Unified Fabric:
Cisco's Innovation for Data Center Networks

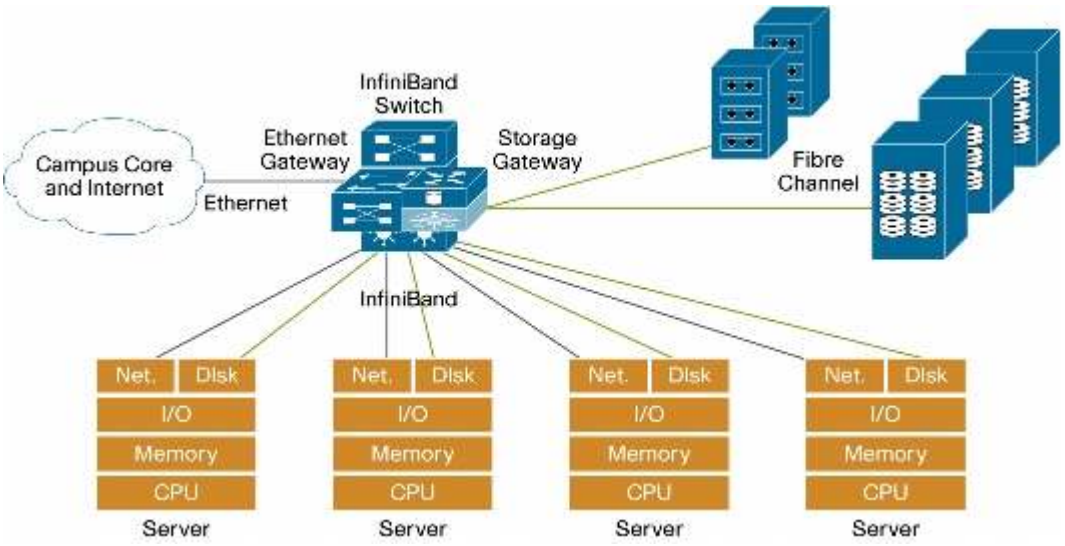
What You Will Learn

Unified Fabric supports new concepts such as IEEE Data Center Bridging enhancements that improve the robustness of Ethernet, which reduce total cost of operations and improves responsiveness by enabling network consolidation. In this document, you will learn about Cisco's commitments in supporting the development of Ethernet based data center networks to provide support for new application demands. You will also see where other approaches have fallen short, and how Unified Fabric promises to be the architecture that fulfills the requirements for the next generation of Ethernet networks in the data center.

Introduction

Ethernet is the predominant network choice for interconnecting resources in the data center. It is ubiquitous and well understood by network engineers and developers worldwide, and Ethernet has withstood the test of time against challengers trying to displace it as the popular option for data center network environments. Emerging demands from applications require additional capabilities in networking infrastructures, resulting in deployment of multiple, separate, application-specific networks. It is common for enterprise data centers to deploy an Ethernet network for IP traffic, one or two storage area networks (SANs) for block mode Fibre Channel traffic, and often an InfiniBand network for high-performance, clustered computing (Figure 1). The combined capital and operating costs for deployment and management of three distinct network types are high, creating an opportunity for consolidation on a unified fabric.

Figure 1. Three Distinct Data Center Networks



When the three types of networks are evaluated technically, Ethernet has the most promise for meeting most of the requirements for all three network types, but it requires some additional capabilities. IEEE Data Center Bridging provides standards-based advancements to Ethernet that enables consolidation of multiple network infrastructures on to a Unified Fabric.

IEEE Data Center Bridging is a collection of architectural extensions to Ethernet that are designed to improve and expand the role of Ethernet networking and management specifically in the data center. There are two main aspects of IEEE Data Center Bridging: extensions to Ethernet supporting I/O consolidation on a unified fabric, with separation and preservation of distinct traffic classes across that fabric; and support for a no-drop service so that traffic requiring guaranteed delivery can be transported over lossless fabrics.

