



ethernet alliance

Ethernet: THE Converged Network

Ethernet Alliance Demonstration as SC'09

Authors:

Amphenol, Cisco, Dell, Fulcrum

Microsystems, Intel, Ixia, JDSU, Mellanox,

NetApp, Panduit, QLogic, Spirent, Tyco

Electronics, Volex



Table of Contents

- I. Executive Summary..... 2
- II. Technologies in the Demonstration..... 2
 - Data Center Bridging..... 3
 - DCBX and ETS 2
 - PFC 3
 - FCoE 4
 - iSCSI and iSCSI over DCB 5
 - iWARP 5
- III. Description of Demonstration Setup..... 7
- IV. Introduction to products in the demonstration 13
- V. Testing methodologies and Testing Results 15
- VI. Conclusions 19

Figures List

- Figure 1: Priority Flow Control..... 4
- Figure 2: FCoE Mapping Illustration (Source: FC-BB-5 Rev 2.0)..... 5
- Figure 3: iWARP Data Flow Diagram..... 6
- Figure 4: Integrated Demonstration Diagram 8
- Figure 5: FCoE Network 9
- Figure 6: iSCSI Over DCB Network 10
- Figure 7: TCP LAN Network..... 11
- Figure 8: iWARP Network 12
- Figure 9: Monitor Maps for Viewing The Test Results 16
- Figure 10: Testing results displayed from monitor F 16
- Figure 11: Testing results displayed from monitor B..... 17
- Figure 12: Testing results displayed from monitor C..... 19

Table List

- Table 1: Traffic Class Priority and Bandwidth Summary..... 9



1. Executive Summary

Continuous reduction of Total Cost of Ownership (TCO) is the ultimate goal for building next generation data center networks. Key technology transitions applicable to future data centers are network convergence and virtualization. The Ethernet Alliance multi-vendor, multi-technology showcase demonstrates a proof-of-concept converged network based on 10GbE showing its ability to provide high performance networking for various traffic types, including LAN, SAN and IPC traffic. This showcase highlights how network convergence takes advantage of high speed Ethernet to deliver client messaging, storage, and server application communications over a unified network while maintaining the same level of high performance delivered in separate networks. In addition data center interconnect technologies such as 10GBASE-T and SFP+ 10GbE Direct Attach Cables will be demonstrated.

2. Technologies in the Demonstration

Data Center Bridging

In order for Ethernet to carry LAN, SAN and IPC traffic together and achieve network convergence, some necessary enhancements are required. These enhancement protocols are summarized as Data Center Bridging (DCB) protocols also referred to as Enhanced Ethernet (EE) which are defined by the IEEE 802.1 data center bridging task group. A converged Ethernet network is built based on the following DCB protocols:

- **DCBX and ETS**

Existing Ethernet standards do not provide adequate capability to control and manage the allocation of network bandwidth to different network traffic sources and/or types (traffic differentiation) or to allow management capabilities to prioritize bandwidth utilization across these sources and traffic types based on business needs. Lacking these complete capabilities, 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.

Overcoming these limitations is a key to enabling Ethernet as the foundation for true converged data center networks supporting LAN, storage, and inter-processor communications.

Enhanced Transmission Selection (ETS) protocol addresses the bandwidth allocation issues among various traffic classes in order to maximize bandwidth utilization. This standard (IEEE 802.1Qaz) specifies the protocol to support allocation of bandwidth amongst priority groups. ETS allows each node to control bandwidth per priority group. Bandwidth allocation is achieved as part of a negotiation process with link peers – this is called DCBX (DCB Capability Exchange Protocol). When the actual load in a priority group doesn't use its allocated bandwidth, ETS will allow other priority groups to use the available bandwidth. The bandwidth-allocation priorities allow sharing of bandwidth between traffic loads while satisfying the strict priority mechanisms already defined in IEEE 802.1Q, requiring minimum latency.

