZ_CH0[4] F16_RCM_RCP0_TBBZ_CH0[5]	Total count of transitions to zero for Tx B2B credits on ch0; these transitions are typically the result of the attached device's withholding of R_RDY primitive from the switch due to congestion in that device.
None[2] None[3] THB_TMM_PORT_FRM_DROP_CNT[4] F16_TMM_PORT_FRM_DROP_CNT[5]	Number of frames dropped in **tolb_path** or **np path**; these drops include all types of frame drops: timeout, offline, abort drops at egress, etc.
None[2] None[3] THB_TMM_PORT_TWAIT_CNT[4] F16_TMM_PORT_TWAIT_CNT[5]	Frame is available to send, but no credit is available; increments every cycle (cycle = 2.35 nanoseconds).
FCP_CNTR_LAF_C3_TIMEOUT_FRAMES_DISCARD[2] AK_FCP_CNTR_LAF_C3_TIMEOUT_FRAMES_DISCARD[3] THB_TMM_TO_CNT_CLASS_3[4] F16_TMM_TO_CNT_CLASS_3[5]	Count of Class 3 Fibre Channel frames dropped as a result of congestion-drop timeout.
FCP_CNTR_RX_WT_AVG_B2B_ZERO[2] AK_FCP_CNTR_RX_WT_AVG_B2B_ZERO[3] FCP_SW_CNTR_RX_WT_AVG_B2B_ZERO[4] FCP_SW_CNTR_RX_WT_AVG_B2B_ZERO5(unable to generate)	Count of the number of times an interface was at zero Rx B2B credits for 100 ms; this status typically indicates that the switch is withholding R_RDY primitive to the device attached on that interface due to congestion in the path to devices with which it is communicating
FCP_CNTR_TX_WT_AVG_B2B_ZERO[2] AK_FCP_CNTR_TX_WT_AVG_B2B_ZERO[3] FCP_SW_CNTR_TX_WT_AVG_B2B_ZERO[4,5]	Count of the number of times that an interface was at zero Tx B2B credits for 100 ms; this status typically indicates congestion at the device attached on that interface.
FCP_CNTR_FORCE_TIMEOUT_ON[2] AK_FCP_CNTR_FORCE_TIMEOUT_ON[3] FCP_SW_CNTR_FORCE_TIMEOUT_ON[4,5]	Count of the number of times the system timeout no-credit-drop threshold has been reached by this port; when a port is at zero Tx B2B credits for the time specified, the port starts to drop frames at line rate.
FCP_CNTR_FORCE_TIMEOUT_OFF[2] AK_FCP_CNTR_FORCE_TIMEOUT_OFF[3] FCP_SW_CNTR_FORCE_TIMEOUT_OFF[4,5]	Count of the number of times that the port has recovered from the system timeout no-credit-drop condition; this status typically means that R_RDY primitive has been returned or possibly that an LR and LRR even has occurred.

FCP_CNTR_LAF_CF_TIMEOUT_FRAMES_DISCARD[2] AK_FCP_CNTR_LAF_CF_TIMEOUT_FRAMES_DISCARD[3] THB_TMM_TO_CNT_CLASS_F[4] F16_TMM_TO_CNT_CLASS_F[5]	Count of Class F Fibre Channel frames dropped due to congestion-drop timeout.
FCP_CNTR_CREDIT_LOSS[2] AK_FCP_CNTR_CREDIT_LOSS[3] FCP_SW_CNTR_CREDIT_LOSS[4,5]	Count of the number of times that creditmon credit loss recovery has been invoked on a port.

[1]: Generation 1 modules are no longer supported by NX-OS 5.0, and later, and are not covered by this white paper.

[2]: DS-X9112, DS-X9124, and DS-X9148 and DS-X9304-18K9 - modules are not covered by this document.

[3]: 8 Gbps DS-X9248-48K9 and DS-X92xx-96K9 modules

[4]: Advanced 8 Gbps DS-X92xx-256K9 module

[5]: Cisco MDS 9700 DS-X9448-768K9 module and MDS 9396S

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 C11-737315-00 07/16