One of the business benefits of network consolidation is cost savings. A homogeneous Ethernet infrastructure would be operationally simpler, tapping into the existing skill set of Ethernet networking engineers and resulting in fewer management tools and shorter startup time for new networks. In addition, a consolidated Data Center network will provide all the existing functions of the Layer 2 networks that it replaces. For Ethernet, this includes support for multicast and broadcast traffic, VLANs, link aggregation, etc.; for Fibre Channel, this includes provision of all the Fibre Channel services such as zoning and name server and support for virtual SANs (VSANs), inter-VSAN routing (IVR), etc.

Evolution of the Data Center Network

Ethernet continues to advance and the data center network continues to evolve, shaped now by the way that applications use the network as a resource. The demands on the network have changed, and it is no longer used simply for traditional client-to-server transactions. For example, deployment of server clusters is increasing, resulting in increased server-to-server traffic. Grid computing also has expanded server-to-server traffic. Increased periodic storage backup requirements have resulted in traffic growth between server farms and storage devices on SANs. In addition, serverless backups between storage devices are now commonplace, increasing disk-to-disk and disk-to-tape traffic. Data center traffic now is being moved from client to server, server to server, server to storage, and storage to storage.

This increase in overall traffic and change in traffic patterns has resulted in greater reliance on the network to deliver the throughput required to support server cluster applications. Now application performance is measured along with performance of the network, which means that bandwidth and latency are both important. There are differences in the types of traffic being sent as well. Client-to-server and server-to-server transactions involve short and bursty transmissions. Most server-to-storage and pure storage applications require long, steady, message flows. This requires a network architecture to be flexible, with network intelligence to support, discover, and respond to changes in network dynamics.

There are also differences in the capability of applications to handle packet drops. Packet drops have unique effects on different protocols, with applications responding in different ways: some applications can tolerate and recover from drops with resends. Ethernet supports these cases, but other applications cannot tolerate any packet loss, requiring some guarantee of end-to-end transmission with no drops. Fibre Channel traffic carried over Ethernet is one example of an application requirement for no-drop service. For Ethernet networks to support applications with no-packet-drop requirements, a method for providing a lossless service class over Ethernet needs to be established. IEEE Data Center Bridging extensions for traffic management provide this capability.

Data center networks also must accommodate large, flat designs. As data center networks continue to expand, with customers adding growing quantities of servers and switches, large existing flat Layer 2 network domains are becoming even larger—this is the norm, not the exception.

Other Options for Network Consolidation

Ethernet is not the only option for data center network consolidation, but it has the greatest chance of succeeding when compared with the other options. Fibre Channel requires a reliable transport with no-drop service during network congestion to avoid retransmission penalties. One challenge to Ethernet was to resolve how to transmit Fibre Channel traffic over Ethernet natively without drops. Priority-based Flow Control (PFC) enables lossless Ethernet, and Fibre Channel over Ethernet (FCoE) allows native Fibre Channel encapsulation.

iSCSI

Ethernet-based Small Computer System Interface over IP (iSCSI) was considered as a replacement for Fibre Channel, allowing consolidation of block storage transfers over Ethernet. Although iSCSI continues to be popular in many storage applications, particularly with small and medium-sized businesses (SMBs), it is not surpassing or replacing the widespread adoption of Fibre Channel for critical storage media transfers in the enterprise. One reason for this is an unwillingness to entrust critical storage media transfer to an Ethernet infrastructure that could not guarantee no-packet-drop service. Another reason that Fibre Channel has not been replaced by iSCSI is iSCSI does not support the native Fibre Channel services or tools on which SAN administrators have come to rely.

Some believe that adding 10 Gigabit Ethernet alone will allow iSCSI to displace FCoE on a pure performance basis. From that perspective, 10 Gigabit Ethernet will also add more speed when FCoE is transported over it. Reliability and integrity of the data are more critical to Fibre Channel than the speed of the link; hence, a lossless Ethernet capability would be attractive. Finally, another challenge with iSCSI has been the TCP/IP overhead that increases the CPU utilization on the server. iSCSI relies on TCP/IP to deliver reliable, in-order storage traffic. TCP/IP offload engines have been applied in network adapter hardware, but this approach can increase the cost of the interface, requiring special application-specific integrated circuits (ASICs).

