DCB Whitepaper

June 17, 2010

Contributors:

Gary Gumanow, Dell
Charlie Lavacchia, NetApp
Chauncey Schwartz, Qlogic
Manoj Wadekar, QLogic

Ethernet Alliance
3855 SW 153rd Drive
Beaverton, OR 97006

www.ethernetalliance.org

Table of Contents

1. Introduction 2
2. Current Ethernet Limitations 4
3. Can Current Ethernet Limitations Be Addressed 6
Can current Ethernet be made lossless? 6
Can traffic differentiation and Quality of Service be delivered over current Ethernet? 6
4. Data Center Bridging Activity 7
5. Specific Standards Addressing Ethernet Limitations 8
802.1Qaz: Priority Groups (Enhanced Transmission Selection) 9
802.1Qaz: DCB Exchange Protocol 9
802.1Qau: Congestion Notification (Persistent Congestion Control) 9
802.1Qbb: Priority-based Flow Control (Transient Congestion Control) 10
Comparing Fibre Channel to Ethernet Flow Control Mechanisms 11
Ethernet Flow Control Mechanisms: 11
Fibre Channel Flow Control Mechanisms: 12
6. DCB IEEE 802.1 Schedules and Timelines 12
7. Use Case: Storage over Ethernet 13
Server-to-storage protocols 13
Strengths and limitations of existing server-to-storage protocols 13
Ethernet Storage 13
DCB Advances Ethernet Storage 14
8. Summary 15

Introduction

The need for digitized data is pervasive across all vertical markets in industries including government and education, financial services, automotive and manufacturing, bio-technology, health care, and high-tech. The ever-increasing need for digital storage requires data centers to strive for the most efficient connection to, processing of, and management of information. The requirement for lower cost, highly available and more efficient networks that provide access between users, servers, and data storage continues to grow unabated.

Until very recent times, data center managers were forced to deploy different types of networks for user-to-server, server–to-server and server-to-storage networks. Each of these network types has its own unique components, management software and administrative training requirements that, taken collectively, diminish opportunities to dramatically improve data center efficiencies in such key areas as performance, space, power costs, and administrative productivity. The increasing demand for storage and the requirements for ready access to information continue to drive the explosion of these disparate networks. At the access layer, network proliferation has increased IT costs with the need for multiple adapters to connect separate cables to the various networks found in the data center. This increase in I/O connectivity requirements has led to the growth of cable and server costs which in turn has increased power and cooling requirements in the data center. Power and cooling of data center equipment is one of the largest annual data center costs, often exceeding the capital equipment expenditure for the equipment itself.

The three most common types of networks found in the data center are:

- Storage Area Networks (SAN)
- Local Area Networks (LAN)
- Inter-process Communication Networks (IPC)

As described earlier, each network evolved uniquely to solve a specific requirement delivering specific characteristics for its unique traffic type:

- The SAN network uses Fibre Channel technology to provide a deterministic, in-order, guaranteed delivery to be sent to/from storage devices.
- The LAN network provides traditional TCP/IP based Ethernet network for best effort data communications.
- The IPC network uses a High Performance Computing (HPC) clustered environment, where multiple servers communicate with each other using high speed, low latency messaging.

IT managers today realize the continued cost and complexity of managing multiple networks is not a sustainable long-term solution. They are now looking for solutions that will enable them to consolidate all of their disparate network traffic onto one consolidated network.