ETS is defined in IEEE 802.1Qaz Task Force. An additional protocol called DCB Capability eXchange Protocol (DCBX) is defined in the same specification. It provides a mechanism for Ethernet devices (Bridges, end stations) to detect DCB capability of a peer device. It also allows configuration and distribution ETS of parameters from one node to another. This simplifies management of DCB nodes significantly, especially when deployed end-to-end in a converged data center. The DCBX protocol uses Link Layer Discovery Protocol (LLDP) defined by IEEE 802.1AB to exchange and discover DCB capabilities.

- **PFC**

One of the fundamental requirements for a high performance storage network is guaranteed data delivery. This requirement must be satisfied for critical storage data to be transported on a converged Ethernet network with minimum latency impact. Another critical enhancement to conventional Ethernet is to enable lossless Ethernet. IEEE 802.3X PAUSE defines how to pause link traffic at a congestion point to avoid packet drop. IEEE 802.1Qbb defines Priority Flow

Control (PFC) which is based on IEEE 802.3X PAUSE and provides granular control of traffic flow. PFC eliminates lost frames due to congestion. PFC enables pausing less sensitive data classes while not affecting traditional LAN protocols operating through different priority classes.

Figure 1 shows how the PFC works in the converged traffic scenario.

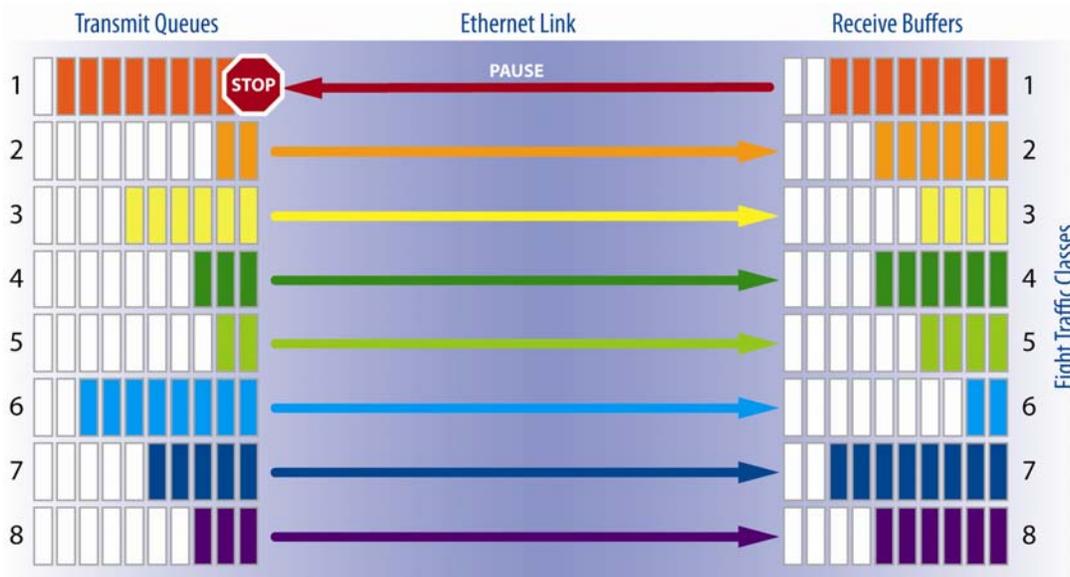


Figure 1: Priority Flow Control

FCoE

FCoE is an ANSI T11 standard for the encapsulation of a complete FC frame into an Ethernet frame. The resulting Ethernet frame is transported over Enhanced Ethernet networks as shown in figure 2. Compared to other mapping technologies, FCoE has the least mapping overhead and maintains the same constructs as native Fibre Channel, thus operating with native Fibre Channel management software. FCoE is based on lossless Ethernet in order to enable buffer-to-buffer credit management and flow control of Fibre Channel packets.