InfiniBand

InfiniBand also has been positioned as a potential candidate technology for data center network consolidation. InfiniBand requires low latency. Although InfiniBand provides gateways to Fibre Channel and Ethernet networks, it still requires the building of yet another, parallel network, and with Ethernet networks being so pervasive, it is unlikely that IT departments would move their Ethernet-based infrastructures onto InfiniBand; this would not be cost effective, as it would be an incremental buildout and require additional administrative overhead, and it could not be operationally supported without extensive training of the Ethernet IP networking staff to learn InfiniBand. Another hurdle that InfiniBand faces is the lack of interconnectivity between different InfiniBand subnets. As data centers scale and require shared resources, InfiniBand subnets would need to connect to each other with InfiniBand routers that don’t exist today. 10 Gigabit Ethernet is a higher bandwidth standard than 10-Gbps InfiniBand. Since InfiniBand uses 10 bits to encode 8 bits of data, there is a 20 percent loss in line rate, limiting the usable bandwidth at the data link layer to 8-Gbps. 10 Gigabit Ethernet can send 10-Gbps of good throughput, achieving the actual line rate of a 10-Gigabit Ethernet link.

Considering that 80 percent of all Server Clusters are performed across Ethernet infrastructures today, the most likely option is that Ethernet can be enhanced to serve a larger percentage of Server Cluster and grid computing applications. Creating Remote Direct Memory Access (RDMA) drivers for 10 Gigabit Ethernet is one development that seems inevitable. Low-latency Ethernet with high throughput directly connecting memory resources is required. The availability of the lossless transport class in IEEE Data Center Bridging also will be beneficial for Server Cluster applications.

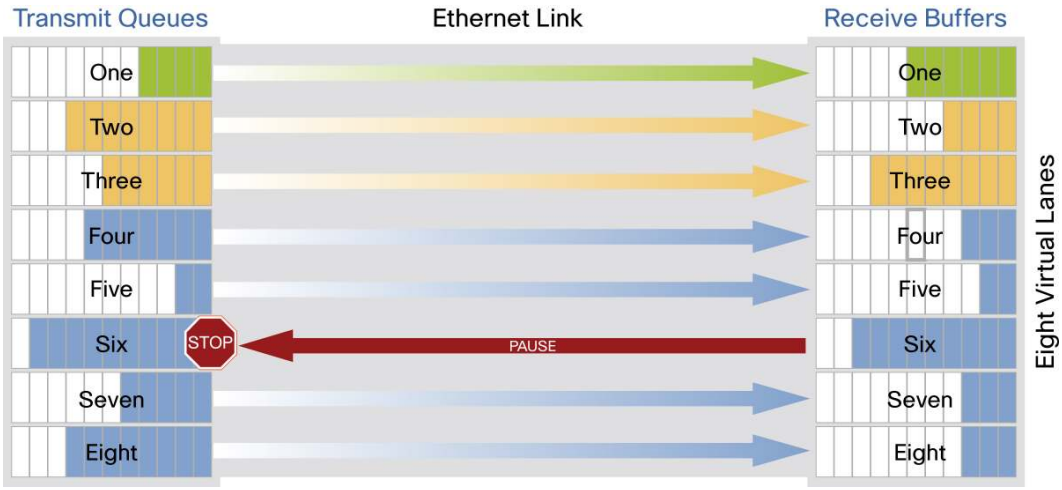
IEEE Data Center Bridging

IEEE Data Center Bridging has been well thought out to take advantage of classical Ethernet's strengths, add several crucial extensions to provide the next-generation infrastructure for data center networks, and deliver the unified fabric promised in the Cisco Data Center 3.0 architecture. The remainder of this document outlines IEEE Data Center Bridging, and describes how each of the main components of the architecture contributes to a robust Ethernet network capable of meeting today's growing application requirements and responding to future data center network needs.

