ethernet alliance

Ethernet Alliance Supercomputing 2008 10 Gigabit Ethernet Converged Data Center Demonstration

November 16, 2008

Cisco, Extreme Networks, Force10 Networks, Fulcrum Microsystems, Ixia, Mellanox, Panduit, Solarflare, Teranetics, Tyco Electronics and QLogic.

Ethernet Alliance | 3855 SW 153rd Drive | Beaverton, OR 97006
www.ethernetalliance.org

Table of Contents

1.0 Introduction

The advantages of IEEE 802.3™ Ethernet are well documented and established for implementation in traditional Local Area Networks (LAN). The role of Ethernet in the LAN data center is evolving beyond the interconnection of clients, servers and switches. The modern Ethernet, specifically 10 Gigabit Ethernet (10 GbE), is becoming the new interconnection technology of choice for Storage Area Networks (SAN) using the Internet Small Computer System Interface (iSCSI) protocol on an IP-based network. The usage of Ethernet infrastructure is being further extended by a relatively new protocol that maps Fibre Channel over Ethernet (FCoE). This allows servers, storage, and management devices to communicate over Ethernet links in a converged data center fabric.

In contrast to other network link types, Ethernet has grown based on its advantage of low component cost (i.e., low cost per link), ease of installation and use, and lower maintenance cost over time to become the pervasive standard for interconnectivity. This has been further advanced with the new 10 GbE capable copper and optical interconnection media. Utilizing the widely adopted RJ45 connector along with reliable low cost CAT6 and CAT6A copper cabling, 10GBASE-T (IEEE Std. 802.3an™-2006) brings incredible speed to LAN data centers. From the optical perspective for 10GbE networks, the new Small Form Factor Pluggable module (SFP+) optical transceiver and copper interconnection technologies lower the cost of short and long reach optical links.

Based on the SFF-8431 Specifications for Enhanced 8.5 and 10 Gigabit, interswitch and server-to-switch interconnection costs via SFP+ direct attached copper assemblies are also reduced.

The advent of server virtualization technology, in partnership with the increasing deployment of multi-core servers, is a watershed event for 10 GbE networking. In the beginning, virtualization was primarily used for testing application deployment and reducing power and cooling infrastructure costs by consolidating servers. In 2009, virtualization will be deployed more pervasively and new models, such as virtual desktop infrastructure (VDI) where desktop applications and services run on the servers in the data center, will allow for the reduction of hardware costs and more cost-efficient manageability. This will have a direct effect on the push toward virtualization standardization.

The IEEE 802.1™ Bridging Working Group will play a significant role in creating a standard virtual switch. These ports (or MAC addresses) can bridge or switch with switch ports in the network, virtual or real, while maintaining support for Layer 2 services and protocols, such as VLANs and access control lists. Virtualization allows organizations to increase their computing power by virtually replicating physical hardware. Futhermore, 10 GbE enables enterprises and data centers to leverage existing copper cabling and deliver optimal performance in any virtualized environment.

The Ethernet Alliance demonstration at the Supercomputing 2008 tradeshow brings together many of the emerging 10GbE technologies into a converged Ethernet data center environment. The demonstration showcases how 10 GbE optical and copper technologies interoperate and interconnect all of the various elements including physical layer communications interface devices (PHYs,) network interface cards (NICs), converged network adapters (CNAs), layer 2 and 3 switches, storage, network test systems and management consoles. The demonstration also highlights several different applications that can be run on an Ethernet converged network such as running real world virtualization, iSCSI storage I/O, low-latency video, and prioritized network traffic that combine a lossless Ethernet environment and a standard Ethernet network.

Ethernet Alliance member companies participating in this demonstration are: Cisco, Extreme Networks, Force10 Networks, Fulcrum Microsystems, Ixia, Mellanox, Panduit, Solarflare, Teranetics, Tyco Electronics and QLogic.

