White Paper

Cisco MDS 9000 Family Quality of Service

Quality of Service (QoS) enables traffic differentiation and prioritization, allowing latency-sensitive applications such as online transaction processing (OLTP) to share storage resources alongside throughput-intensive applications such as data warehousing.

INTRODUCTION

This white paper serves as a guide to the advanced QoS and traffic engineering features present in Cisco® MDS 9000 Family switches. It describes enhancements to QoS in second-generation Cisco line cards and supervisors and includes example storage area network (SAN) designs in which enabling QoS will provide better overall service.

WHY IS QoS NEEDED?

The primary goal of QoS is to provide priority for traffic flows to and from specific devices. In this context, priority means providing lower latency and higher bandwidth connections with more controlled jitter.

An underlying principle of Fibre Channel switching is that the network guarantees that no frames will be dropped. If this is the case, why do we need QoS at all? Switches today provide high-performance, non-blocking, non-oversubscribed crossbar switch fabrics. The Cisco MDS 9513 Multilayer Director can switch more than a billion frames per second. Why would users ever need QoS when a switch fabric provides seemingly endless amounts of frame-switching capacity?

The answer is simple: congestion.

Congestion occurs for two basic reasons:

- Congestion will occur if multiple senders are contending with a smaller number of receivers. If the aggregate rate of traffic transmitted by senders exceeds the size of the connection to the receivers, blocking will occur (Figure 1).
- Any time there is a speed mismatch between senders and receivers, buffering will occur. Buffers are a finite resource on switches, typically in the range of 16 buffers (32 KB) to 255 buffers (512 KB) per port. When these buffers are full, blocking occurs (Figure 2).

Figure 1. Congestion Caused by Senders Outnumbering Receivers

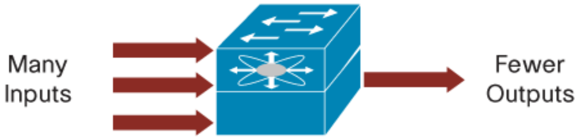
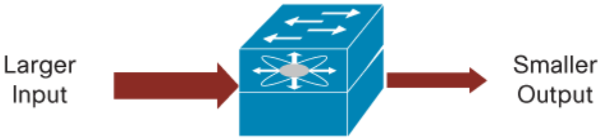


Figure 2. Congestion Caused by Speed Mismatch between Senders & Receivers



Many organizations consolidate their SAN infrastructure in order to realize cost savings and increased management efficiencies by pooling disparate storage resources into one single physical storage fabric. Managing contention for resources is an important aspect in realizing the business benefits associated with storage consolidation. If storage resources become congested because noncritical business applications cause time-sensitive mission-critical applications to become slowed down, the cost benefits associated with SAN consolidation quickly disappear.

There is no automatic quick fix to alleviate congestion and blocking. It is possible to add more buffering to a switch, but additional buffers will not remove the congestion; they will simply increase the time it takes for congestion to turn into blocking. Virtual output queuing (VOQ) can be used to prevent one blocked receiver from affecting traffic being sent to other noncongested receivers ("head-of-line blocking"), but it does not do anything for traffic being sent to the congested device.

SAN traffic can be categorized as a large number of devices (hosts) communicating with a smaller number of devices (storage ports). Put another way, SAN designs are almost always over-subscribed, with more host-attached ports than storage-attached ports. What is important is making sure that the oversubscription does not impact the performance of mission-critical time-sensitive applications.

QoS IN THE CISCO MDS 9000 FAMILY

Cisco MDS 9000 Family switches were among the first SAN switches to offer QoS. Enabled in November 2003 with Cisco MDS 9000 SAN-OS Software Release 1.3, QoS-enabled switches provided traffic differentiation and prioritization, enabling latency-sensitive applications such as OLTP to share common storage resources alongside throughput-intensive applications such as data warehousing.

QoS can be used alongside other traffic engineering features such as Fiber Channel Congestion Control (FCC) and ingress port-rate limiting and can be configured to apply different policies at different times of day using the command scheduler built into Cisco MDS 9000 SAN-OS Software.