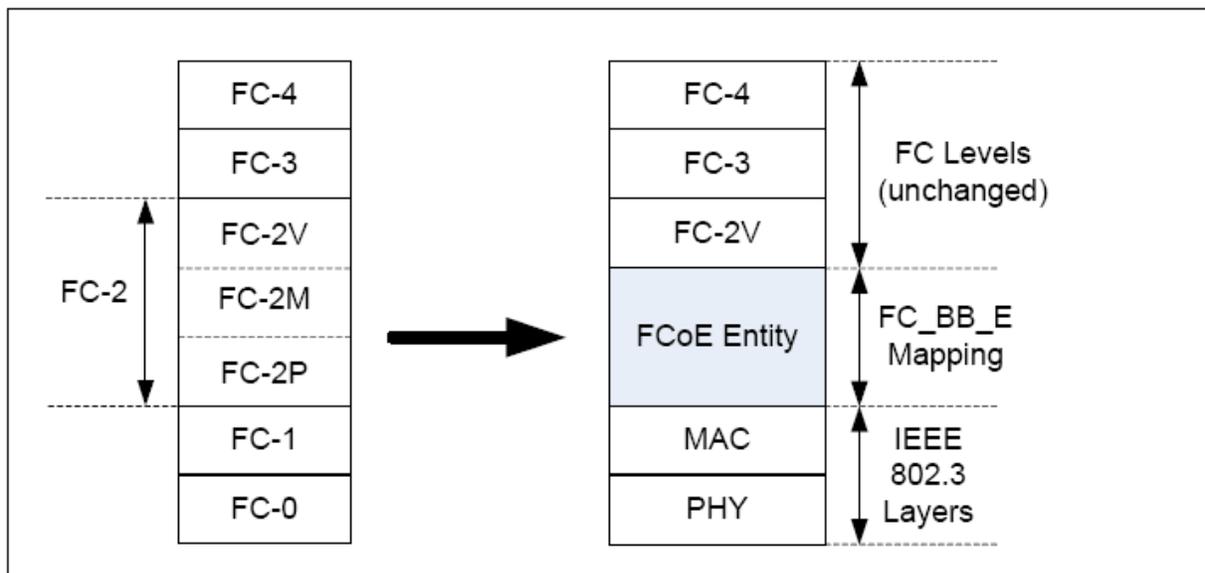


Figure 2: FCoE Mapping Illustration (Source: FC-BB-5 Rev 2.0)

iSCSI and iSCSI over DCB

iSCSI, an Ethernet standard since 2003, is the encapsulation of SCSI commands transported via Ethernet over a TCP/IP network, and is by nature, a loss-less storage fabric. Inherent in iSCSI's design is recovery from dropped packets or over-subscribed, heavy network traffic patterns. So why would iSCSI need the assist of Data Center Bridging (DCB)? iSCSI over DCB reduces latency in networks which are over-subscribed, and provides a predictable and certain application responsiveness, eliminating Ethernet's dependence on TCP/IP (or SCTP) for the retransmission of dropped Ethernet frames. iSCSI over DCB adds the reliability that Enterprise customers need for their data center storage needs.

iWARP

iWARP (Internet Wide Area RDMA Protocol) is a low-latency RDMA over Ethernet solution. The specification defines how the RDMA (Remote Direct Memory Access) protocol runs over TCP/IP. iWARP data flow (see figure 3) delivers improved performance by:

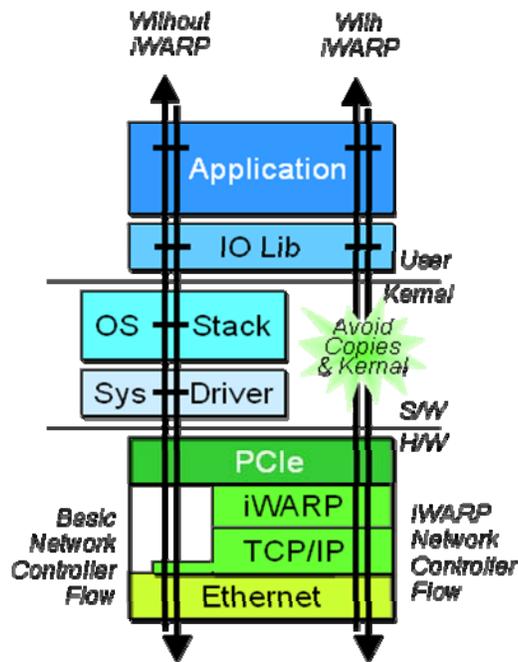


Figure 3: iWARP Data Flow Diagram

- **Eliminating Intermediate Buffer Copies:** Data is placed directly in application buffers vs. being copied multiple times to driver and network stack buffers, thus freeing up memory bandwidth and CPU compute cycles for the application.
- **Delivering a Kernel-bypass Solution:** Placing data directly in user space avoids kernel-to-user context switches which adds additional latency and consumes additional CPU cycles that could otherwise be used for application processing.
- **Accelerated TCP/IP (Transport) Processing:** TCP/IP processing is done in silicon/hardware vs. operating system network stack software, thereby freeing up valuable CPU cycles for application compute processing.

By bringing RDMA to Ethernet, iWARP lends itself to environments that require low-latency performance in an Ethernet ecosystem, including HPC (High Performance Computing) Clusters, Financial Services, Enterprise Data Centers, and Clouds. All of which value Ethernet as an existing, reliable, and proven IT environment that uses heterogeneous equipment and widely-deployed management tools.



3. Description of Demonstration Setup

Figure 4 illustrates the entire integrated demonstration network. The converged Ethernet demonstration contains:

- Three DCB capable 10GbE switches serving as the foundation of the converged Ethernet network
- One 10GBASE-T switch connected to the converged network
- Four 10GbE servers installed with Converged Network Adapter (CNA), a single I/O adapter that supports both FCoE and TCP/IP traffic. One of these servers is running virtual applications on top of a VMware hypervisor
- Two load balancing clusters supporting the high performance computing traffic in the converged network
- Two 10GBASE-T servers providing high performance TCP LAN traffic
- A unified storage system running Ethernet (NFS), Fibre Channel and Fibre Channel over Ethernet protocols simultaneously within a single array to further demonstrate storage system convergence
- A high performance iSCSI storage system supporting iSCSI over DCB storage traffic
- Multiple cabling technologies that support 10GbE links: SFP+ Directed Attached Copper, CAT6A and multimode optical cables

In addition to the network components, the demonstration also includes advanced testing tools for measuring and verifying DCB and other advanced technologies in the converged Ethernet network. These tools provide the following information regarding the demonstration:

- Load generation, measurement statistics, and packet capture, host/target SCSI emulation and performance measurement, and virtualization application modules
- Trace capture and analysis and I/O performance test modules
- Virtualization application test modules