2.0 Logical view of the Converged Data Center

The converged data center configuration is installed in two network racks. The data center contains several blocks of different Ethernet network technologies that are interconnected into a single fabric. Figure 1, shows the logical blocks of the data center and their interconnection technologies. The data center contains:

1. An iSCSI network running via 10GBASE-T on 100 meters of copper cabling
2. An Ethernet and FCoE network running over 10 GbE optical interconnect (XFP and SFP+)
3. Three server clusters:
    a. A cluster running virtualized I/O and storage blocks traffic with an iSCSI storage device
    b. A second cluster running a low latency video demonstration over SFP+ 10 GbE optical interconnect
    c. A third cluster with a converged network adapter driving 10 GbE traffic into the network via a SFP+ optical interconnect
4. A multi-vendor, 10 GbE layer-2 switching infrastructure with multiple interconnection technologies: SFP+ optical and copper, XFP optical, and 10GBASE-T. The switch infrastructure connects all three server clusters and the management consoles.
5. Management consoles that provides real time access to all of the virtualization, storage I/O, FCoE and Ethernet traffic measurements and statistics.
6. A demonstration chassis with the latest 10 GbE copper and fiber interconnect including CX4 and quad SFP (QSFP) active cable samples.

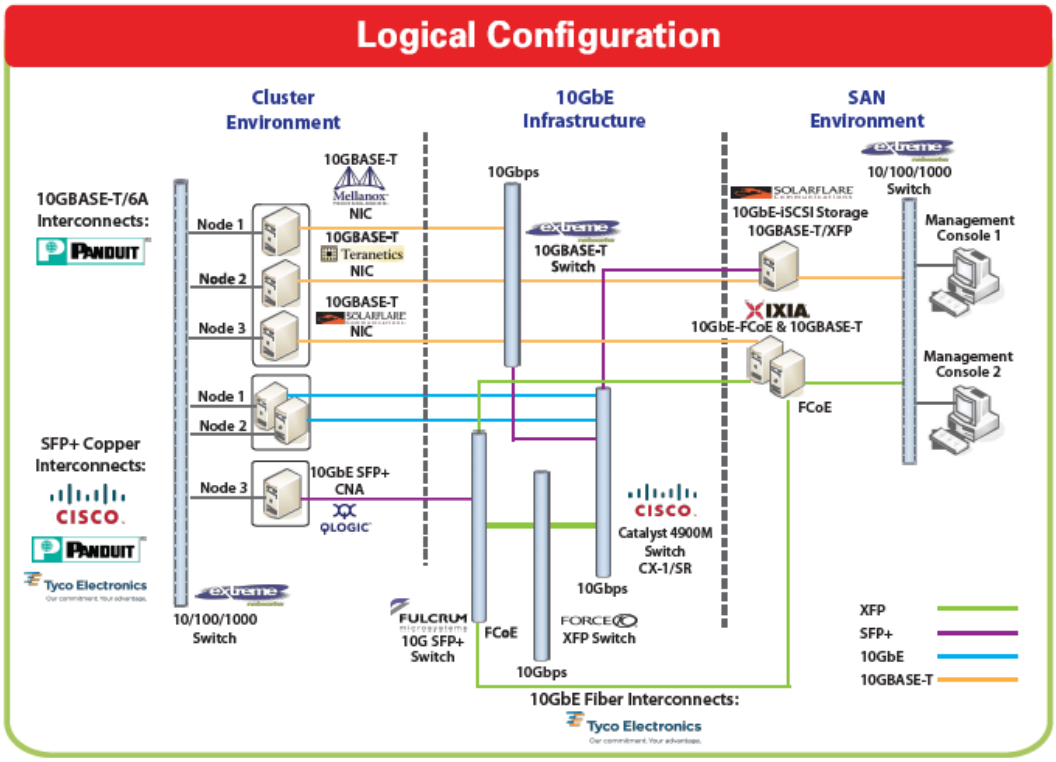

Figure 1 – Converged data center demonstration

3.0 Traffic Flows within the Converged Data Center Demonstration

An explanation of the various traffic patterns and content in the converged data center is provided in Figures 2 and 3. A legend is provided in the figures to emphasize the traffic flows.