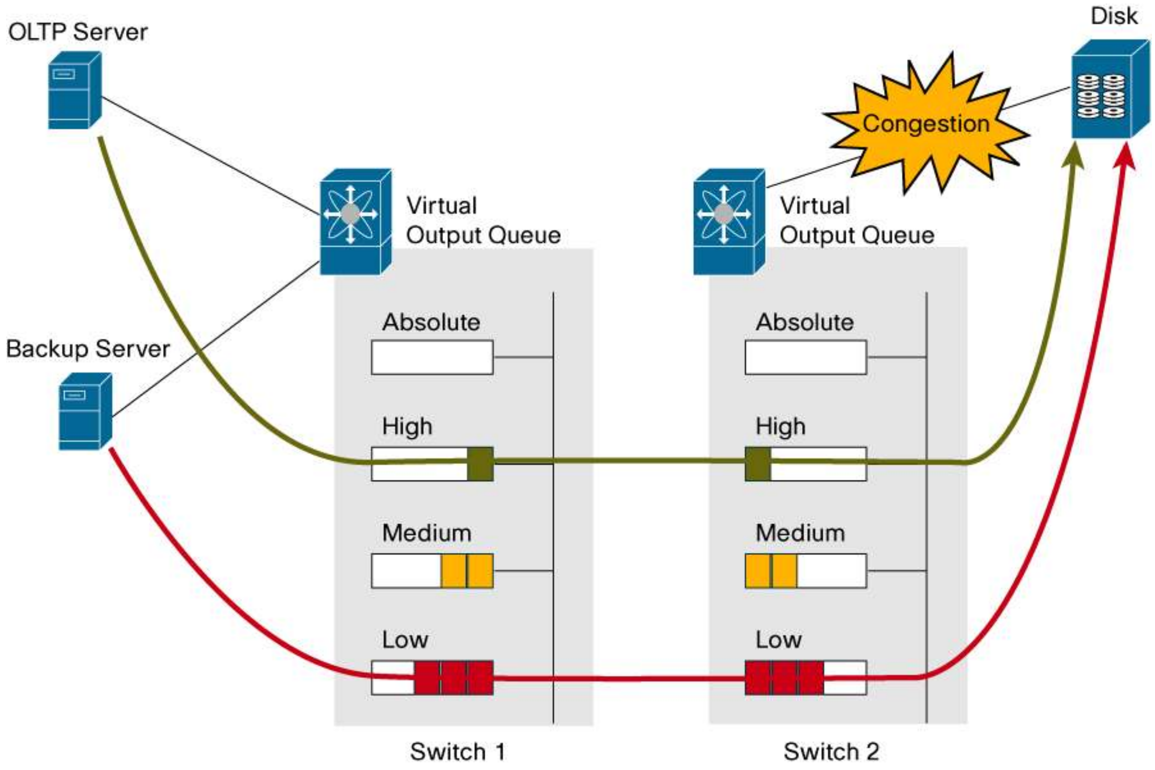
QoS capabilities have been enhanced with the introduction of second-generation Cisco MDS 9000 Family line-card modules and supervisors. The following paragraphs describe the implementation differences and ramifications. The semantics for QoS configuration are identical, regardless of generation.

QoS on First-Generation Line Cards and Supervisors

QoS on first-generation Cisco MDS switch hardware is effective at mitigating the negative effects of traffic contention on ports connected to storage arrays, where one or more application servers are contending for access to storage resources. For QoS to be effective, hosts and storage arrays need to be connected to different switches. QoS policies function only when there is congestion on end-port devices; QoS cannot do anything to alleviate congestion on Inter-Switch Links (ISLs) in the core of the SAN.

Figure 3 shows how QoS could be deployed on first-generation switches, line cards, and supervisors. The two switches in the figure are connected with one or more ISLs.

Figure 3. QoS Deployment on First-Generation Switches, Line Cards, and Supervisors



In this diagram, OLTP traffic arriving at switch 1 is marked with a high priority level through classification (class map) and marking (policy map). Similarly, the backup traffic is marked with a low priority level. Traffic is queued within a virtual output queue corresponding to the frame priority. Frames from each virtual output queue are forwarded to switch 2.