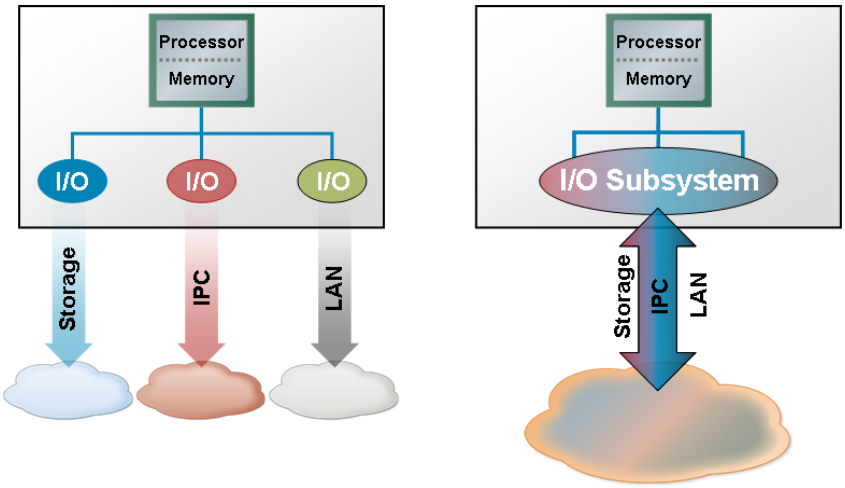

Figure 1 – Today's data center with multiple networks and with a consolidated network

This network I/O consolidation simplifies the data center and significantly saves on capital, as well as, operational expenditures. A consolidated network infrastructure gives IT managers the opportunity to:

- Reduce costs of deployment and operations (Manage one fabric, Ethernet)
- Reduce hardware costs (cables, interfaces, and switches)
- Deploy much simpler, more manageable cabling infrastructure
- Decrease network congestion simplifying server room management
- Reduce power and cooling requirements

When the underlying bases of the three types of data center networks are evaluated technically, Ethernet has the most promise for meeting most if not all of the unique requirements for all three network types. 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 and has been deployed for over twenty five years.

This paper highlights the fundamental enhancements that Ethernet is undergoing in an industry wide move to provide the medium for a single network to be able to transport SAN, LAN and IPC data effectively.

Current Ethernet Limitations

Ethernet was initially designed to provide end-user access to compute resources over networks of unknown and unpredictable reliability and distance. As designed, Ethernet also depended upon higher level protocols such as TCP to provide for flow control, congestion and error recovery. As the

underlying physical components of Ethernet networks became capable of operating at increasingly higher levels of performance, i.e. from Kilobits to Gigabits per second, Ethernet continued to evolve to support these higher speeds while continuing to depend on the higher layers for these functions.

Most significant among these limitations are:

- Ethernet provides all classes, and/or types of traffic the same access to bandwidth. Ethernet has for many years provided mechanisms for Quality of Service (QoS) (IEEE 802.1p), but did not go far enough to distinguish between types or classes of traffic. A telephony application or customer support agent might have the same priority as bandwidth for a mission critical application or as a high priority file transfer. Lacking this level of QoS, data center managers must either over provision network bandwidth for peak loads, accept customer complaints during these periods, or manage traffic prioritization at the source side by limiting the amount of non-priority traffic entering the network.

- Fibre Channel (FC) provides a "buffer-to-buffer credit" that ensures pakcets will not be dropped due to congestion in the network. However, Ethernet's only mechanism for providing such "lossless" transport of traffic is to either use link-level flow control (802.13x) or leave the retransmission to higher protocols, like TCP. Proposed storage convergence protocols like FCoE need "lossless" characteristics for the Traffic Class over this FCoE traffic is being carried. Applying 802.3x PAUSE mechanism to whole link is not appropriate – because even though it will help FCoE, it may not be appropriate for other traffic classes (e.g IPC traffic) that can share same converged IO pipe" Ethernet utilizes upper level layer protocols (TCP) to manage end-to-end data delivery and integrity. When the amount of data entering the network exceeds network capacity Ethernet networks become what is known as over-subscribed, and will "drop" data frames in certain circumstances. The upper-level protocols, TCP for example, request the dropped data packets to be retransmitted. These retransmits happen quickly (25 milliseconds) but can contribute to the lack of consistent response times. These inconsistencies limit Ethernet's ability to service applications that are either response time sensitive i.e. customer facing on-line transaction processing systems (OLTP) or applications that are dependent on isochronous or near-isochronous communications such as video streaming over distance.