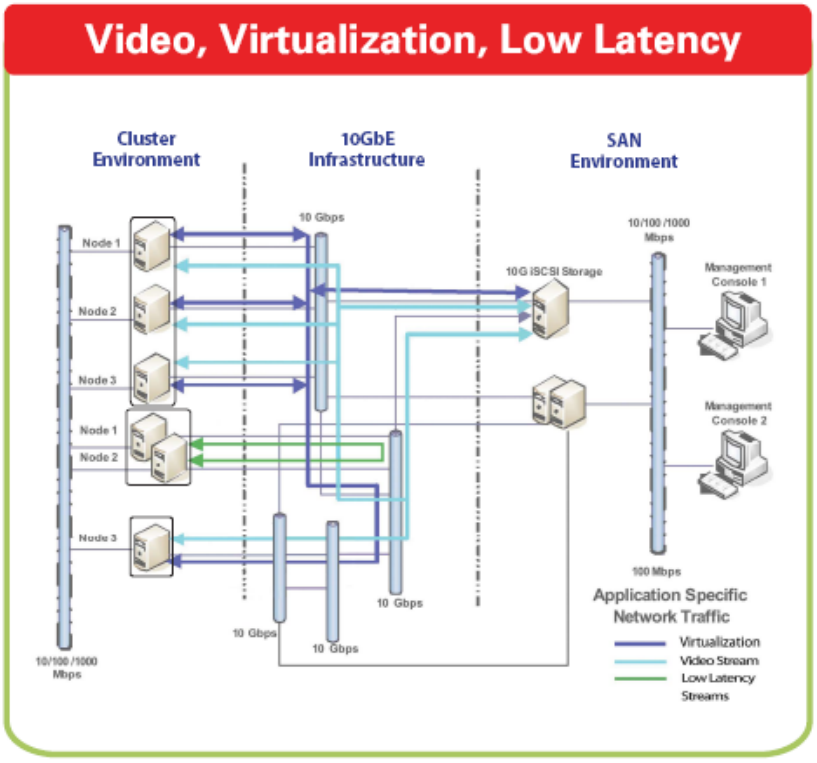

Figure 2 – Video, virtualization and low latency traffic flows

3.1 Virtualization

As shown in Figure 2, the virtualization demonstration has servers running application-to-application data, streaming video, VMware® VMotion and iSCSI. A 10GBASE-T infrastructure using Solarflare's Solarstorm® SFC4000 Ethernet Controller server adapter and Mellanox Connect EN 10GBASE-T adapter operates as an iSCSI target storage network. The servers are running VMware ESX3.5. The demonstration exists in a multi-vendor switch environment showing a combination of converged networking over 10GBASE-T, SFP+ and XFP interconnect.

3.2 Video Application

The video demonstration has a video stream running through the Cisco® Catalyst® 4900M switch with SFP+ copper connections the server and SFP+ optical connections to the network. VideoLAN VLC media player is the application used to stream the video from the server to an end station.

3.3 Low Latency

The low latency demonstration has Solarflare 10 GbE XFP NICs in two servers running Red Hat® Enterprise Linux 5 (RHEL5) connected via a Cisco Catalyst 4900M switch. On each server is Solarflare's Open Onload (OOL) middleware that allows the TCP/IP stack to run in application space. A Pallas message passing interface (MPI) benchmark – now known as Intel MPI benchmark – which is a suite of MPI benchmarks used to measure the most important MPI functions is run between the two-node clusters. The benchmark shows the relative contribution in latency of the OOL-enabled NICs and Cisco Catalyst 4900M switch under the various benchmark loads.