When the frames arrive at Switch 2, OLTP requests are queued in the high-priority virtual output queue. If the output storage port is congested, requests from the OLTP server will be serviced at a higher rate than that of requests from the backup server. This is because the scheduler services OLTP traffic (in the high-priority queue) more frequently than backup traffic in the low-priority queue. The latency experienced by the OLTP server is independent of the volume of low-priority traffic and congestion on the output interface.

Deficit weighted round robin (DWRR) scheduling is used to help ensure that high-priority traffic is treated more quickly than low-priority traffic. For example, DWRR weights of 70:20:10 (high:medium:low) imply that the high-priority queue is serviced at seven times the rate of the low-priority queue. This guarantees lower latency and higher throughput to high-priority traffic, where there is congestion on an output port.

Note that when an output port is not congested, there is no queuing, and all traffic is treated equally.

QoS on Second-Generation Line Cards and Supervisors

On second-generation Cisco MDS 9000 Family line-card modules and supervisors, QoS has been enhanced to provide traffic differentiation and prioritization regardless of where the congestion occurs. Second-generation Cisco MDS 9000 Family line-card modules and supervisors can enforce QoS policies for congestion on end ports across multiswitch SAN fabrics, end ports in a single-switch fabric, and congestion on ISLs.

The primary difference is that first-generation Cisco MDS 9000 Family line-card modules and supervisors enforce QoS on ingress, whereas second-generation Cisco MDS 9000 Family line-card modules and supervisors enforce QoS on egress. First-generation line-card modules and supervisors could only enforce a given QoS policy map where traffic destined for a congested output port was arriving on the same input port. This typically required incoming traffic to be aggregated onto common input ports such as ISLs, hence the multiswitch requirement. Second-generation line-card modules and supervisors enforce QoS on egress, removing this limitation and enabling QoS to be effective in single-switch environments and where there is congestion on ISLs.

Table 1 lists first-generation Fibre Channel line cards.

Table 1. First-Generation Fibre Channel Line Cards

Cisco Part Number	Description
DS-X9016	16-port 1/2-Gbps Fibre Channel module
DS-X9032	32-port 1/2-Gbps Fibre Channel module
DS-X9032-SSM	32-port 1/2-Gbps Fibre Channel storage services module
DS-X9302-14K9	2-port 1-Gigabit Ethernet IPS, 14-port 1/2-Gbps Fibre Channel module

Table 2 lists second-generation Fibre Channel line cards.

Table 2. Second-Generation Fibre Channel Line Cards

Cisco Part Number	Description
DS-X9112	12-port 1/2/4-Gbps Fibre Channel module
DS-X9124	24-port 1/2/4-Gbps Fibre Channel module
DS-X9148	48-port 1/2/4-Gbps Fibre Channel module
DS-X9704	4-port 10-Gbps Fibre Channel module

QoS CONFIGURATION

QoS is a licensed feature and requires an Enterprise Package license installed on all switches where QoS is enabled. Although QoS is a licensed feature, users can try licensed features for up to 120 days using a license grace period.

QoS configuration should be consistent across multiple switches to help ensure that all switches are enforcing a common policy for traffic in both send and receive directions.

QoS is configured in an identical manner regardless of whether the switch has first-generation or second-generation line cards present. QoS can be deployed in any one of three ways depending on the complexity of the QoS policy desired:

- Virtual SAN (VSAN)–based QoS
- Zone-based QoS
- Individual QoS policies matching individual devices

VSAN-Based QoS

VSAN-based QoS enables QoS priority to be assigned on a per-VSAN basis.

VSAN-based QoS enables individual VSANs to be configured with a high, medium, or low QoS priority. All traffic flow between devices in that VSAN will inherit the desired priority.

VSAN-based QoS is useful when multiple VSANs (for example, production, development, and test) share common ISLs between data centers, where priority should always be given to production traffic in the event of congestion.

Zone-Based QoS

Where more granular QoS is required, QoS priority can be assigned on a per-zone basis.

Enhanced zoning on Cisco MDS 9000 Family switches can be used to associate a QoS priority with a zone. All communication between devices common to that zone is mapped to the configured QoS priority (high, medium, or low).