Overcoming these limitations is the key to enabling Ethernet as the foundation for true converged data center networks supporting all three types of data center networks and their unique requirements described earlier.

Can Current Ethernet Limitations be Addressed

Can current Ethernet be made lossless?

Ethernet is a robust networking technology. if traffic burdens and usage remained constant, Ethernet networks could be planned and designed to not drop a single packet due to congestion. However, networks do not remain constant. Increasing need for access and increasing demand for new services has wreaked havoc on network managers' best intentions for a planned, well designed network.

Ethernet and Fibre Channel switch implementations have finite buffers and must be designed to cope with full buffer situations during periods of high data traffic. Fibre Channel switches are designed using a buffer credit mechanism to eliminate dropped packets. In contrast, it is common for Ethernet switches to be over-subscribed for the stated bandwidth of the switch, and to drop packets when they run out of buffer space.

The IEEE 802.3x PAUSE mechanisms can be used to provide "flow control" of actual data traffic, but the link level nature of PAUSE can cause head-of-line blocking and other fairness concerns. The PAUSE mechanisms alone provide no ability to treat certain traffic differently.

Higher level protocols such as TCP/IP have adapted to the intent of IEEE 802 based networks by incorporating end-to-end congestion avoidance and flow control algorithms (window based protocol, etc.).

Can traffic differentiation and Quality of Service be delivered over current Ethernet?

The existing IEEE 802.1p/Q standards provide classification of traffic flows with 3 bit tagging. This classification is used by network devices (like bridges, routers) to queue different class traffic into different queues. The standard specifies strict priority scheduling of these queues. This allows higher priority traffic to be serviced before the lower priority queues – thus achieving lower latency and also lower drop probability for priority traffic; however, this creates unfairness issues for other queues because higher priority queues use more bandwidth starving the lower priority queues. The standard does permit the use of other scheduling algorithms. However, because the specific behavior is not specified, no single implementation exists resulting in interoperability issues.

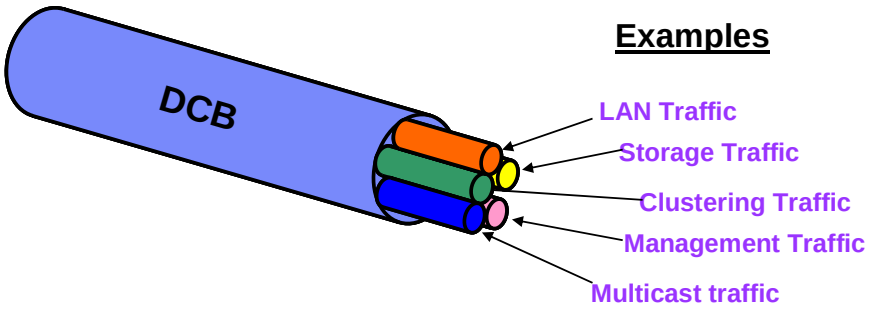

Figure 3 - Traffic differentiation

Data Center Bridging Activity

To accommodate the upper layers of network protocols that need more stringent delivery requirements such as Fibre Channel to run over Ethernet networks, the IEEE 802.1 Working Group created the Data Center Bridging (DCB) task force.

This task force is focused on enhancing Ethernet to address the key issues identified in the preceding paragraphs. The task force is defining a collection of Ethernet architectural extensions designed to improve and expand the role of Ethernet networking and management specifically in the data center. There are four main functional areas for these architecture extensions:

- Congestion Notification (CN) provides end to end congestion management for upper layer protocols that do not already have congestion control mechanisms built in; e.g. Fibre Channel over Ethernet (FCoE). It is also expected to benefit protocols such as TCP that do have native congestion management. These enhancements will help make Ethernet a "lossless" network technology

- Priority-based Flow Control (PFC) provides a link level flow control mechanism that can be controlled independently for each priority. The goal of this mechanism is to ensure zero loss due to congestion in DCB networks. These enhancements will help deliver Quality of Service over Ethernet.