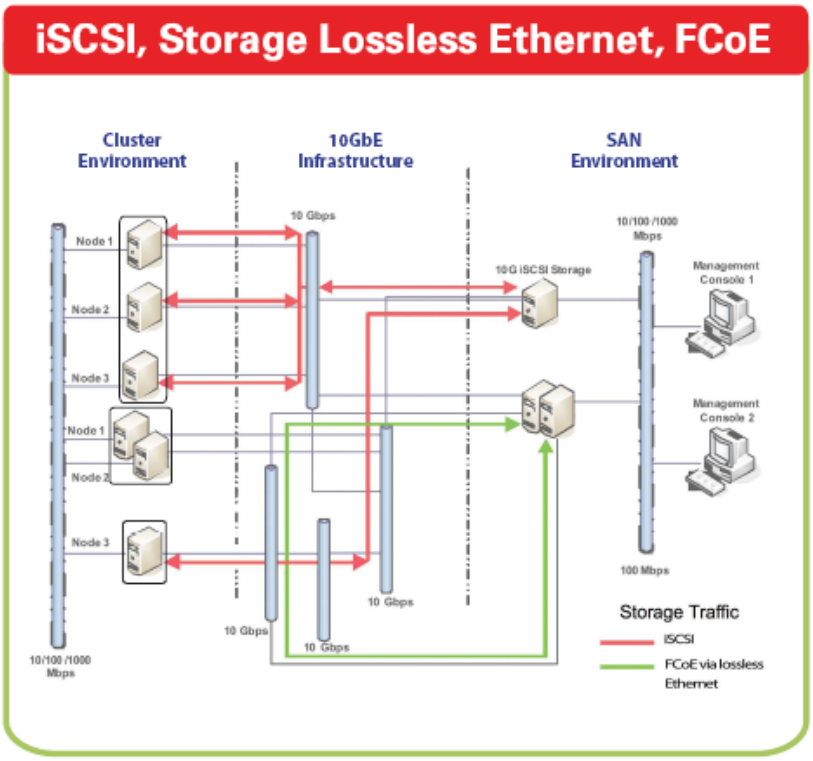

Figure 3 – Storage and lossless Ethernet traffic flows

3.4 iSCSI

Figure 4 depicts the storage element of the demonstration. The iSCSI storage network demonstration operates with Solarflare, Mellanox, Teranetics and QLogic adapters in the servers and multi-vendor switch infrastructure using 10GBASE-T and SFP+/XFP interconnects.

3.5 FCoE and Lossless Ethernet

The lossless Ethernet demonstration compares IEEE Std. 802.3x™-1997, also known as PAUSE flow control, and IEEE P802.1Qbb also known as Priority-based Flow Control (PFC). A comparison is established by using two duplicate network traffic profiles transmitted and received over two sets of four ports. The Ixia 10GbE IxYukon™ load module generates monitors and analyzes identical traffic profiles for both sets of four ports connected to a Fulcrum "Monaco" FocalPoint® switch. The first set of four ports on the switch are configured to accept standard Ethernet and FCoE traffic with PFC enabled. The second set of four ports are configured for standard Ethernet traffic with PAUSE enabled.

The goal of the demonstration is to measure the actual throughput of the FCoE traffic using the PFC support and compare that to the actual throughput using PAUSE flow control. The comparison between the two methods is conducted with deliberate congestion such as traffic oversubscription on one of the switch ports in the test. The target goal is to achieve a guaranteed 8 Gbps throughput for the FCoE traffic.

Figure 4 shows the traffic profiles used for both sets of four ports in the demonstration. The goal is to deliberately congest (i.e. oversubscribe) the Ixia test port 2 in the PFC environment in the first four port set and in PAUSE flow control in the second four port set. The Ixia test port 2 has traffic coming to it from test ports 1 and 3. Test port 2 is being hit with 12 Gbps of traffic to deliberately create congestion. The FCoE traffic is configured for a higher priority than the standard Ethernet traffic for the entire traffic profile.

In the second, four port set, the switch port connected to the Ixia test port 6 is congested with the exact same traffic pattern as the first set of four ports. The only difference is that PAUSE flow control is used. The Ixia test port 6 has traffic coming to it from test ports 5 and 7. Test port 6 is being hit with 12 Gbps of traffic to deliberately create congestion on port 6.