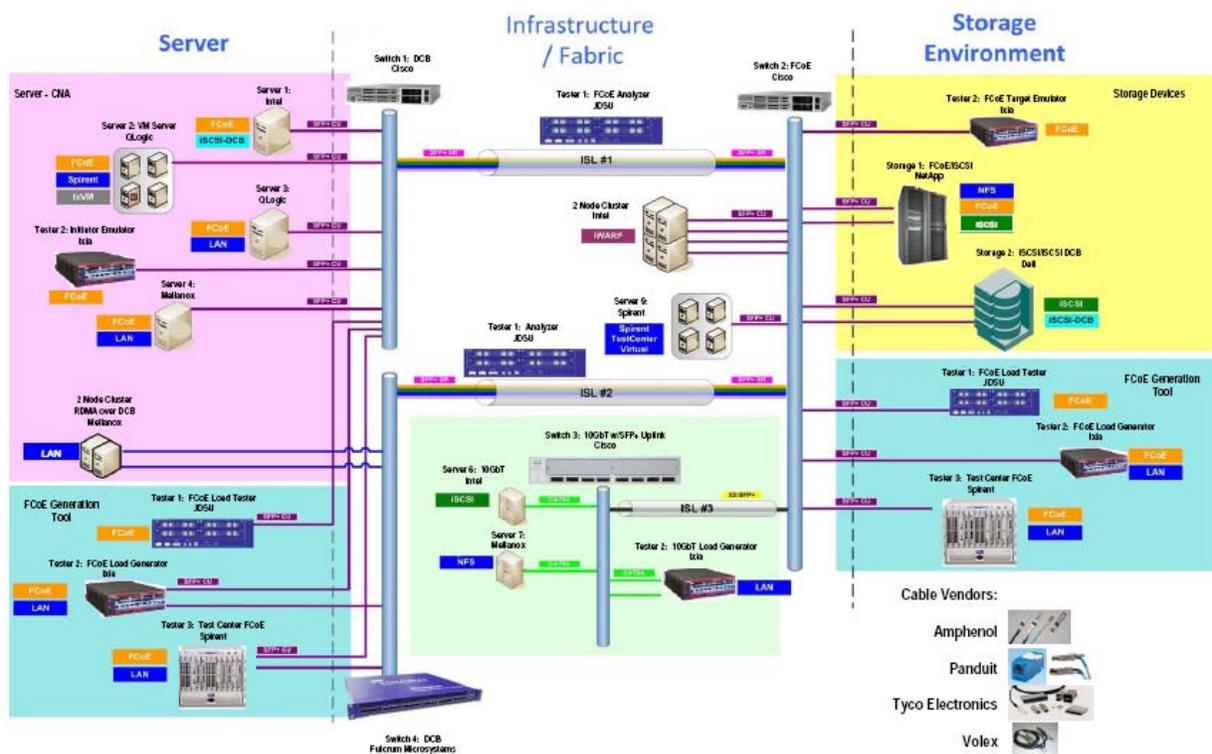


Figure 4: Integrated Demonstration Diagram

In this network, three types of traffic: FCoE, iSCSI over DCB, and TCP are simultaneously transported across the network to establish a converged traffic scenario at ISL#1 and ISL#2 links sharing the 10GbE bandwidth. The network structures for each individual traffic type are illustrated in Figure 5 to 8. Two key DCB features are demonstrated here:

- Utilizing PFC to enable lossless data over Ethernet for critical FCoE and iSCSI storage data
- Utilizing ETS to maximize utilization of the 10GbE bandwidth to achieve high I/O and low latency of each traffic class

Table 1 shows the class of services and priority group assignments to different traffic types. To differentiate the FCoE traffic generated from server-storage data communication and from the load generation tools, they are assigned to different traffic classes.

Traffic Type	Class of Services	Priority	Priority Group Bandwidth	PFC enabled
FCoE (server/storage)	3	1	35%	Yes
iSCSI over DCB	2	2	35%	Yes
FCoE Load Generation	4	4	20%	Yes
TCP applications (iSCSI, virtual applications)	N/A	0	10%	No