- Enhanced Transmission Selection (ETS) provides a common management framework for assignment of bandwidth to different traffic classes. These enhancements will also assist in providing Quality of Service over Ethernet.

- A discovery and capability exchange protocol is used for conveying capabilities and configuration of the above features between neighbors ensuring a consistent configuration across the network. This protocol is expected to leverage current functionality provided by 802.1AB (LLDP).

Data Center Bridging networks (bridges and end nodes) will be characterized by limited bandwidth-delay and limited hop-count.

Specific Standards Addressing Ethernet Limitations

802.1Qaz: Priority Groups (Enhanced Transmission Selection)

When multiple traffic types are being transmitted over a single link, there needs to be assurance that each traffic type obtains the bandwidth that has been allocated for it; e.g. for a server that has LAN, SAN and IPC connections and they are being consolidated on a single Ethernet link, data center management should be able to guarantee and provision the bandwidth allocated to a particular traffic type, like SAN traffic, while at the same time conditionally restraining those traffic types from exceeding their allocated bandwidth.

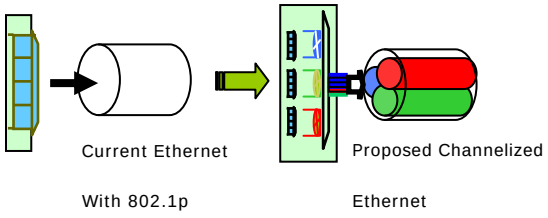

Figure 4 - Channelized Ethernet Link: Virtual Interfaces

When multiple traffic types are consolidated onto a single link there is no inherent prioritization of traffic between these types; however, each traffic type needs to maintain their current usage model of a single interface with multiple traffic classes supported. Each type also needs to maintain its bandwidth allocations for a given virtual interface (VI) independent of traffic on other VIs. Data Center Bridging physical links provide multiple virtual interfaces for different traffic types – LAN, SAN, IPC etc. Each traffic type will map into this virtual interface comprised of multiple traffic classes and each virtual interface will be guaranteed its own bandwidth. These virtual interfaces do not need a separate identifier on the physical link. the virtual interfaces are constructed by each device associating multiple 802.1p User Priority values to a single virtual interface. The following table shows an example.

	Application	User Priorities	*BW Share*
VIF 1	LAN Traffic	0, 3,4,5	*40%*
VIF 2	SAN Traffic	1, 6	*40%*
VIF 3	*IPC Traffic*	*2, 7*	*20%*

Table 1 - Virtual Interface configuration on Data Center Ethernet Link

By providing bandwidth guarantees to various traffic types on a converged link, data center managers will be able to maintain the low latency behavior certain traffic types require.

This channelization of physical links allows various traffic types to map to a virtual interface and be configured according to each group's requirements. For instance, SAN traffic is mapped to Virtual Interface 2 in the example above. All User Priorities in Virtual Interface 2 are guaranteed to have Congestion Management features enabled to provide "no drop" behavior.

802.1Qaz: DCB Exchange Protocol

The new features and capabilities of Data Center Bridging will need to operate within and across multiple network domains with varying configurations. Achieving interoperability across these environments requires that link partners exchange information about their capabilities and configuration and then select and accept feature configurations with their link partners.

This aspect of the 802.1Qaz standard defines the configuration of link parameters for Data Center Bridging functions and includes a protocol to exchange (send and receive) Data Center Bridging parameters between peers; to set local "operational" parameters based on the received Data Center Bridging parameters; and to detect conflicting parameters.

802.1Qau: Congestion Notification (Persistent Congestion Control)

This standard specifies protocols, procedures and managed objects that support congestion management of long-lived data flows within network domains supporting services that require limited bandwidth delay i.e. high-speed short-range networks such as data centers, backplane fabrics, single and multi-chassis interconnects, computing clusters, and storage networks.