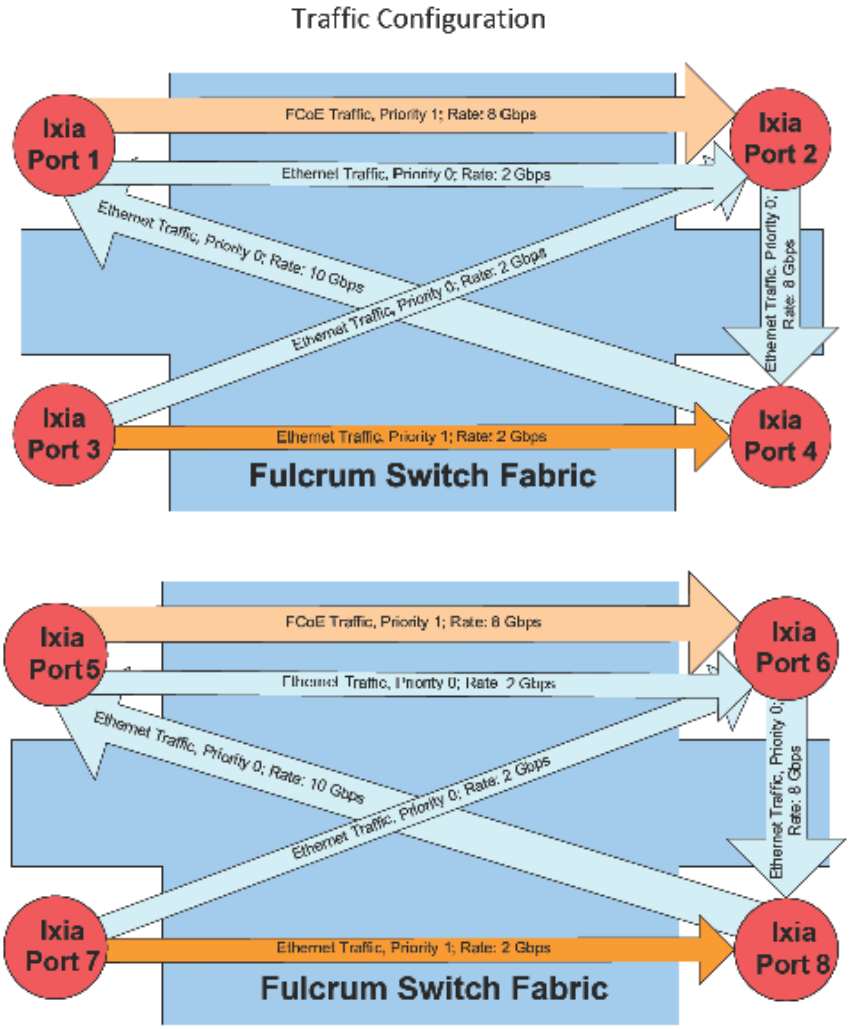

Figure 4 - Test Traffic and Topology for Testing Lossless Ethernet

Figure 5 shows how the Fulcrum switch fabric reacts to Ixia test port 2 being congested with 12 Gbps of traffic load. Only the Ethernet traffic with priority of zero is flow controlled by the switch fabric. Once Ixia port 1 and port 3 receive PFC frames from Fulcrum switch fabric for the lower priority (priority 0) traffic destined for the congested port, they throttle down the standard Ethernet traffic from the offered 2 Gbps to 1 Gbps. The throttle-down message is shown in the figure with the red arrow Pause for the stream being flow controlled. The FCoE traffic with a priority of 1 on Ixia test port 1 is allowed by the switch fabric to flow at the throughput target of 8 Gbps to Ixia test port 2. Since there are no other congestions created, no further flow control is needed, and ports 1, 2 and 3 all end up receiving full line rate 10 Gbps of traffic.

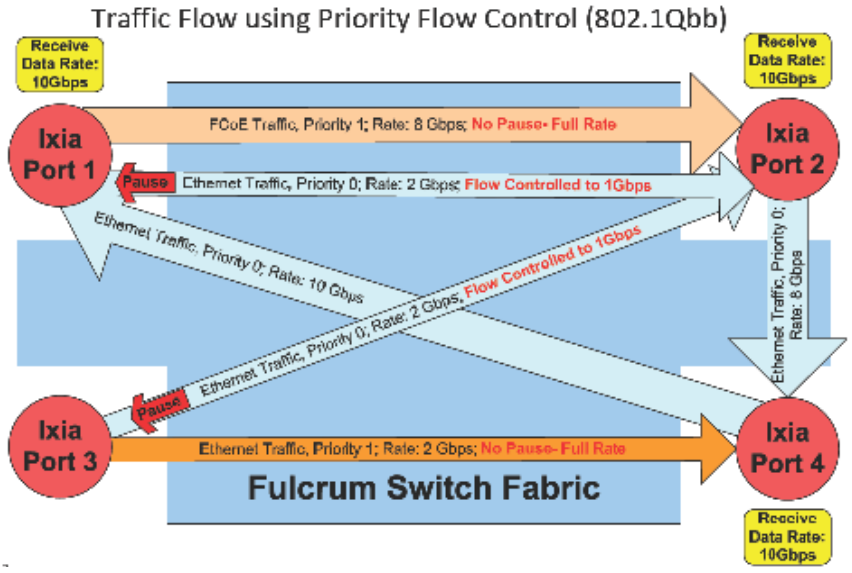

Figure 5 - Priority-based Flow Control guarantees the FCoE throughput at 8Gbps

In Figure 6, it can be seen that the Fulcrum switch fabric, due to congestion at Ixia test port 6, pauses all traffic that it is receiving from test ports 5 and 7. In this scenario, both the FCoE and the standard Ethernet traffic are throttled back by the switch fabric using PAUSE. IEEE 802.3x PAUSE sends a Pause message to all streams. This behavior flow controls all of the traffic out of a port to the congested and non-congested destinations and results in lower throughput of FCoE traffic for the non-congested ports compared to the IEEE P802.1Qbb PFC configuration.