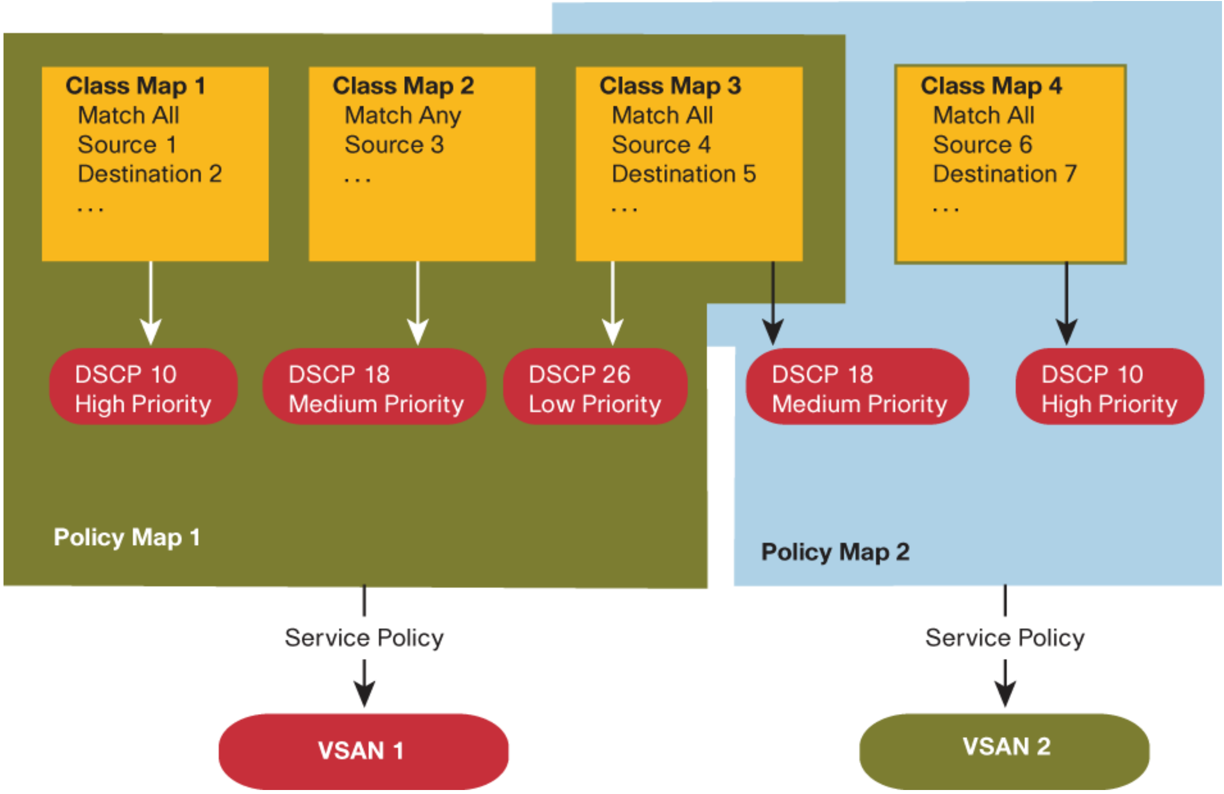
Zone-based QoS is useful where different classes of application servers are in a single VSAN and providing a higher priority to time-sensitive mission-critical applications sharing common storage resources with less critical hosts and applications is desirable. An example would be sharing storage array resources between an OLTP application and a server for common workgroup storage. QoS could be used to assign priority to the OLTP application such that if the connectivity to the storage array is congested, the OLTP application has priority over the workgroup servers.

Zone-based QoS is dependent on enhanced zoning and the VSAN operating in interop mode 0.

Individual QoS Policies Matching Individual Devices

Where maximum flexibility is desired, QoS policy can be defined on a per-device basis, with individual policies applied to different devices and VSANs (Figure 4). QoS class maps are used to classify traffic. Classifications are mapped to policies in QoS policy maps. Policies are assigned to VSANs using QoS service maps.

