Ethernet Enhancements Supporting I/O Consolidation

The drive to consolidate multiple networks

Many IT organizations today operate multiple parallel networks: one for IP networking, one for storage, and in High-Performance Computing (HPC) environments, one for Inter-Process Communication (IPC). These networks cost organizations in terms of the additional capital equipment they require, the cost and complexity of cabling at the rack level, administrative cost, and the additional power and cooling expense imposed by multiple redundant interfaces and transceivers for each server.

I/O consolidation promises to support all three types of traffic on a single networks. One of the primary enablers is 10 Gigabit Ethernet, a technology with the bandwidth and latency characteristics sufficient to support multiple traffic flows on the same link. Standards committees are developing technologies including Fibre Channel over Ethernet (FCoE) that will enable the convergence of both IP and Fibre Channel networks. With FCoE expected to be standardized by the INCITS T11.3 working group in the second half of 2008, I/O consolidation is quickly becoming a reality.

While FCoE can operate on existing Ethernet networks, several enhancements that are currently in progress can improve how well consolidated I/O traffic is handled on the fabric. This brief describes how some of these enhancements will help the network to manage congestion, bursts of traffic, and multiple flows on the same cable.

Fibre Channel over Ethernet

Fibre Channel over Ethernet transports native Fibre Channel frames over an Ethernet infrastructure, allowing existing Fibre Channel management modes to stay intact. One FCoE prerequisite is for the underlying network fabric to be lossless. This requirement can be satisfied using today's Ethernet technology, however enhancements being developed today will make FCoE even more effective in the near future.

The need for Fibre Channel to operate over a lossless network dates back to its heritage as a replacement for SCSI. The Small Computer Systems Interface is a mechanism that provides a low-overhead, high-performance parallel interface that is efficient in handling storage traffic within a chassis. Fibre Channel was developed to overcome the distance and switching limitations inherent in SCSI. Rather than developing an entirely new device-access method, Fibre Channel carries SCSI as its higher-level protocol. SCSI does not respond well to lost frames, which can result in significant delays when recovering from a loss. Because Fibre Channel carries SCSI, it inherits the requirement for an underlying lossless network.

Traditional IP networks manage congestion by dropping packets and allowing upper-level protocols to detect the loss and manage the congestion appropriately. The Fibre Channel protocol uses link-level flow control, where buffer-to-buffer credits control the temporary storage of frames to help absorb congestion. This mechanism results in ports 'borrowing' buffers from upstream ports until the congestion mitigates.

Minimally, Fibre Channel over Ethernet deployments can make use of the IEEE 802.3x PAUSE functionality. The PAUSE control frame can be issued by a congested port to instruct a transmitting port to temporarily halt traffic for a period of time, preventing packets from being dropped.

While the Ethernet PAUSE mechanism can be used to make a network lossless, it can result in congestion trees. If a single downstream port is congested, it may issue a PAUSE frame on several upstream ports, which in turn can cause further upstream ports to issue PAUSE frames when their respective buffers reach a threshold. This behavior leads to the desire to partition traffic so that congestion spreading caused by a protocol such as FCoE doesn't affect other protocols in the network.

Priority Flow Control

Priority Flow Control (PFC), sometimes called Per-Priority PAUSE, is a proposal by Cisco Systems to enable PAUSE capability on the user priorities or classes of service that

are defined by the IEEE 802.1p specification. IEEE 802.1p defines a three-bit field that can be used to create up to eight user priorities within a single physical link.

With the ability to enable PAUSE on a per-user-priority basis, administrators can create lossless lanes for Fibre Channel while retaining packet-drop congestion management for IP traffic. As Figure 1 illustrates, a link divided into eight lanes can use PAUSE on a single lane without affecting traffic on the others. This reduces the overall impact of a PAUSE command because it affects only a fraction of the link's traffic.

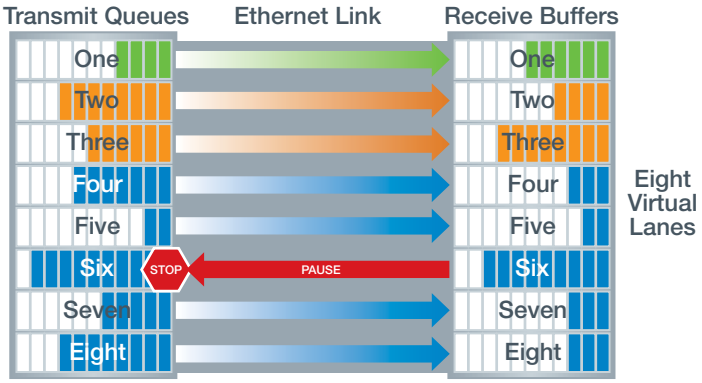


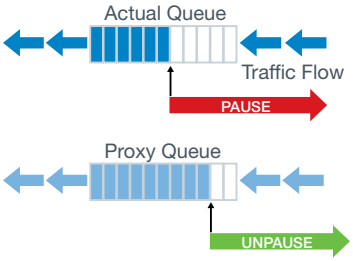
Figure 1: A link using Priority Flow Control can have one lane paused without affecting traffic on the other lanes.