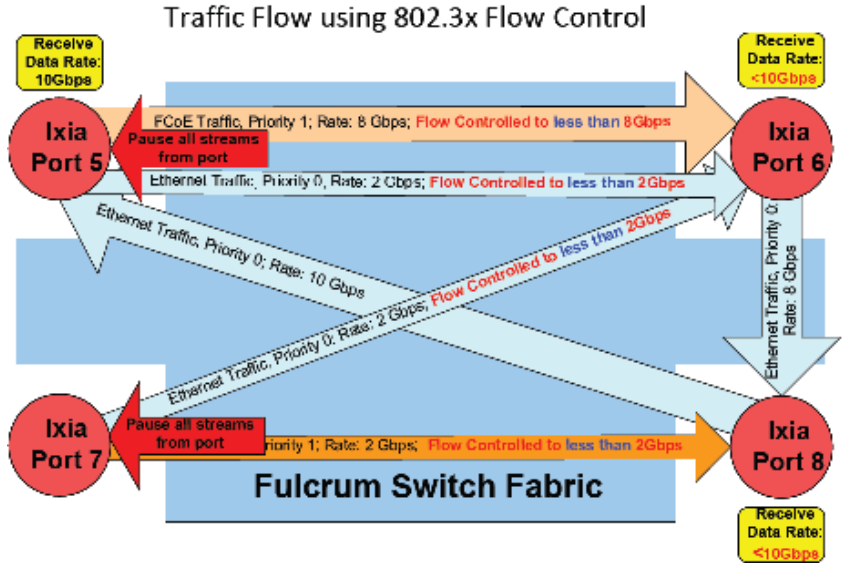

Figure 6 – 802.3x PAUSE Flow Control delivers < 8 Gbps of FCoE traffic

In Figure 7, the FCoE traffic using 1516 byte frame lengths and the IEEE P802.1Qbb PFC data throughput was measured at 7.9 Gbps (green bar). The standard Ethernet traffic using IEEE 802.3x PAUSE flow control data throughput was measured at 5.7 Gbps (red bar). Frame overhead contributed to the final throughput values. The inter-frame gap and preamble are not included in the throughput numbers shown in figure 7.

