Learn About Data Center Bridging

Introduction to General Concepts

Data Center
A physical or virtual group of network devices and maintenance equipment used to store, manage, and transport large amounts of electronic information, usually with redundant resources (power, ports, environmental controls, and so on) to mitigate network outages.

A bridge connects two points that are separated by a perilous obstacle, such as a deep, rocky gorge, or a raging river. Thanks to the bridge, cars and other traffic on the path travel between the two points and overcome the obstacle safely, as if it posed no danger at all.

In the network world, data center bridging (DCB) also connects two points that are separated by a perilous obstacle. The two points to be connected are a device on the storage area network (SAN) segment and a device on the Ethernet segment of a converged data center network. The traffic that travels between these two points is not cars, but storage data that crosses the converged network, travelling between the SAN device and the Ethernet device. And the perilous obstacle is the Ethernet network itself—at least, as far as the SAN traffic is concerned (see Figure 1)."

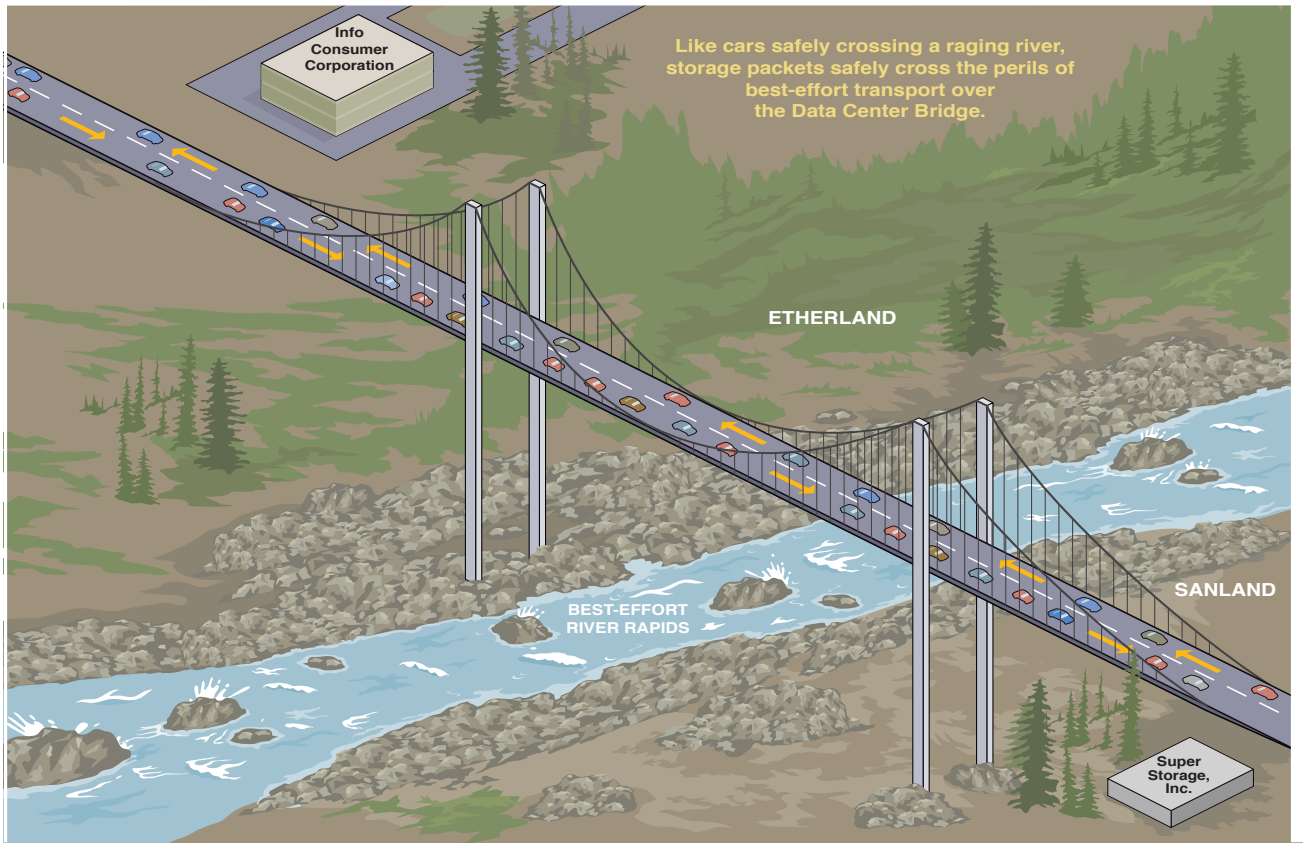

Figure 1 Storage Traffic Uses Data Center Bridging to Traverse a Converged Ethernet Network

Lossless transport
Lossless transport means that no frames are dropped because of network congestion. Lossless transport does not mean that no frames are ever dropped for any reason. Network failure conditions such as device failure or link failure can still cause frame loss.