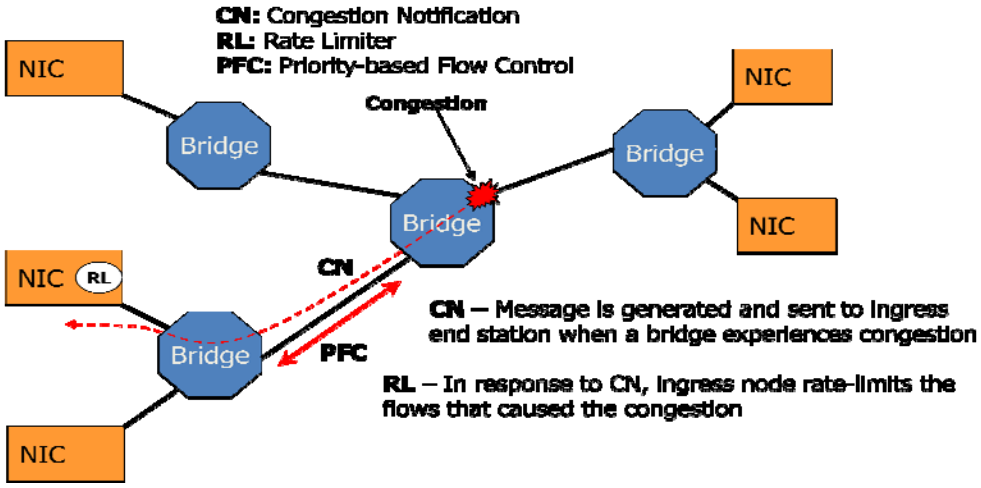

Figure 5 - Congestion Notification

802.1Qbb: Priority-based Flow Control (Transient Congestion Control)

Although the Congestion Notification mechanism provides control of queue at the CP and reduces packet loss due to overflow, it is not guaranteed that this mechanism will ensure zero packet loss in case of congestion. This is due to the fact that control loop delay for the congestion notifications causes variable amounts of data to be on the wire before sources can start acting upon the congestion information. This data on the wire can cause overflow in the congested queue. Switches respond by dropping packets.

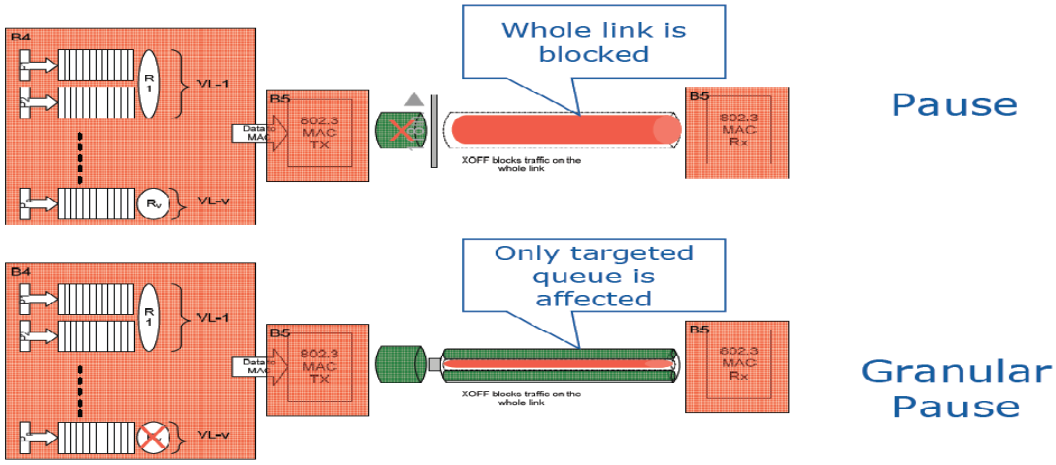

Figure 6 - Granular Link Flow Control

To avoid this, a switch can resort to link level flow control. However, since this network is a consolidated network, data center managers cannot flow control the whole link, only the Virtual Interfaces that require “no-drop” behavior can be flow controlled.