Figure 4. Per-Device QoS Policy Definition



Traffic classification is performed using QoS class maps. QoS class maps can match frames using multiple-criteria matches ("match all") or single-criteria matches ("match any"), based on any of the following commands:

- **source-wwn**—Match frames based on the World Wide Name of the device sending traffic
- **destination-wwn**—Match frames based on the World Wide Name of the device receiving traffic
- **source-address**—Match frames based on the Fibre Channel ID (FC_ID) of the device sending traffic
- **destination-address**—Match frames based on the FC_ID of the device receiving traffic
- **input-interface**—Match frames based on the interface on which the frames arrive
- **destination-device-alias**—Match frames based on the alias of the device receiving traffic

QoS policy maps are used to provide an ordered mapping for QoS class maps to service levels. A single policy map can contain multiple class maps associated to a single service level (high, medium, or low) or differentiated services code point (DSCP). All DSCP values except 46 (expedite forwarding) are allowed.

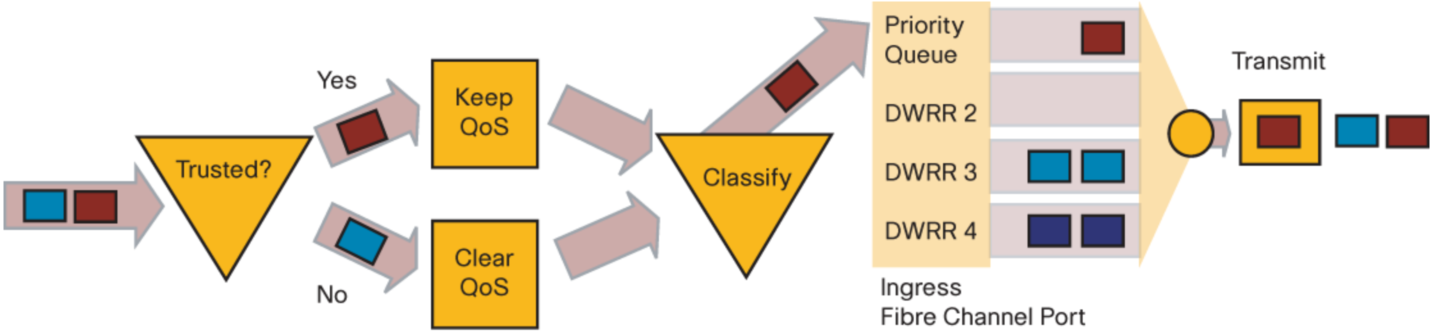
QoS service maps are used to associate QoS policy maps to VSANs.

SCHEDULING FRAMES FOR TRANSMISSION

When frames first arrive into the switch, existing QoS information in the frame headers can be considered to be either trusted or untrusted. All Cisco MDS 9000 Family switches default to not trust any QoS information on frames arriving at the switch, unless the frame arrives on a trunked E_Port (TE_Port). Existing QoS information in a frame is always preserved when a frame is forwarded, even if there are QoS policies that match the frame; QoS policies only alter the forwarding behavior within the switch and never alter the actual frame itself.

Where a frame arrives on a trusted port, the switch copies down QoS information in the frame header. Unless a QoS service map that matches the frame exists, the QoS service level from the incoming frame is used when the frame is forwarded. For frames arriving on untrusted ports, unless a QoS service map that matches the frame exists, it will be treated at the default service level (Figure 5).

Figure 5. QoS Processing Steps



Frames are mapped to one of four virtual output queues per output port, based on the QoS service level or DSCP mapping, as shown in Table 3.

Table 3. Mapping of Frames to Virtual Output Queues

Service Level	DSCP	PHB (per FC-FS-2 Standard)	Mapped to Queue
Absolute	46	EF (expedite forwarding)	Priority queue
High	10, 12, 14	AF1 (assured forwarding 1)	DWRR1 queue
Medium	18, 20, 22	AF2 (assured forwarding 2)	DWRR2 queue
Low	26, 28, 30	AF3 (assured forwarding 3)	DWRR3 queue
	34, 36, 38	AF4 (assured forwarding 4)	
Default	All others	All others	

Four virtual output queues are available for each output port, corresponding to four QoS levels per port: one priority queue and three queues serviced using DWRR scheduling. Frames queued in the priority queue will be scheduled for transmission prior to any frames queued in the DWRR queues. Frames queued in the three DWRR queues are scheduled using whatever relative weightings are configured. The default relative weightings are 33:33:33; however, these are configurable items and can be set to any ratio at all.