Ethernet networks are designed to transport best-effort traffic. Ethernet networks tolerate frame loss and retransmission, and sustain and compensate for packet collisions, errors, and out-of-order packet delivery. However, SAN traffic, such as Fibre Channel (FC) traffic that is encapsulated in Ethernet (Fibre Channel over Ethernet, or FCoE), requires lossless transport not only across the SAN, but also across the Ethernet portion of the converged data center network. But in its natural state, an Ethernet network doesn't ensure lossless data transport, or even encourage it.

To solve the problem of how to deliver lossless storage traffic across a converged Ethernet data center network, the IEEE 802.1 working group developed DCB, a set of open-standards Ethernet enhancements to the IEEE 802.1 bridging specifications. DCB tames the perilous obstacle of lossy transport by forcing the Ethernet network to behave in an entirely unnatural manner – following the rules storage traffic requires for lossless transport – instead of allowing packets to drop and collide and retransmit in a Wild West packet shootout on the switched Layer 2 Ethernet links.

DCB enables you to treat different types of traffic in different ways on the same physical Ethernet link. For storage traffic such as FCoE or iSCSI, DCB provides a safe bridge that transports traffic losslessly over perilous Ethernet networks without affecting the way that the Ethernet network handles best-effort traffic. (Lossless transport means that the network drops no frames because of network congestion that overflows a switch's queue buffers. DCB can't protect data against frame loss caused by network issues such as device failures or link failures.) DCB relieves the effects of network congestion like an extra-strength nasal spray opens stuffed sinuses, by using queue management techniques to prevent queue overflow (and thus frame drops), and bandwidth allocation enhancements to utilize port bandwidth as efficiently as possible.

Data Center Network Convergence

Network convergence in a data center is the concept of carrying standard, best-effort Ethernet traffic while at the same time carrying storage traffic that requires lossless transport, using a common Ethernet infrastructure as a unified fabric.

The desire to converge Ethernet and storage networks drove the development of the DCB standards. The primary driver was the need to transport FCoE traffic across Ethernet networks. (FCoE is native Fibre Channel frames encapsulated in Ethernet. The Ethernet network uses the Ethernet frame headers to forward and handle traffic appropriately.)

The DCB extensions to Ethernet standards support not only the transport of storage traffic such as FCoE and iSCSI, but also the transport of any traffic that requires lossless handling. The amount of lossless traffic a converged network can handle depends on characteristics such as bandwidth, traffic load, frame size, and physical distance between devices.

Using separate networks for Ethernet and storage traffic requires separate sets of switches, links, wires, power supplies, and so on. Network convergence boasts several money and space saving benefits, such as reducing the number of devices required instead of using completely separate networks, reducing the number of interfaces required to transport traffic, reducing cabling complexity, and reducing administration activities such as network management, maintenance, and upgrades. Converged networks also save on power and cooling costs.

Data Center Bridging

The DCB enhancements make Ethernet a viable infrastructure for storage and other traffic that requires lossless transport by providing the level of class of service, sometimes known as quality of service, needed to transport lossless traffic. The DCB task group developed four specifications that help eliminate frame loss due to network congestion:

Flow Control Mechanism
A method of regulating traffic to avoid dropping frames during periods of network congestion. Flow control stops and resumes the transmission of network traffic between two connected peer devices to prevent output queues from overflowing and dropping frames during periods of congestion.