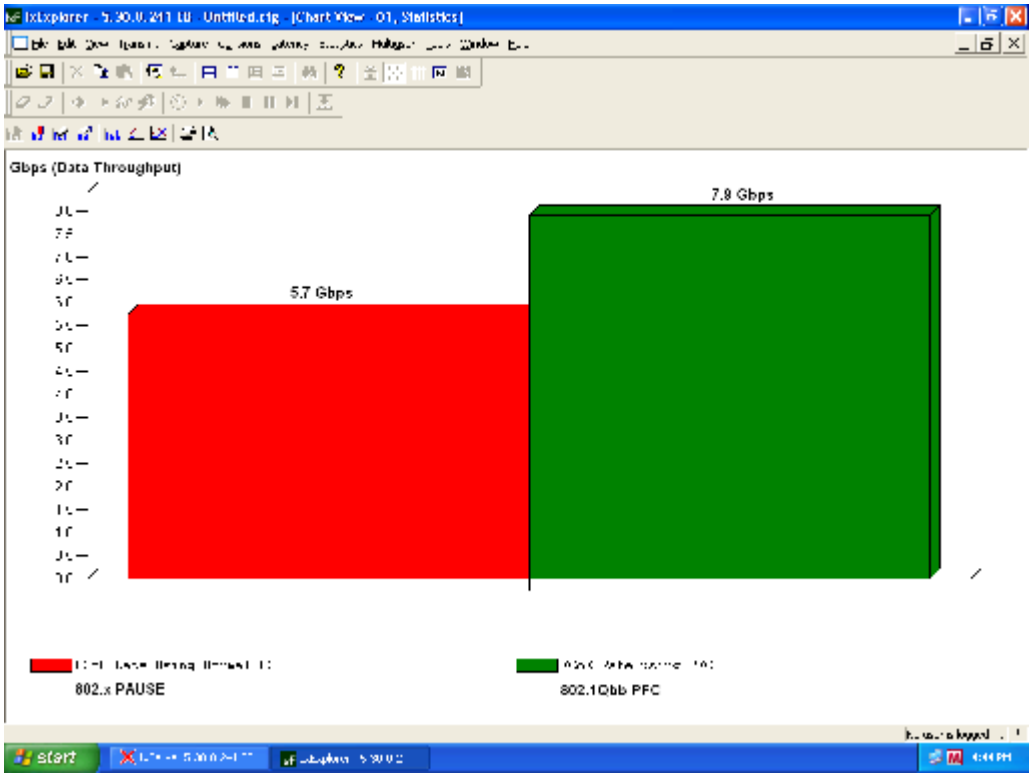

Figure 7 - Comparison of FCoE traffic with PFC versus PAUSE

4.0 Layer 2 and 3 10 GbE Switch Infrastructure

The multi-vendor switched network consists of 10 GbE switch ports all of which are line-rate capable.

The Cisco Catalyst 4900M series is part of the Cisco data center switching portfolio. The Catalyst 4900M is a two-rack-unit (2RU) switch designed for 10/100/1000 and 10 GbE access devices, and provides 10GBASE-T and SFP+ copper and optical connectivity for the demonstration network.

The Extreme Networks Summit® X650 switch provides 10GBASE-T and SFP+ copper and optical connectivity for the demonstration network. The demonstration uses the Summit X650-24t with 24 ports of 10GBASE-T in the front and optional Versatile Interface Modules (VIM) modules for high-speed stacking and/or additional ports in the back. In the demonstration, the front ports are used to connect with servers using standard CAT6A copper cables to the servers. The 8 port SFP+ VIM provides fiber connectivity to the Cisco 4900 switch using 10GBASE-SR.

The Extreme Networks Summit X350 switches separately provide gigabit 1000BASE-T copper connectivity for the demo network.

Force10 Networks provided a 24-port 10 GbE switch with XFP optical transceiver interfaces. This switch is configured for non-blocking layer-2 operation.

Fulcrum Microsystems provided a layer-2 reference design switch with 24 SFP+ interfaces. The switch is configured for both FCoE and Ethernet on eight ports and standard Ethernet on the remaining ports. The demonstration shows low latency performance and Priority-based Flow Control per IEEE 802.1Qbb.

5.0 GbE Adapters

Recent advances in processor cores and PCI Express host bus architecture allows today's servers to support very high bandwidths. In addition, operating systems and applications are able to take advantage of these multi-core architectures furthering the need for 10 GbE. Add to this the enhancements brought about by virtual environments, converged networks, etc. Some of these elements are included in the data center demonstration.

5.1 10GBASE-T network interface cards (NICs)

Unshielded twisted-pair (UTP) copper cabling remains a popular medium of choice for network managers for horizontal wiring. More than 85 percent of the copper cabling currently inside buildings is Category (CAT) 5e or better.

10GBASE-T NICs provide high-speed networking solutions for horizontal copper applications and high-performance networking in the following areas:

- HPC data centers
- Enterprise server farms/data centers
- Local uplinks, aggregation links and inter-switch links
- Other applications that can use in-building structured cabling.

Solarflare designs and develops standards-compliant 10 GbE controllers and 10GBASE-T transceivers that are optimized for all key virtualization platforms. Combined on a server adapter, Solarflare's 10Xpress 10GBASE-T PHY and Solarstorm 10GbE controller are used in the data center demonstration to support video stream, application-to-application streaming, virtualization and iSCSI storage traffic.

Mellanox ConnectX EN 10GBASE-T adapter delivers two ports of RJ45 connectivity on a single adapter. The ConnectX EN is used in the data center demonstration to support video stream, application-to-application streaming, virtualization and iSCSI storage traffic.