Priority-based Flow Control: IEEE 802.1Qbb

Link sharing is critical to I/O consolidation. For link sharing to succeed, large bursts from one traffic type must not affect other traffic types, large queues of traffic from one traffic type must not starve other traffic types' resources, and optimization for one traffic type must not create large latency for small messages of other traffic types. The Ethernet pause mechanism can be used to control the effects of one traffic type on another. Priority-based Flow Control (PFC) is an enhancement to the pause mechanism (Figure 2).

Figure 2. Priority-based Flow Control



The current Ethernet pause option stops all traffic on a link; essentially it is a link pause for the entire link. PFC creates eight separate virtual links on the physical link and allows any of these links to be paused and restarted independently. This approach enables the network to create a no-drop class of service for an individual virtual link that can coexist with other traffic types on the same interface. PFC allows differentiated quality-of-service (QoS) policies for the eight unique virtual links. PFC also plays a primary role when used with an arbiter for intraswitch fabrics, linking ingress ports to egress port resources (see "Lossless Fabric" later in this document; IEEE 802.1Qbb and http://www.ieee802.org/1/files/public/docs2007/new-cm-barrass-pause-proposal.pdf).

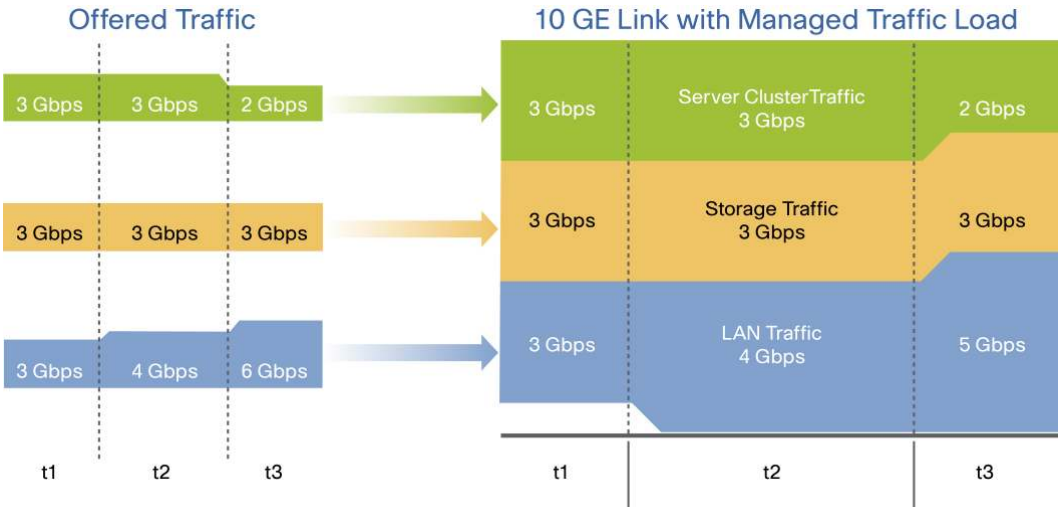
Enhanced Transmission Selection: IEEE 802.1Qaz

PFC can create eight distinct virtual link types on a physical link, and it can be advantageous to have different traffic classes defined within each virtual link. Traffic within the same PFC IEEE 802.1p class can be grouped together and yet treated differently within each group. Enhanced Transmission Selection (ETS) provides prioritized processing based on bandwidth allocation, low latency, or best effort, resulting in per-group traffic class allocation. Extending the virtual link concept, the network interface controller (NIC) provides virtual interface queues: one for each traffic class. Each virtual interface queue is accountable for managing its allotted bandwidth for its traffic group, but has flexibility within the group to dynamically manage the traffic. For example, virtual link 3 for the IP class of traffic may have a high-priority designation and a best effort within that same class, with the virtual link 3 class sharing a percentage of

the overall link with other traffic classes. ETS allows differentiation among traffic of the same priority class, thus creating priority groups (Figure 3).

Today's IEEE 802.1p implementation specifies a strict scheduling of queues based on priority. With ETS, a flexible, drop-free scheduler for the queues can prioritize traffic according to the IEEE 802.1p traffic classes and the traffic treatment hierarchy designated within each priority group. The capability to apply differentiated treatment to different traffic within the same priority class is enabled by implementing ETS (see IEEE 802.1Qaz http://www.ieee802.org/1/pages/802.1az.html).