- Priority-based Flow Control (PFC, IEEE 802.1Qbb) – A link-level flow control mechanism that pauses traffic on an Ethernet link to prevent frame loss caused by network congestion. PFC divides a physical link into eight virtual "lanes" of traffic. PFC controls the classes of flows assigned to each lane of traffic independently, so that if a lane of traffic is paused, the other lanes of traffic on the link are not paused. Each virtual lane is called a priority. This enables one physical link to carry traffic that requires lossless transport at the same time as carrying loss-tolerant Ethernet traffic. PFC works by telling the directly connected peer device to temporarily stop (pause) transmitting the congestion-causing traffic flow.
- Enhanced Transmission Selection (ETS, IEEE 802.1Qaz) – A bandwidth management mechanism that enables you to allocate port bandwidth in a way that maximizes bandwidth utilization for all flows on the link. ETS allows a port to share and reallocate bandwidth dynamically among its flows while at the same time guaranteeing a minimum amount of bandwidth to every flow.
- Quantized Congestion Notification (QCN, IEEE 802.1Qau) – A congestion management mechanism that sends a congestion notification message through the network to the ultimate source of the congestion. Instead of pausing transmission from the connected peer (as PFC does), QCN tries to stop congestion at its source—the network edge where the "end host" originates the congestion-causing flow. The idea is that instead of pushing a flow control message through the network one device at a time (like PFC), QCN tries to find the cause of congestion and stop the flow at the source.
- Data Center Bridging Exchange Protocol (DCBX)—The mechanism DCB devices use to communicate with each other. Communication includes exchanging DCB state and configuration information (for example, which lanes of traffic PFC is configured to pause), and DCBX can even allow a DCB device to configure a connected peer. DCBX is an extension of Link Layer Discovery Protocol (LLDP, IEEE 802.1AB).

PFC, ETS, and DCBX are mandatory to support lossless transport over Ethernet. QCN is optional and is rarely implemented.

In addition to the three mandatory enhancements, DCB requires:

Buffer
A portion of physical memory used to store data temporarily while data is moved from one location to another. Data is often stored in buffers when it arrives from an input device (such as a connected peer device) and awaits forwarding to its destination.

- 10-Gbps (or greater), full-duplex Ethernet interfaces (device interfaces that do not require DCB can be lower speed).
- Proper buffer management. Although buffer management is not part of the DCB standards, lossless traffic requires sufficient port buffer space to store frames during periods of congestion.

PFC

PFC is an enhancement to a flow control mechanism called Ethernet PAUSE (IEEE 802.3X). However, Ethernet PAUSE does not work well for lossless traffic flows because it pauses all of the traffic on a link during periods of congestion. So one congested flow pauses all of the other flows on the same link, even if those flows are not experiencing congestion.

Because PFC divides a link into eight virtual lanes of traffic (eight priorities), you can choose flows within a physical Ethernet link and pause them individually, without affecting traffic in other virtual lanes. Each lane of traffic (priority) maps to one of the eight IEEE 802.1p code point values in the 3-bit Priority Code Point (PCP) field in the Ethernet frame VLAN header. The code point values identify traffic by priority, and all traffic on a link that requires the same treatment should use the same priority. (For example, all FCoE traffic on a network might use priority 3, which is IEEE 802.1p code point 011.)

IEEE 802.1p Priority
The class-of-service value in the 3-bit Priority Code Point (PCP) field in the Ethernet frame VLAN header (the 802.1Q tag). Priority values range from 0 to 7 (IEEE 802.1p code points 000 through 111). The priority value identifies different types of traffic and allows the switch to differentiate the way it handles different types of traffic.

Devices use the priority to map incoming traffic to class-of-service and DCB properties. Enabling PFC on a priority programs an interface to pause traffic with that priority code point value in its Ethernet VLAN header during periods of congestion. (Traffic that is not paused behaves as normal best-effort Ethernet traffic.) Pausing the traffic on one priority does not affect traffic on other priorities on the link, so lossless and best-effort flows can use the same link without affecting each others' access to port resources.

PFC works through communication between peer devices on directly connected interfaces. When the output queue on the peer receiving the flow fills to a certain threshold, the receiving device asks the sending device to temporarily stop transmitting (pause) the flow. Pausing the flow prevents the queue from overflowing and dropping frames. When the congested queue empties below another threshold, the receiving device asks the sending device to resume transmitting the flow. Devices that support PFC must have port buffers that are deep enough to store frames while the flow is paused.

PFC must be configured on all of the device interfaces in the path of the flows that you want to be lossless. (It doesn't do you any good to pause traffic on one device and let that same traffic drop because of congestion on another device in the path.) PFC uses DCBX to communicate with its directly connected peer.

ETS

ETS manages and shares bandwidth dynamically among the flows (traffic classes) on a port. ETS creates a flexible bandwidth allocation hierarchy by organizing priorities into groups called priority groups. The priorities within each priority group should require similar class-of-service treatment. This is called hierarchical scheduling because it creates a bandwidth allocation hierarchy, as shown in Figure 2.