IEEE 802.3x defines link level flow control and specifies protocols, procedures and managed objects that enable flow control on IEEE 802.3 full-duplex links. IEEE 802.1Qbb enhances this basic mechanism to achieve per priority flow control by adding priority information in the flow control packets.

This mechanism, in conjunction with other Data Center Bridging technologies, enables support for higher layer protocols that are highly loss sensitive while not affecting the operation of traditional LAN protocols utilizing other priorities. In addition, Priority-based Flow Control complements 802.1Qau Congestion Notification in Data Center Bridging-enabled networks.

Comparing Fibre Channel to Ethernet Flow Control Mechanisms

The following sections compare the two very different flow control mechanisms for Fibre Channel and Ethernet networking technologies that must be accommodated in a converged network.

Ethernet Flow Control Mechanisms:

The mechanisms that control flow in an Ethernet network are characterized as PAUSE frames. PAUSE based flow control is timing dependent, reactive and must take into account several delay products such as link speed, media delay, software delay, MAC delay, PHY delay, etc. in order to properly time the high-water mark of internal receive buffers that trigger a PAUSE frame. This is a very complex engineering task and is not done to the same level of rigor in every design.

If Station 1 needs to send a PAUSE frame to Station 2 because RX1 has reached its queue high-water mark then it must wait (worst case) until a MAX MTU frame has been sent from TX1, then send the PAUSE frame from TX1 accumulating delays in High Level, Interface and Cable on Station 1 and Interface, High Level and (worst case) MAX MTU reception on Station 2

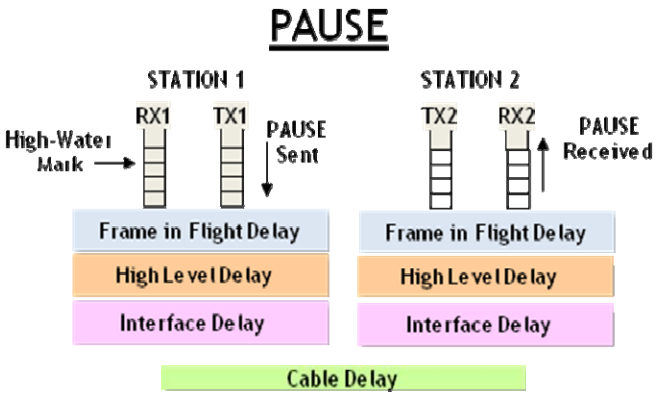

Figure 7 – Pause Frame Ethernet Flow Control

The MAX DELAY would be the sum of all these delays. This would drive the queue buffer and high-water mark values for both stations which may not be the same for each station in heterogeneous networks. In addition, each type of interface/cable combination has different delay characteristics (10GBASE-T not the same as SFP+/SR, for example). One must be very careful to arrive at delay values that provide true lossless behavior. This depends entirely on the robustness of the implementation.

Fibre Channel Flow Control Mechanisms:

Fibre Channel, on the other hand, was purpose built for storage from the ground up, with a credit-based link-level flow control model. All of the "buffers" are of a known quantity, and all of the intermediaries have a known delay, greatly reducing buffer overflow issues. While FCoE SANs can be built with current Ethernet fabrics, providing the mission critical reliability and data integrity required by enterprise data centers requires enhancements to Ethernet to create fully lossless fabrics. Credit based flow control is not timing dependent, is proactive and is easier to design. Credit based flow control is used by FC, InfiniBand and PCI Express today. Only Ethernet uses PAUSE based flow control.