5.2 SFP+ converged network adapters (CNAs)

The QLogic 8000 Series is a family of 10 GbE Ethernet-to-PCIe Converged Network Adapters (CNAs) that support NAS, FCoE and iSCSI technologies to converge data and storage networking. This functionality is highlighted in the data center demonstration where the 8000 Series CNA is shown streaming video over 10 GbE via iSCSI. The QLogic 8000 Series CNA offers flexible connectivity options through implementation of an SFP+ interconnect.

6.0 Interconnection Technologies

There are a number of physical media that can be used to transmit 10 GbE network traffic. This data center demonstration uses OM3 multimode fiber (MMF), Category 6A twisted pair cabling, and SFP+ direct attach copper twinax cable.

6.1 10BASE-T Copper Interconnects

The demonstration shows the capability of running 10 GbE over Category 6A UTP cabling with the familiar RJ45 connector. Category 6A cabling provides a structured cabling infrastructure that allows an operating reach of up to 100 meters. In order to achieve 10 GbE transmission over UTP, the cabling requires a bandwidth of 500 MHz. At these frequencies there is a propensity for signal to couple from one cable to an adjacent cable known as alien crosstalk. This can cause excessive noise and packet loss in the transmission. The cabling system within itself must provide alien crosstalk suppression for successful 10GbE operation over standard Ethernet copper connects for extended lengths of up to 100 meters. This demonstration is conducted using Panduit TX6 10GIG Cat6A cabling tightly bundled and spooled in a 100 meter configuration to prove error-free operation in a worse case environment.

6.2 SFP+

SFP+ direct attach copper uses high frequency parallel shielded pair cable, which is known as twinax cable. The cable is factory terminated to an SFP style type module which acts a hot-pluggable connector. This type of interconnect is commonly referred to as 10GBASE-CX1 in the industry. Unlike more familiar UTP and RJ45 interconnects which can be field terminated, all twinax SFP+ cable assemblies are factory terminated.

The SFP+ direct attach copper cable assembly is a low cost SFP+ alternative for short reach 10 GbE applications. The design allows for high speed serial data transmission up to 11.1Gbps in each direction and is defined by the SFF-8431 specification. This specification supports copper SFP+ cable assemblies up to 7 meters in length. SFP+ copper assemblies are hot-pluggable and the programmed EEPROM signature enables the host system to differentiate between a copper cable assembly and a fiber optic module. The mechanical design of the die cast cable plug and external EMI skirt ensure that EMI radiation is sufficiently suppressed. The low power consumption (<1.0 W) and low heat characteristic make the passive copper cable assembly an attraction solution for intra-rack and rack-to-rack interconnect.

The SFP+ direct attach copper cable assemblies used in this demonstration provide a high speed 10Gbps passive connections. The passive design has no signal amplification in the cable assembly.

6.3 Active Optical Cables

As the applications for 10 GbE technologies grow, new cable types are emerging onto the market. Active optical cable assemblies use state-of-the art technology to provide cost effective high data throughput interconnects. The cables incorporate electrical-to-optical (E/O) and optical-to-electrical (O/E) conversion built into the connector shell to yield an improvement in PCB real estate utilization. These optical transceivers are incorporated into standard copper connectors such as the CX4 and QSFP; however since they are internally terminated, there is no optical connector to clean.

7.0 Summary

The Ethernet Alliance converged data center demonstration shows the deployment of several types of mixed traffic over a multi-vendor 10 GbE infrastructure with physical links that consist of both UTP copper, twinax copper and fiber optic media. The SAN/LAN network is converged at Layer 2 using Ethernet as the common transport protocol. The demonstration data center runs 10 GbE application-to-application traffic, iSCSI storage and low latency traffic with a streaming video environment from iSCSI storage, FCoE over a lossless Ethernet network and converged traffic types via virtualized servers. The Ethernet Alliance demonstration shows a physical data center built with equipment from multiple vendors, switches, network interface card (NICs), converged network adapters (CNAs) and cabling systems that supports a complex yet realistic set of 10 GbE application environments.

8.0 About the Ethernet Alliance

The charter of the Ethernet Alliance is to support, promote and educate the industry regarding existing and emerging Ethernet technologies based upon IEEE 802® Ethernet standards. The Ethernet Alliance promotes interoperability events, such as this converged data center demonstration. Interoperability demonstration events bring together Ethernet Alliance member companies and their solutions to work together in educating the industry on the latest advancements in Ethernet technology deployment. For more information on the Ethernet Alliance, visit www.ethernetalliance.org.