Priority Group
One or more priorities that are bound together to receive port resource allocations. Priorities in a priority group should have similar traffic handling requirements with respect to latency and frame loss (for example, priorities that require lossless transport can be grouped together, or priorities that require only best-effort transport can be grouped together).

- ETS allocates the available port bandwidth to priority groups (the available port bandwidth is the bandwidth remaining after servicing strict-high priority traffic).
- ETS allocates the bandwidth each priority group receives to the priorities in the group.
- ETS allocates port bandwidth dynamically to priority groups and to the priorities in each priority group, as bandwidth is needed.

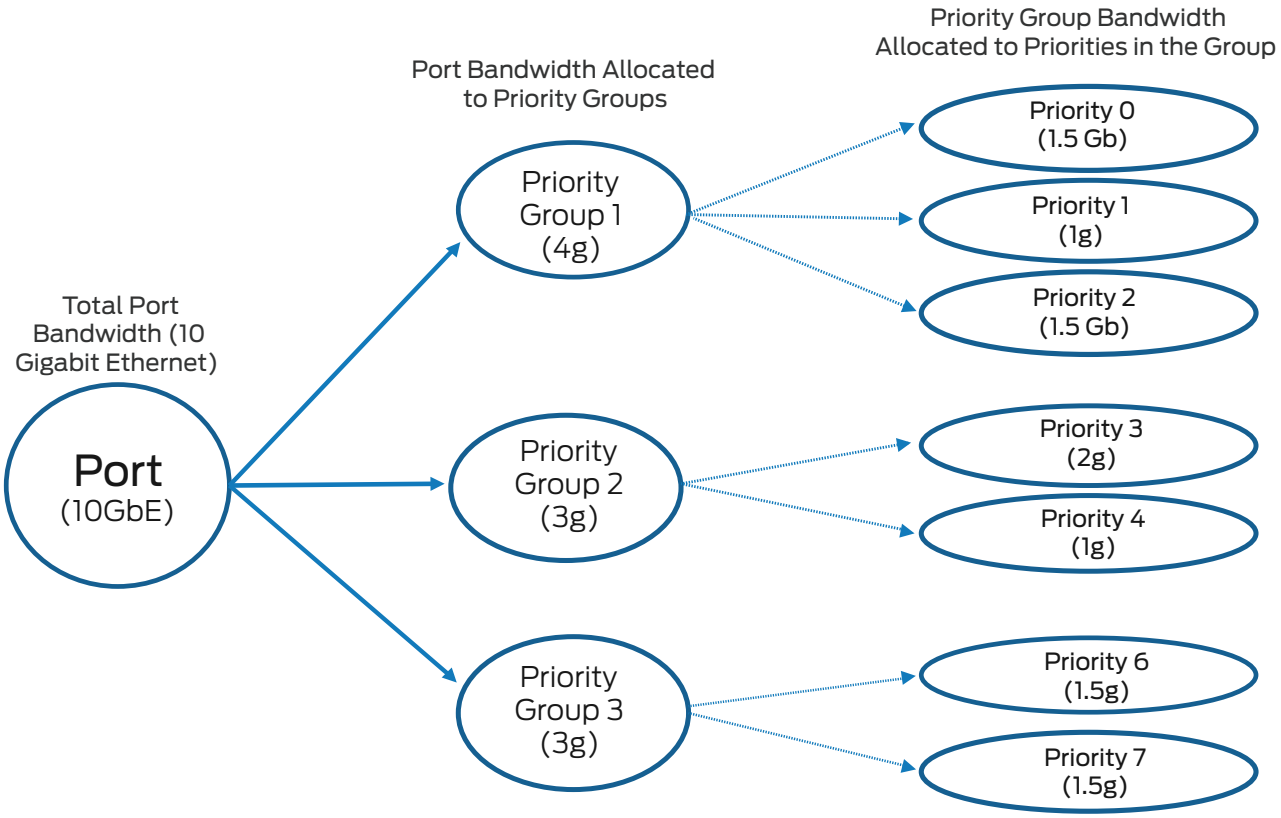

Figure 2 Port Bandwidth Scheduling Hierarchy (Enhanced Transmission Selection)

ETS guarantees a minimum amount of bandwidth to flows, which ensures that storage traffic receives the minimum amount of bandwidth required for lossless transport. However, if a priority group does not need all of its allocated bandwidth, the leftover bandwidth is not wasted because other priority groups on the port can use it. If a priority does not need all of its allocated bandwidth, other priorities within the priority group can use the leftover bandwidth. One benefit of dynamic bandwidth sharing is that bursty traffic can take unused bandwidth from other priorities when the traffic rate is high, and when the burst traffic load is light, other priorities in the priority group can use the leftover bandwidth.