The IEEE 802.1p standard allows allocating resources according to user priority. Products with PFC functionality will allow administrators to assign resources, including buffers and queues based on user priority, resulting in a higher level of service for critical traffic where congestion has the most impact.

Delayed Drop

Delayed Drop is an enhancement being developed by Nuova Systems. Delayed Drop uses the PAUSE mechanism to reduce packet drop on short-lived traffic bursts while triggering upper-layer congestion control through packet drops to handle long-term congestion. This technology allows congestion to spread only for short-term bursts and alleviates it for long-term congestion. Delayed Drop can be individually enabled on each user priority.

As Figure 2 illustrates, Delayed Drop uses a proxy queue to measure the duration of traffic bursts. During normal operation, the proxy queue mimics the actual queue when packets are added or drained. When a burst of traffic causes the actual queue to fill to its high-water mark, that queue issues a PAUSE to stop incoming packets. The proxy queue, meanwhile, simulates the continued receipt of packets. When the proxy queue is filled, the transmitter is unPAUSED, which in turn causes frame drops. This is the condition necessary to stimulate the TCP flow control mechanism for long-lived streams.



Actual Queue	Proxy Queue
Adds a frame	Adds a frame
Issues a PAUSE	Adds frames at line rate
Drains a frame	Drains a frame
Is Empty	Drains frames at line rate

Figure 2: Delayed Drop is implemented using a proxy queue that measures the length of congestion.

During short-term congestion, both queues drain fast enough that the actual queue releases the PAUSE on its own. During long-term congestion, the proxy queue fills to its high-water mark, and it releases the PAUSE. The actual queue begins dropping packets, and the congestion is managed through higher-level protocols.

Congestion Notification

Congestion notification is a form a traffic management that pushes congestion to the edge of a network by notifying upstream rate limiters to shape the traffic causing congestion. As of July 2007, the IEEE 802.1Qau working group unanimously approved a baseline proposal that accepts the Cisco Quantized Congestion Notification (QCN) proposal. This model defines a two-point architecture with a congestion point and a reaction point. In this architecture, congestion is measured at the congestion point and rate-limiting, or back pressure, is imposed on the reaction point to shape traffic and reduce the impact of congestion.

Figure 3 illustrates an aggregation-level switch that has sent control frames to two access-level switches asking them to throttle back their traffic. This maintains integrity of the network's core by impacting only the parts of the network causing the congestion.

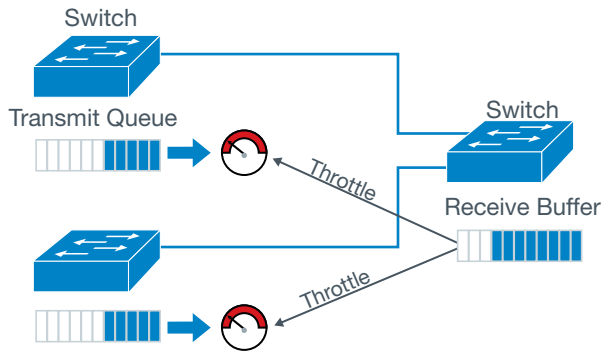


Figure 3: Congestion notification protects the network's core by pushing it to the edge. In this case, a congestion point throttles back two reaction points.

Enhanced Transmission Selection

IEEE 802.1Qaz is a proposed standard that specifies enhanced transmission selection to allocate bandwidth among different traffic classes. When a given load in a traffic class doesn't fully utilize its allocated bandwidth, enhanced transmission selection allows other traffic classes to use the available bandwidth. This helps accommodate the bursty nature of some traffic classes while maintaining bandwidth guarantees.

A real-world example is sharing FCoE, IPC, and classic LAN traffic on the same physical link. For consolidation to be effective, storage traffic must be guaranteed a minimum bandwidth, and traffic requiring low latency must be guaranteed a sufficiently high priority.

Figure 4 illustrates three classes of traffic being offered to a 10 Gigabit Ethernet link. The three classes have three priorities. For lowest latency, HPC traffic is given a 'high' priority. Storage traffic is set to 'medium-high,' and LAN traffic is set to 'medium.' These priorities are generic terms that represent the managed priority objects set by network administrators. In practice, many more adjustments can be made. The three classes of traffic have guaranteed bandwidth levels assigned to them: HPC and FCoE traffic are each assigned 30 percent, and LAN traffic is assigned 40 percent of the link's bandwidth. The offered load is shaped onto the actual link as follows:

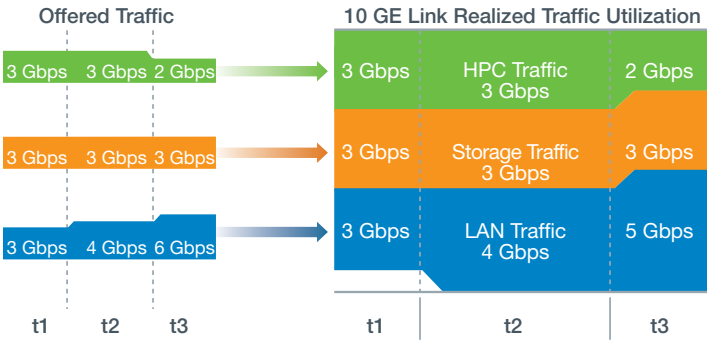


Figure 4: Three classes of traffic with fluctuating bandwidth requirements are shaped onto a single physical link using IEEE 802.1Qaz Enhanced Transmission Selection.