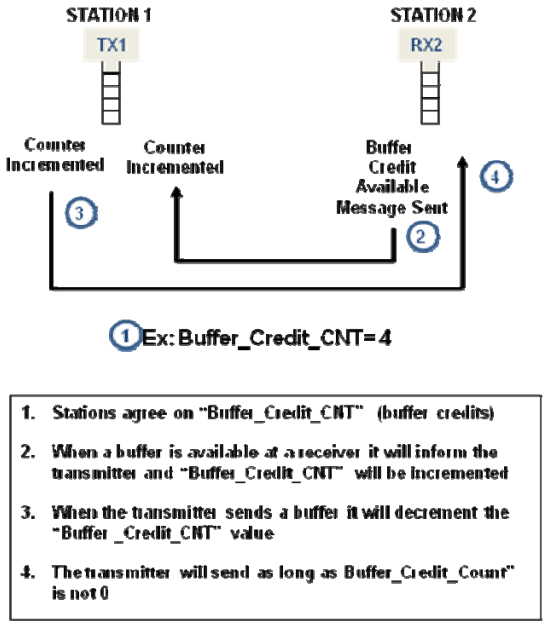

Figure 8 – Credit Based Flow Control Mechanism

DCB IEEE 802.1 Schedules and Timelines

- **802.1Qaz - Enhanced Transmission Selection & Discovery and Capability Exchange**
  - In Work Group ballots
  - Targeting standardization 2010.
- **802.1Qbb - Priority-based Flow Control**
  - In Sponsor Ballot
  - Targeting Standardization 2010
- **802.1Qau - Congestion Notification (Persistent Congestion Control)**
  - Approved Standard

Use Case: Storage over Ethernet

Server-to-storage protocols

In general, server-to-storage applications are intolerant of extended and/or non-deterministic I/O latencies. Many server workloads expect response times in the 1-10 millisecond range and many servers are configured to assume that, in the case of lengthy delays e.g. sixty seconds without a response, I/O operations have been lost and so will typically cause the server application to cease execution. These same applications are simply intolerant of dropped packets as these cause I/O queue backups that will result in the elongated response times and application termination described above.

Strengths and limitations of existing server-to-storage protocols

Fibre Channel is the prevailing protocol standard deployed for server-to-storage connectivity today. While Fibre Channel can be deployed in point-to-point mode i.e. direct attached storage (DAS), the majority of Fibre Channel deployments support "storage area networks" or SANs that enable resource sharing across a larger and more disparate set of server and storage assets than the direct connect model will typically support.

Fibre Channel deployments today run over a dedicated network using dedicated single-use components i.e. Host Bus Adapters (HBAs) providing server-to-storage connectivity, specialized FC switches and directors, special cabling, etc. and require specialized training, experience and products to set up and manage.

Fibre Channel today offers speeds up to 8 Gigabits per second. As such, Fibre Channel works extremely well in providing low-latency at relatively high-performance levels, and meets most of the I/O intensive workloads found in today's data center. While 16GFC is now defined, Ethernet is currently at 10G and will be moving to 40G shortly. A significant issue with Fibre Channel networks is a lack of QoS capability that forces over-provisioning of Fibre Channel network resources to address peak consumption periods. Data center management is seeking ways to increase efficiency and reduce cost. A converged network reduces cost by eliminating the number of host adapters by combining NIC and HBA function into one adapter and eliminating redundant infrastructure cost by using one cable infrastructure to support both storage and IP traffic.

Ethernet Storage

iSCSI, standardized in February 2003 by the IETF, is the server-to-storage protocol designed to transport SCSI block storage commands over Ethernet using TCP/IP. iSCSI was designed to take advantage of all the manageability, ease-of-use, high availability and guaranteed delivery mechanisms provided by TCP/IP, while providing a seamless path from 1 Gigabit Ethernet to 10 Gigabit Ethernet and beyond. iSCSI deployments are observed today primarily over Gigabit Ethernet networks today while 10 Gigabit iSCSI deployments have seen an uptick over the past 18 months.