Dynamically sharing bandwidth while guaranteeing minimum amounts of bandwidth to flows at the same time is how ETS increases bandwidth utilization on a port and keeps link throughput high, yet ETS also supports lossless flows while preserving as much port bandwidth as possible for best-effort flows. Essentially, you get to have your cake and eat it too—you guarantee bandwidth for lossless flows, but you don't prevent other flows from using the bandwidth if the lossless traffic isn't using it. Each flow receives the maximum possible bandwidth without impacting flows that require lossless transport.

Available Port Bandwidth The port bandwidth remaining after servicing strict-high priority traffic. For example, if a 10-Gbps port has 1 Gpbs of strict-high priority traffic, the available port bandwidth is 9 Gbps.

QCN

Although few vendors and networks implement QCN today, it is one of the four DCB standards. Unlike PFC, ETS, and DCBX, QCN is not mandatory for supporting lossless transport across Ethernet networks.

The idea behind QCN is to stop congestion at the source. When congestion occurs, QCN traces it back to the network edge device (the "end host") that originated the congestion-causing flow. QCN sends the end host a message to reduce or pause flow transmission, thus shutting off the congestion at its source.

For QCN to work, every device in the network data path must support it. (Otherwise, the chain of communication required for QCN to find the source of the congestion is broken.) QCN works best in situations when congestion is sustained for relatively long periods of time.

However, intermittent congestion can cause practical problems for QCN, which is one reason that few networks use it. One characteristic problem is that by the time a QCN message propagates through the network and finds the source of congestion, and the end host reduces or pauses the flow that caused the congestion, the congestion no longer exists because the congestion was not sustained. So the reduced flow does not relieve congestion because there is no longer congestion to relieve, and the links are underutilized because the flow has been reduced or paused without need.

Another problem occurs if the network includes an FCoE Forwarder (FCF) switch, as converged FC SAN and Ethernet networks often do. QCN does not work when FCoE traffic enters the FC SAN because the FC SAN strips out the Ethernet encapsulation (so the Ethernet headers are lost) when it converts FCoE frames into FC frames. QCN works by identifying the source MAC address of the end host, which is how QCN learns the source of the congestion-causing flow. But for FC frames that have been stripped of Ethernet encapsulation, the FCF replaces the source Ethernet MAC address with the FCF address as the new source MAC address. Because the source MAC address of the flow has changed, QCN cannot identify the source of a flow that causes congestion, so QCN does not work.

Even in networks on which traffic does not enter an FC SAN, the congestion state often changes too fast for QCN to be useful, so depending on network traffic characteristics, QCN has the potential to create more chaos than it cures.

DCBX

In a DCB network, connected peer devices need to know and to negotiate the state of each other's DCB configuration. Each DCB interface uses DCBX to exchange (communicate) the state of its DCB capabilities and applications to its connected peer interface. Because DCBX is an extension of LLDP, if you disable LLDP on an interface, DCBX cannot run on that interface.

Each interface that uses DCBX advertises its configuration for each DCB application (DCBX application protocol TLV exchange). A DCB application is a Layer 2 application (such as FCoE) or a Layer 4 application (such as iSCSI). DCBX can:

- Discover the DCB capabilities of directly connected peers.
- Detect DCB feature misconfigurations or mismatches between directly connected peers. (Not all DCB feature configurations must match to ensure lossless transport. For example, ETS configuration can be different on each peer. However, the PFC configuration must be the same so that the same traffic is treated losslessly on each device.)
- Configure DCB features on directly connected peers (if the peer is configured as "willing" to change its configuration).

Buffer Management

Although buffer management is not part of the DCB standards, it is critical to manage port buffers correctly to ensure that there is enough buffer space to support lossless queues. Without proper buffer management, PFC does not work, because if buffers overflow, frames drop, and transport is not lossless. Port buffers need to be deep enough to store:

- All of the frames sent during the time it takes to send the PFC pause message to the connected peer device.
- All of the frames that are already on the wire (link) when the connected peer receives the pause message and stops transmitting the flow.