- In time period *t1*, each traffic priority offers a load of 3 Gbps, with each priority offering less than or equal to its guaranteed bandwidth. The result is a 90 percent link utilization during this period.
- In time period *t2*, a burst of LAN traffic brings this category to an offered load of 4 Gbps. The other two categories continue to offer 3 Gbps of bandwidth each, bringing the link utilization to 100 percent.
- In time period *t3*, two events occur: HPC traffic drops to 2 Gbps, and the LAN traffic increases to 6 Gbps. The FCoE traffic remains at 3 Gbps, bringing the offered load to 11 Gbps. At this point, the link is oversubscribed and bandwidth allocation comes into play. The 802.1Qaz standard allows the LAN class of traffic to borrow some unused bandwidth from the higher-priority HPC class. This allows 5 Gbps of LAN traffic, which is less than the offered load, but more than its guaranteed bandwidth.

This Ethernet enhancement ensures that bandwidth allocation is provided while taking into account both traffic priority and available bandwidth.

Data Center Bridging Capability Exchange Protocol

The enhancements presented so far enable I/O consolidation on an Ethernet network. For these enhancements to work seamlessly with existing Ethernet deployments, a management protocol that dynamically discovers the capabilities of its peers is needed. Data Center Bridging Capability Exchange Protocol (DCBCXP) provides the needed configuration and discovery functionality`

Consider the typical Ethernet network depicted in Figure 5. Multiple devices of varying capabilities are connected to each other. In such a network there are two distinct scenarios where dynamic configuration management is required. First, devices need to discover the edge of the enhanced Ethernet cloud. This means that each edge switch needs to learn that it is connected to a legacy switch. Also, servers need to learn

whether or not they are connected to Enhanced Ethernet device. Second, within the Enhanced Ethernet cloud, devices need to discover the capabilities of its peers.

DCBCXP utilizes the link-layer discovery protocol and defines new Type-Length Values (TLVs) for capability exchange parameters. The protocol state machines handle local operational configuration for each feature by comparing and synchronizing with the peer's feature settings. The protocol provides the following capabilities:

- *Enhanced Ethernet-related peer discovery* allows Enhanced Ethernet devices to discover whether a peer device supports particular features such as Congestion Notification or Priority Flow control. Appropriate action can be based on the compatible features.
- *Mismatched configuration detection* can determine if the bandwidth groups and bandwidth allocations on both ends of the link are the same. This information can be used to notify the device manager in case of a conflict.
- *Enhanced Ethernet link peer configuration* can be used by switch devices to push configuration parameters uplink to Server adapters or vice versa.

DCBCXP enables the critical function of discovering the Enhanced Ethernet cloud. It helps individual devices to learn appropriate peer capabilities. Furthermore, it allows for incremental deployment of Enhanced Ethernet components and at the same time it enables seamless operation with legacy Ethernet devices.

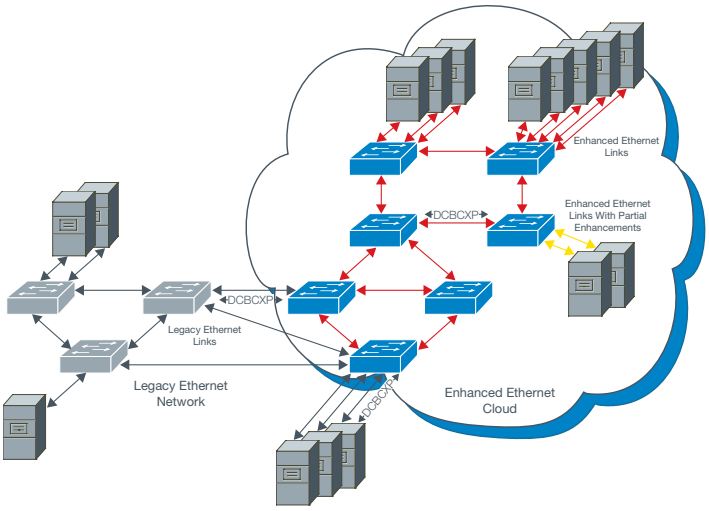


Figure 5: Data center bridging capability exchange protocol (DCBCXP) allows Ethernet Enhancements to work seamlessly with existing networks by determining the extent of the Enhanced Ethernet cloud.

Improving datacenter networks

Although I/O consolidation is possible with today's Ethernet, enhancements proposed by companies including Nuova Systems promise to make it an even more effective I/O consolidation technology in the near future. The enhancements discussed in this brief will improve the ability of datacenter networks to implement FCoE by improving network management, quality of service, and congestion management.