By using Ethernet as its underlying transport layer, iSCSI addresses two key data center manager issues – reducing both the cost and complexity of deploying Fibre Channel networks. The first aspect is addressed by the fact that iSCSI can run over existing Ethernet networks and the fact that iSCSI can use software initiators that consume server cycles and memory atop existing Ethernet NICs or LOM (LAN on

Motherboard) chips, instead of requiring dedicated and relatively expensive Host Bus Adaptors to provide connectivity. Specialized TOE (TCP/IP Offload Engine) and HBA cards are available for iSCSI deployments and address server cycle and memory offload; however, with improvements in server processors, and memory architectures, (think Moore's law), the need for specialized cards has diminished over time. The second aspect is addressed by the sheer ubiquity of Ethernet network management competency and toolsets that exist today within most data centers.

iSCSI adoption is growing and has found considerable adoption in small and medium size enterprises, as well as finding adoption in cost sensitive applications within larger enterprises. Data center managers with time-sensitive applications, such as OLTP, and Monte Carlo Portfolio Analysis, are uncertain about the latency resulting from iSCSI's dependency on an Ethernet layer with Ethernet's inherent behaviors that can result in dropped packets and extended I/O latencies in lossy environments. Also, as 10G Ethernet brings the hope of converged networking, there is a new need for finer-grained QoS mechanisms that can address each of the previously separate networks.

DCB Advances Ethernet Storage

Within iSCSI's inherent design is the possibility of Ethernet packet loss – in other words, designed around a lossy network. iSCSI leverages TCP/IP to mitigate this by providing guaranteed delivery mechanisms. The result is that if a packet is dropped, there is some network latency as the packet is resent. While this currently is mitigated with network simplification (appropriate provisioning, minimal router hops, VLANs, etc.), the introduction of DCB will enable iSCSI to take full advantage of the lossless nature and deterministic latency of DCB enabled Ethernet networks. In addition, the finer grained QoS capabilities provided by DCB will better enable iSCSI, as well as other Ethernet based storage protocols, in the movement toward converged networking environments.

Data Center Bridging enhancements to Ethernet networks also enables FCoE, a new protocol specification being developed in the INCITS T11 committee. FCoE will allow Fibre Channel commands to be transmitted natively over Data Center Bridging enabled 10 Gigabit Ethernet networks. FCoE specifies the mechanism for encapsulating FC frames within Ethernet frames and is targeted at data center storage area networks (SAN). It preserves the capital investment and operational expertise of existing deployments of Fibre Channel SANs while allowing them to coexist with a converged infrastructure. The FCoE specification was ratified in June 2009.

As described above, the intrinsic improvements in DCB Ethernet for storage deployments (lossless, QoS, low-latency) enables data center managers to finally plan for a singular converged Ethernet based network for all data center communications including user-to-server, server–to-server and server-to-storage networks. By deploying a converged network, data center managers will gain the broad benefits of: cost reduction through the use of common components between endpoints (servers, storage, etc.) and switches; the ability to leverage a common management platform, personnel, knowledge and training across the data center. A platform with a future that today looks to provide a performance upgrade path to 100 Gb/s; a reduction in energy costs resulting from a reduction in dedicated single-use components such as FC only HBAs and switches.

Summary

The ever increasing demand for information is forcing data center managers to find new ways to simplify and drive costs out of their infrastructure. These newly defined set of enhancements to the Ethernet protocol based on 10Gb Ethernet and collectively knows as Data Center Bridging, provide these data center managers an opportunity to converge their previously separate user-to-server, server–to-server and server-to-storage networks into a common network. These Data Center Bridging enhancements (802.1Qaz – Enhanced Transmission Selection, 802.1Qaz – Data Center Bridging Exchange Protocol, 802.Qau – Congestion Notification; and 802.1Qbb – Priority Flow Control) enable Ethernet to now meet the most stringent I/O demands within today's data center while providing data center managers the means to further reduce deployment, management and ongoing operational costs.

About Ethernet Alliance

The Ethernet Alliance is a community of Ethernet end users, system and component vendors, industry experts and university and government professionals who are committed to the continued success and expansion of Ethernet. The Ethernet Alliance brings Ethernet standards to life by supporting activities that span from incubation of new Ethernet technologies to interoperability demonstrations, certification and education.