The amount of buffer space needed to prevent frame loss due to congestion depends on the propagation delay caused by the length of the cable between the connected peers, the speed of the interface, the size of the frames, and the processing speed of the device. A DCB interface must send the PFC pause message to the connected peer before the congested output queue overfills and drops frames, and soon enough to store the traffic that arrives between the time the interface sends the pause frame and the time the wire is cleared of traffic on the paused priority.

For example, Juniper Networks' QFX Series switches automatically set a threshold for sending PFC pause frames to accommodate delay from cables as long as 150 meters (492 feet) and to accommodate large frames that might be on the wire when the interface sends the pause message. This ensures that the receiving interface sends a pause frame in enough time to allow the sender to pause transmission and the receiver to store the traffic on the wire before the buffers overflow.

Problems Addressed by Using DCB

DCB solves many of the problems experienced when attempting to converge Ethernet and SAN networks:

- Lossless Transport – DCB forces Ethernet networks, which are designed for best-effort traffic, to jump through hoops of fire like a tame circus lion to transport lossless storage traffic. Until the DCB standards were created, there was no practical, cost-effective way to guarantee lossless delivery across an Ethernet network. Because FC storage traffic absolutely requires lossless transport, DCB solved the problem of how to transport FC storage data across an Ethernet network.
- Lower Cost – Using the same Ethernet switches, cabling, and power resources for both standard Ethernet traffic and lossless storage traffic reduces equipment, power, and maintenance costs. Ethernet equipment is also usually less expensive than SAN equipment.
- Easier Management and Maintenance – Maintaining, upgrading, and building out one converged fabric is simpler than managing separate fabrics.
- Traffic Control – ETS controls traffic at a finer-grained level than other methods, so network administrators can fine-tune bandwidth allocations to different traffic types.

Juniper Networks DCB Implementation

The Juniper Networks implementation of DCB standards on data center switches supports the three mandatory DCB enhancements to Ethernet: PFC, ETS, and DCBX. QCN is not supported.

Juniper Networks data center switches enable DCBX by default on an interface if the directly connected peer device also supports DCBX. If the directly connected peer of an interface does not support DCBX, Juniper switches disable DCBX on the interface by default. Juniper Networks data center switches support IEEE DCBX (Organizationally Unique Identifier 0x0080c2) and DCBX version 1.01 (subtype 2, OUI 0x001b21).

Interfaces on which DCBX is enabled automatically negotiate the PFC and ETS administrative state and configuration with the directly connected peer. Also, if an interface carries FCoE traffic and that interface advertises no DCB applications other than FCoE, then DCBX also negotiates the FCoE application state with the connected peer. You can configure DCBX for other applications (for example, iSCSI) that you want DCBX to advertise on a given interface. If you explicitly configure any applications, you must also configure the FCoE application if you want DCBX to advertise it. (The explicit configuration overrides the default FCoE advertisement.)

Juniper Networks data center switch interfaces can use DCBX to program the directly connected peer if the peer is configured as "willing" to be programmed. Juniper Networks data center switches are not "willing" and cannot be programmed by the connected peer.

Juniper Networks data center switches support two lossless priorities (classes of traffic) by default, and can support up to six lossless priorities.

The following are some of the featured data center switching devices offered by Juniper Networks:

- QFX Series Switches—The QFX5100, QFX3500, and QFX3600 data center switches are high-performance, low-latency, 10GbE/40GbE devices that can be used as standalone top-of-rack switches, as Node devices in a QFabric system, and as components of other virtual fabric architectures. QFX Series switches are optimized for virtualized data center environments and support the key DCB standards (PFC, ETS, and DCBX). All of the switches feature deep port buffers to support lossless transport, and redundant components for carrier-class reliability. QFX3500 switches also offer native FC interfaces, so you can configure a QFX3500 switch as an FCoE-FC gateway and connect directly to an FC switch in an FC SAN. (QFX3500 switches do not provide FC services and are not FCFs.)
- QFabric Systems—A QFabric system consists of multiple components working together as a single, high-performance, carrier-class, fabric switching solution. QFabric systems flatten the data center network to a single tier with a single point-of-management for all of the QFabric components, and provide non-blocking, any-to-any (full-mesh) connectivity in the data center. QFabric systems scale from a few hundred ports to more than 6,000 ports, so QFabric can grow with your data center and provide a rock-solid foundation for a cloud-ready, virtualized network. QFabric systems have four types of components:

- Node Devices—QFX Series switches that connect to networked data center devices to provide network access.
- Interconnect Devices—High-speed transport devices that interconnect all QFabric system Node devices in a full-mesh topology.
- Director Devices—Devices that provide control and management services to the QFabric system.
- Virtual Chassis Control Plane—Devices (often Juniper Networks EX4200 switches in a virtual chassis configuration) that provide interconnections to all QFabric system devices and processes.