The amount of ingress buffering per port is also a configurable option, with options varying for buffering for 12 frames up to 4096 frames. In most cases, there is no need to change ingress buffer levels beyond the default.

OTHER TRAFFIC ENGINEERING FEATURES

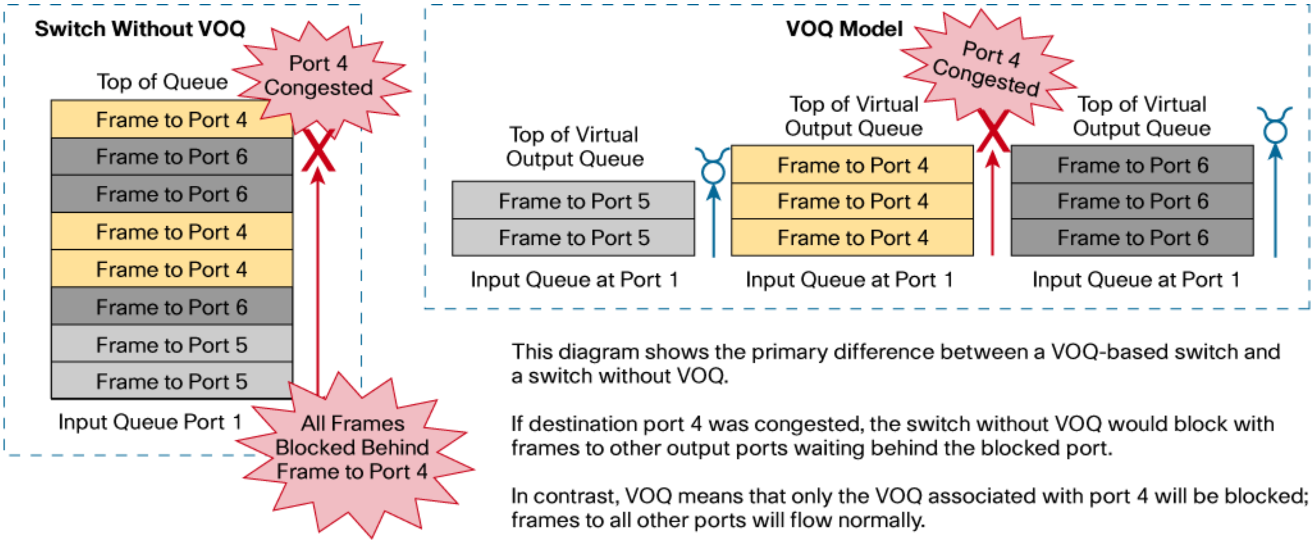
VOQ

All Cisco MDS 9000 Family switches use the VOQ queuing technique (Figure 6). This is a standard feature that is always enabled.

VOQ helps mitigates a phenomenon called “head-of-line blocking”. Blocking occurs when congestion on an output port blocks frames from being sent. Fibre Channel switches provide a guarantee that no frames will ever be dropped, so when an output port is congested, switches have to block until such time as the frame can be transmitted. Head-of-line blocking occurs when the frame at the head of the queue cannot be sent due to congestion at its output port. In this situation, frames behind this frame are “blocked” from being sent to their destination, even though their respective output ports are not congested.

VOQ prevents head-of-line blocking by utilizing multiple virtual output queues. Individual virtual output queues might block, but traffic queued for different (nonblocked) destinations can continue to flow without getting delayed behind frames waiting for the blocking to clear on a congested output port.

Figure 6. Comparison of Switches with and without VOQ



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.

Fairness

All Cisco MDS 9000 Family switches switch frames in a fair manner. Unless QoS is enabled, fairness helps ensure that frames destined for a congested output port will always be given an equal share of the output port capacity. With fairness, no single input device can use more than its share of output port resources and deprive other input devices.

Fairness is always guaranteed, regardless of what line-card modules are used for inputs and outputs.

Fibre Channel Congestion Control

All Cisco MDS 9000 Family switches support Fibre Channel Congestion Control (FCC). FCC is used to prevent blocking on an output port caused by sending devices transmitting traffic to receiving devices more quickly than the receiving devices can handle that traffic.