Figure 3. Enhanced Transmission Selection

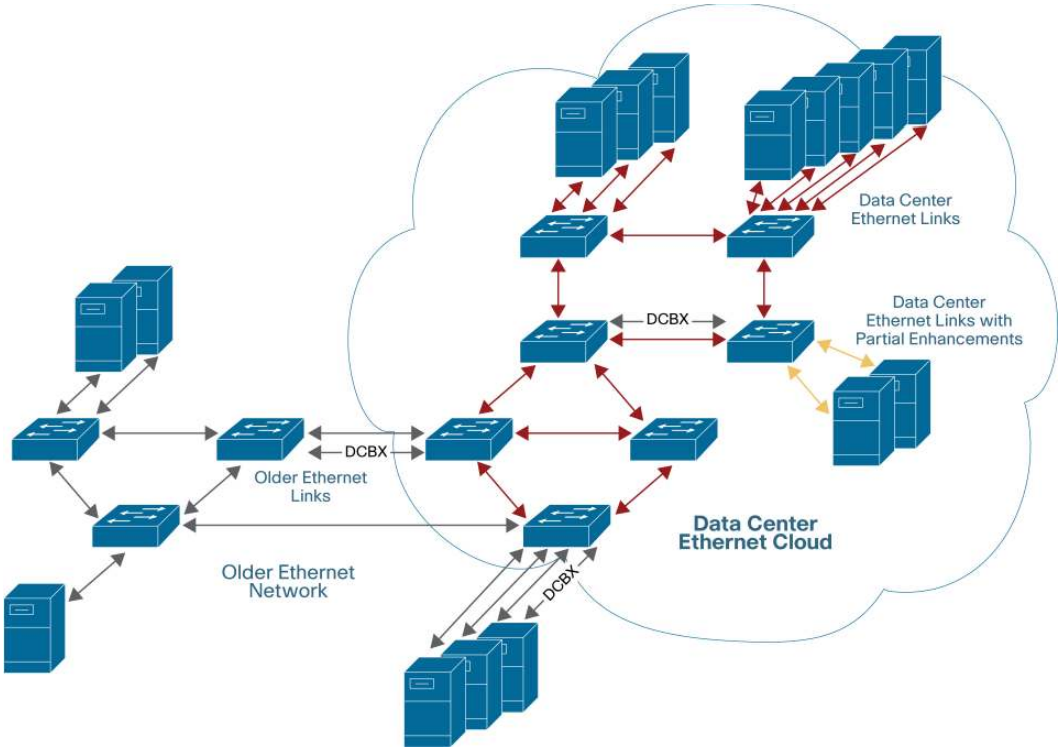


Data Center Bridging Exchange Protocol

Data Center Bridging Exchange (DCBX) Protocol is a discovery and capability exchange protocol developed by Cisco, Nuova, and Intel that is used by IEEE Data Center Bridges to discover peers and exchange configuration information between DCB compliant bridges (Figure 4). The following parameters of the IEEE Data Center Bridge can be exchanged with DCBX (see http://www.ieee802.org/1/files/public/docs2008/az-wadekar-dcbcxp-overview-rev0.2.pdf):

- Priority groups in ETS
- PFC
- Congestion Notification
- Applications
- Logical link-down
- Network interface virtualization