Summary

DCB transforms the dream of a converged data center network into a reality by delivering the enhancements that Ethernet networks need to support the lossless transport of storage traffic. Thanks to the DCB enhancements, network designers can create converged data centers that not only consolidate lossless storage and best-effort Ethernet traffic onto one network to make management easier, but also save money by using the same devices, cables, power supplies, and other resources for both standard Ethernet traffic and SAN traffic.

References and Suggested Reading

Juniper Networks' QFX Series DCB overview:
http://www.juniper.net/techpubs/en_US/junos13.2/topics/concept/fibre-channel-cee-features-understanding.html

Juniper Networks' QFX Series ETS overview (document includes links to configuration examples):
http://www.juniper.net/techpubs/en_US/junos13.2/topics/concept/cos-qfx-series-schedulers-hierarchical-ets-understanding.html

Juniper Networks' QFX Series flow control overview (document includes links to configuration examples):
http://www.juniper.net/techpubs/en_US/junos13.2/topics/concept/cos-qfx-series-congestion-notification-understanding.html

Juniper Networks' QFX Series DCBX overview (document includes links to configuration examples):
http://www.juniper.net/techpubs/en_US/junos13.2/topics/concept/fibre-channel-dcbx-understanding.html

Juniper Networks' QFX Series and QFabric systems technical documentation page:
http://www.juniper.net/techpubs/en_US/junos13.2/information-products/pathway-pages/qfx-series/13.2X51/index.html

Juniper Networks' QFX Series traffic management page (includes documents on PFC, ETS, and DCBX):
http://www.juniper.net/techpubs/en_US/junos13.2/information-products/pathway-pages/qfx-series/traffic-management.html

IEEE Data Center Bridging Task Group:
http://www.ieee802.org/1/pages/dcbridges.html

IEEE PFC (IEEE 802.1Qbb) home page:
http://www.ieee802.org/1/pages/802.1bb.html

IEEE ETS (IEEE 802.1Qaz) home page:
http://www.ieee802.org/1/pages/802.1az.html

IEEE QCN (IEEE 802.1Qau) home page:
http://www.ieee802.org/1/pages/802.1au.html

IEEE LLDP (IEEE 802.1AB) home pages:
http://www.ieee802.org/1/pages/802.1ab.html

http://www.ieee802.org/1/pages/802.1AB-rev.html

IEEE DCBX version 1.01 specification:
http://www.ieee802.org/1/files/public/docs2008/az-wadekar-dcbx-capability-exchange-discovery-protocol-1108-v1.01.pdf

IEEE DCBX specification (requires IEEE password access):
http://www.ieee802.org/1/files/private/az-drafts/d2/802-1az-d2-4.pdf

Storage Networking Industry Association (SNIA):
http://www.snia.org/

Fibre Channel Industry Association:
http://www.fibrechannel.org/

InterNational Committee for Information Technology Standards (INCITS) T11 Home Page (this is the committee responsible for Fibre Channel interface standards):
http://www.t11.org/index.html

Learn About Data Center Bridging

by Steve Levine

Data Center Bridging (DCB) is a collection of Institute of Electrical and Electronics Engineers (IEEE) enhancements to the IEEE Ethernet standards. When network designers need to transport lossless storage area network (SAN) traffic across Ethernet networks that are designed to drop frames during periods of congestion, they rely on DCB. DCB forces the Ethernet network to play nicely with the storage traffic and give it the lossless treatment it needs, even though the nature of Ethernet networks is to drop frames during periods of congestion, not preserve them.

Steve Levine is a Staff Engineering Technical Writer at Juniper Networks and DCB Diva with more than a quarter century of experience developing technical documentation about networking, semiconductor, computer, and communication technologies.

For more information *see*:
juniper.net/documentation

ISBN: 978-1-936779-81-9 Version History: First Edition, March 2014 2 3 4 5 6 7 8 9

ISBN 978-1936779819
50500
9 781936 779819