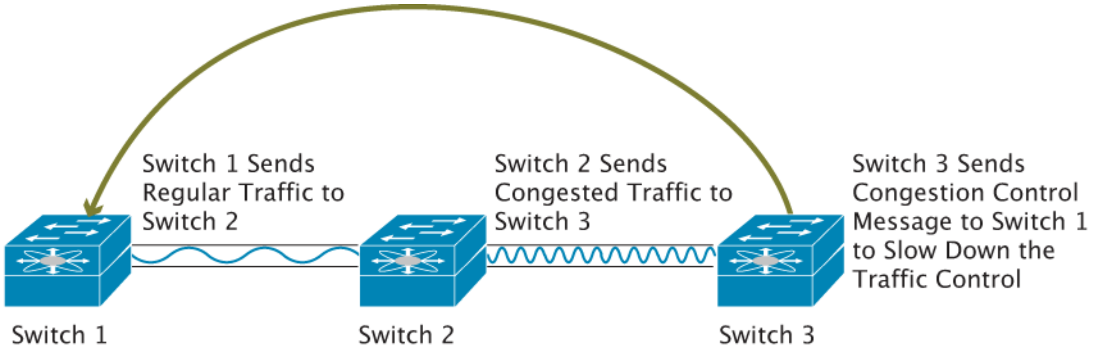
There are two common cases where this can happen:

- Speed mismatch between senders and receivers in a fabric (for example, a 2-Gbps device transmitting to a 1-Gbps device)
- Receivers that cannot maintain sustained performance (for example, an LTO-2 tape drive connected to the fabric at 2 Gbps but only capable of sustaining write throughput of 40 MBps [0.4 Gbps])

Where congestion and blocking exist on one output port, unless the sending device slows down, blocking can spread to additional ports and devices. The cause is related to an underlying principle of Fibre Channel switching: switches must guarantee that no frames will ever be dropped.

FCC can be used to mitigate congestion and blocking from spreading by using signaling to tell devices in the path to slow down the sending device that is causing the congestion (see Figure 7).

Figure 7. FCC Congestion Mitigation and Blocking


Switch 1 Sends Regular Traffic to Switch 2
Switch 2 Sends Congested Traffic to Switch 3
Switch 3 Sends Congestion Control Message to Switch 1 to Slow Down the Traffic Control
Switch 1
Switch 2
Switch 3

When FCC detects congestion on an output port, it generates an FCC edge quench frame, with a destination address of the device causing the congestion (the sender). Switches that receive FCC edge quench will inspect the FCC frame and determine if it is targeted at a directly attached device (the sending device causing the congestion is directly attached). If it is, the switch will instigate rate limiting on the input port attached to the sending device causing congestion, thereby rate-limiting the sender to the rate at which the receiver can receive.

FCC works on an active-loop feedback system. While there is congestion, the switch with the congested output port will continue to send FCC edge quench frames. The switch connected to the sending device causing the congestion will continue to rate-limit the flow until the congestion has gone away.

FCC and FCC edge quench frames are completely compatible with existing Fibre Channel standards and fabrics and can be used in mixed-vendor fabrics.

Ingress Port-Rate Limiting

Ingress port-rate limiting is a feature that can be used to provide precise control over the amount of bandwidth available to individual devices attached to Cisco MDS 9000 Family switches. It works by limiting the rate at which the switch processes frames sent by attached devices, rate-limiting the pace at which buffer credits are returned to the devices sending traffic. Rate-limiting can be configured anywhere from 1 percent to 100 percent of the port speed in increments of 1 percent , with the default being 100 percent.

Ingress port-rate limiting is available on Cisco MDS 9100 Series multilayer fabric switches, the Cisco MDS 9216i Multilayer Fabric Switch, Fibre Channel ports on MPS-14/2 modules, and all second-generation Cisco MDS line-card modules. It requires an Enterprise Feature license to be enabled.

CONCLUSION

Cisco Systems® continues to improve its multiprotocol intelligent SAN switches. Providing industry-leading availability, scalability, security, management, and investment protection, Cisco MDS 9000 Family switches allows you to deploy high-performance SANs with low TCO, layering a rich set of intelligent features onto a high-performance, protocol-independent switch fabric.

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

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

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

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

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

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

Printed in the USA