Figure 4. Data Center Bridging Exchange Protocol

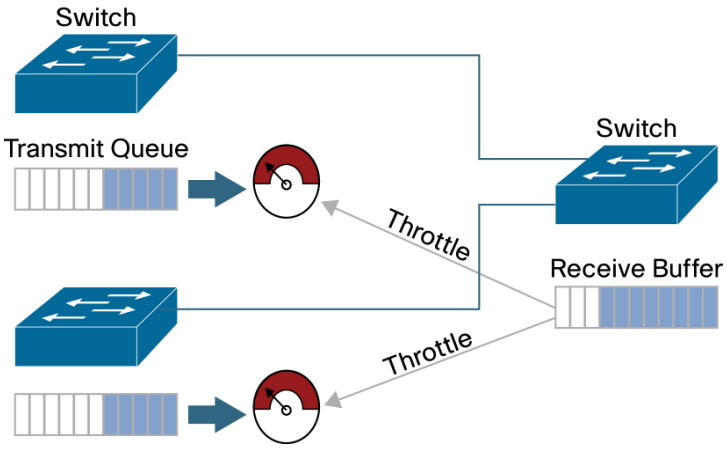


Congestion Notification: IEEE 802.1Qau

Congestion Notification is traffic management that pushes congestion to the edge of the network by instructing rate limiters to shape the traffic causing the congestion. The IEEE 802.1Qau working group accepted the Cisco proposal for Congestion Notification, which defines an architecture for actively managing traffic flows to avoid traffic jams.

Congestion is measured at the congestion point, and if congestion is encountered, rate limiting, or back pressure, is imposed at the reaction point to shape traffic and reduce the effects of the congestion on the rest of the network. In this architecture, an aggregation-level switch can send control frames to two access-level switches asking them to throttle back their traffic (Figure 5). This approach maintains the integrity of the network's core and affects only the parts of the network causing the congestion, close to the source (see IEEE 802.1Qau http://www.ieee802.org/1/pages/802.1Qau.html).

Figure 5. Congestion Notification



Cisco Unified Fabric Related Standards and Enhancements

In addition to IEEE Data Center Bridging, Cisco Nexus data center switches includes other enhancements, such as standards-based layer-2 multi-pathing and Fibre Channel over Ethernet (FCoE), as well as a lossless fabric to enable a Unified Fabric to be constructed. These concepts are briefly discussed below.

Layer 2 Multi-pathing

Equal-cost multi-path routing is performed today at Layer 3. Standards bodies are proposing several alternatives for achieving equal-cost multi-pathing at Layer 2. Transparent interconnection of lots of links (TRILL) is a solution proposed in the IETF standards group, and shortest-path bridging (IEEE 802.1Qat) is being evaluated by IEEE.

Layer 2 multi-pathing (L2MP) increases bisectional bandwidth by enabling multiple parallel paths between nodes, resulting in higher bandwidth in the interconnect network with lower latencies. Based on the traffic patterns of large server farms, L2MP will augment the performance of these networks. Load balancing of traffic among alternative equal-cost paths will improve application performance and network resiliency. IEEE Data Center Bridging with L2MP will enable use of all available connections between nodes, eliminate single-path constraints, and enable data center operators to make dynamic topology changes without worrying about convergence effects when a path is removed or added.

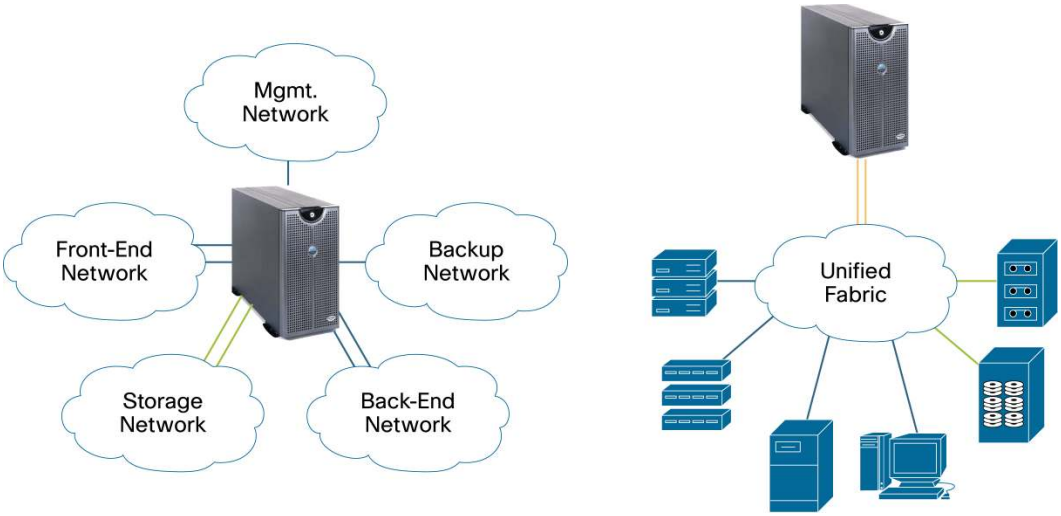
I/O Consolidation

The continuing expansion of 10 Gigabit Ethernet supports mixing of traffic types between servers and switched networks. IEEE Data Center Bridging extensions (PFC, ETS, DCBX, and Congestion Notification) enable a 10 Gigabit Ethernet connection to support multiple traffic types simultaneously, preserving the respective traffic treatments. With these extensions, the same 10 Gigabit Ethernet link will also be able to support Fibre Channel storage traffic by offering a no-drop capability for FCoE traffic. Consolidated I/O at the server supporting FCoE allows all hosts to have access to storage resources over the common unified fabric.

Changes in server architectures are influencing the move to Unified Fabric. The adoption of Peripheral Component Interconnect Express (PCI-Express) over PCI and PCI-X has enabled servers to overcome the I/O bottleneck at the PCI bus. This fundamental change allows servers to utilize a full 10 Gigabit Ethernet interface. At the same time, servers are using higher-density chips, quad cores, and multiprocessor platforms, resulting in more demand for greater bandwidths into and out of the servers. With multiple processors, cores, and virtual machines existing on single servers, 10 Gigabit Ethernet will be more widely adopted, and a method for managing multiple traffic types simultaneously will be critical to allow traffic sharing on larger consolidated I/O connections.

A consolidated I/O link can present multi-protocol traffic to a unified fabric on a single cable. A unified fabric is a single, multipurpose Ethernet transport that can transmit IP and Fibre Channel traffic simultaneously across the same interface and the same switch fabric preserving differentiated classes of service. Use cases include multi-protocol transport, FCoE, and RDMA over low-latency Ethernet.

Figure 6. Unified Fabric



When a product has implemented the required IEEE Data Center Bridging standards-based specifications, it should be compatible with other products that have implemented the same set of specifications. The minimal requirements for Data Center Bridging endpoints (hosts, targets, servers, etc.) are PFC, ETS, and DCBX. A Cisco data center bridging capable switch, in addition to IEEE data center bridging, also supports a lossless intra-switch fabric architecture capable of providing a no-drop service and L2MP.

Lossless Fabric

Although not defined within IEEE Data Center Bridging, a data center switch must provide a lossless architecture to ensure that the lossless transmission service class will not drop a frame. To support FCoE, a lossless fabric is mandatory to help ensure that storage traffic has no-drop service. To create a lossless Ethernet fabric with multi-protocol support, two elements are required: a priority-based pause mechanism (PFC) and an intelligent switch fabric arbitration mechanism that ties ingress port traffic to egress port resources to honor any pause requirements.

Today's standard pause mechanism in Ethernet halts all traffic types on the link. PFC is crucial for I/O consolidation as it creates up to eight separate logical links over the same physical link, thus allowing any of the eight traffic types to be paused independently while allowing the other traffic types to flow without interruption. PFC uses a priority-based pause mechanism to select the IEEE 802.1p traffic type to be paused on a physical link. The capability to invoke pause for differentiated traffic types enables the traffic to be consolidated over a single interface.

PFC provides a no-drop option on each logical link with its capability to halt independent logical traffic types. PFC (as well as the standard pause mechanism) makes a link lossless, but that is not enough to make a network a lossless fabric. In addition to no-drop service on the link, a way to tie the ingress port pause behavior to the egress port resources is required across the intra-switch fabric using PFC. To make the network lossless, each switch needs to associate the resources of the ingress links with the resources of the egress links. Logically tying egress port resource availability to the ingress port traffic allows arbitration to occur to help ensure that no packets are dropped—which is the definition of a lossless switch fabric architecture. This lossless Ethernet intra-switch fabric behavior provides the required no-drop service that emulates the buffer credit management system seen in Fibre Channel switches today.

Fibre Channel over Ethernet (FCoE)

To transport Fibre Channel storage traffic or any other application that requires lossless service over an Ethernet network and achieve a unified fabric, a lossless service class is required. Fibre Channel storage traffic requires no-drop capability. A no-drop traffic class can be created using IEEE Data Center Bridging and a lossless Ethernet switch fabric.

The FCoE protocol maps native Fibre Channel frames over Ethernet, independent of the native Ethernet forwarding scheme. It allows an evolutionary approach to I/O consolidation by preserving all Fibre Channel constructs.

INCITS T11 is writing the standard for FCoE. This group will mandate that a lossless Ethernet network is required to support FCoE. Although the standard pause mechanism (as well as PFC) makes a link lossless, pause or PFC alone is not enough to make a network lossless. To make the network lossless, each switch needs to correlate the buffers of the incoming links with the buffers of the egress links and tie them to the pause implementation. This is a capability of a platform architecture that has nothing to do with protocols. The Cisco® Nexus 5000 Series Switches provide this capability today. The Cisco® Nexus 7000 Series Switches will provide this lossless network capability in the future.

Unified Fabric is designed primarily to enhance data center networks and most of the enhancements will be offered in Cisco products that are found in the data center. Since Unified Fabric has numerous components—IEEE Data Center Bridging and IETF TRILL—at least a portion of the enhancements will likely be added to certain platforms. In addition, not all IT departments will want to run a lossless service over a converged Ethernet network or deploy a Unified Fabric, and in these cases classical Ethernet will continue to be deployed as it is today.

Conclusion

Unified Fabric delivers the architecture that meets the promise of the Cisco Data Center 3.0 vision today. Ethernet is the obvious choice for a single, converged fabric that can support multiple traffic types. Ethernet has a broad base of global engineering and operational expertise from its ubiquitous presence in data centers worldwide. FCoE is the first use case for a Unified Fabric. IEEE Data Center Bridging extensions delivers lossless Ethernet capability that accommodates Fibre Channel's no-drop requirement.

The creation of a lossless service over an Ethernet fabric benefits FCoE, RDMA, iSCSI and applications such as real-time video to get delivery guarantees from no-drop virtual links mixed with other contending applications. The L2MP innovation will benefit any data center Ethernet network, regardless of whether SAN traffic is converged on it. Increased bandwidth and lower latency with L2MP will yield performance gains for all applications. IEEE Data Center Bridging extensions, lossless capability, and L2MP are provided in a family of next-generation switches, from Cisco including the Cisco Nexus 7000 and Cisco Nexus 5000 Series, integral pieces of the data center unified fabric architecture.

With these new data center extensions, IT departments gain several benefits: a new flexible method for consolidating I/O over Ethernet on the same network fabric, as opposed to supporting separate networks; a method for delivering a lossless traffic class on Ethernet; and greater utilization of bandwidth between nodes using equal-cost multi-pathing at Layer 2 for higher bisectional traffic support. The way that IT departments choose to consolidate differentiated traffic types over the same interface, cable, and switch will evolve over time. Unified Fabric offers the flexibility to choose what to run over a consolidated interface, link, and switch fabric and when to make that move.

For More Information

Visit

IEEE Data Center Bridging: http://www.cisco.com/en/US/netsol/ns783/index.html

Priority Flow Control IEEE 802.1Qbb: http://www.ieee802.org/1/pages/802.1Qbb.html

Enhanced Transmission Selection IEEE 802.1Qaz: http://www.ieee802.org/1/pages/802.1Qaz.html

Congestion Notification IEEE 802.1Qau: http://www.ieee802.org/1/pages/802.1Qau.html

DCBX explanation: http://www.ieee802.org/1/files/public/docs2008/az-wadekar-dcbcxp-overview-rev0.2.pdf

Related Standards

IETF Transparent Interconnection of Lots of Links (TRILL): http://www.ietf.org/html.charters/trill-charter.html

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV
Amsterdam, The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at www.cisco.com/go/offices.

Printed in USA C11-462422-01 07/09