Table 1: Traffic Class Priority and Bandwidth Summary

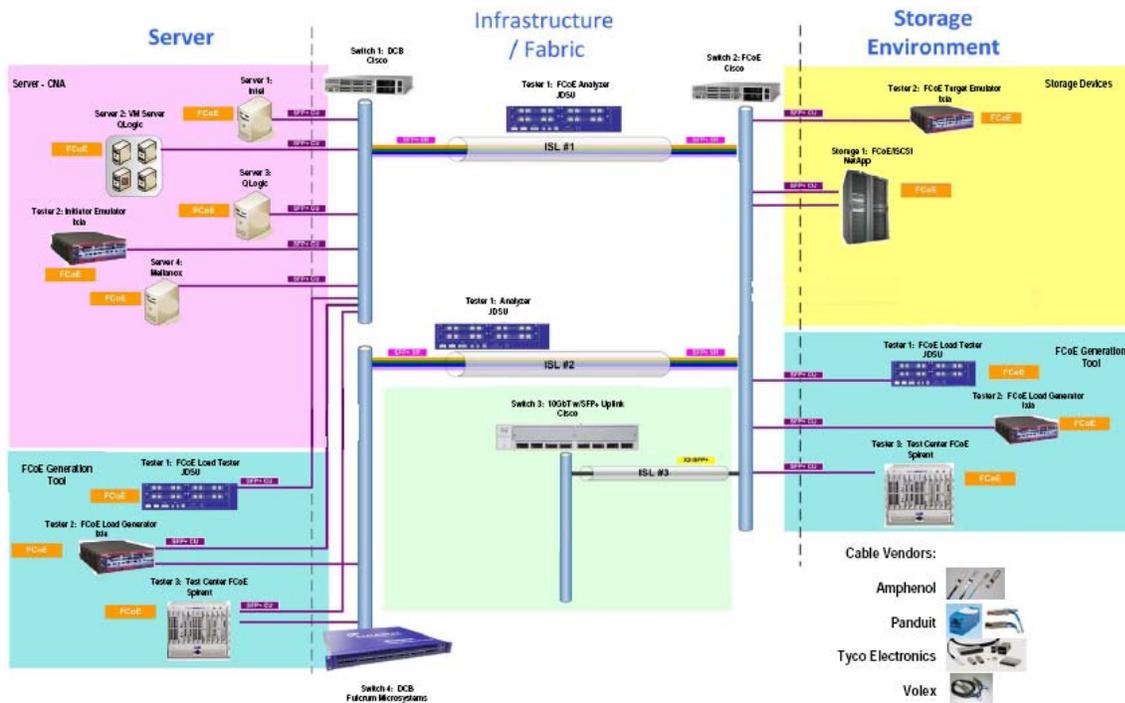


Figure 5: FCoE Network

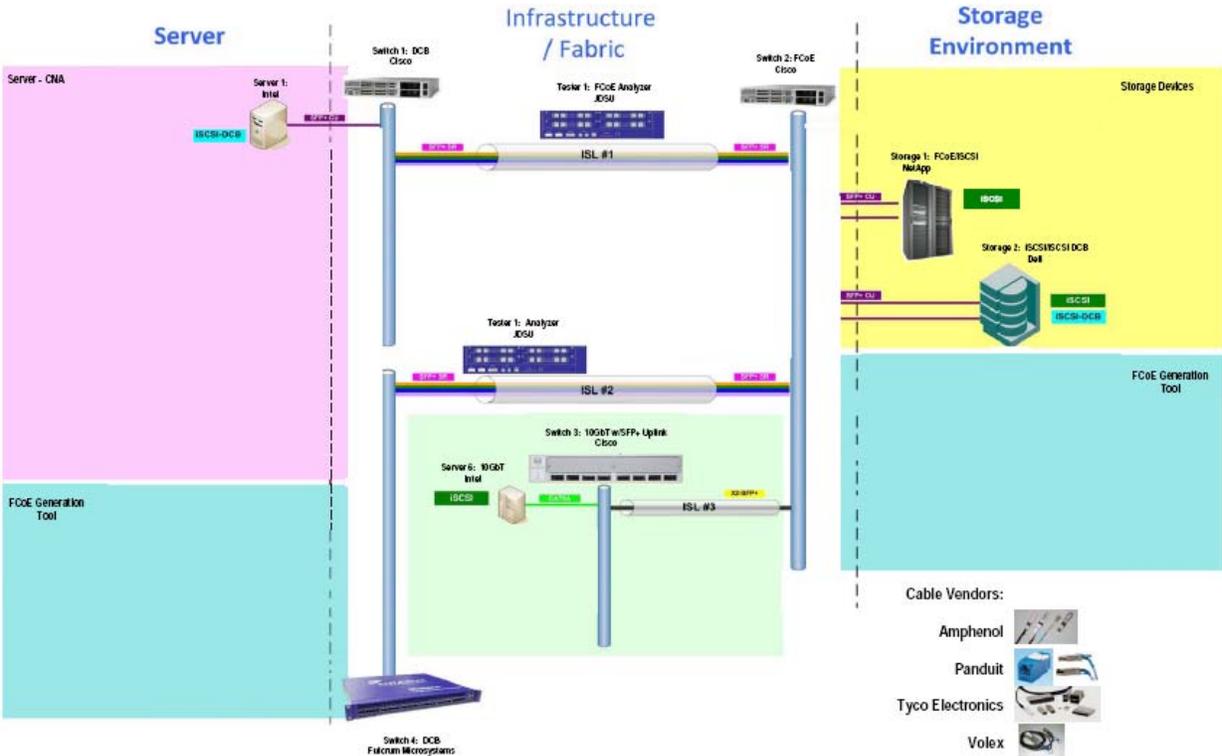


Figure 6: iSCSI Over DCB Network

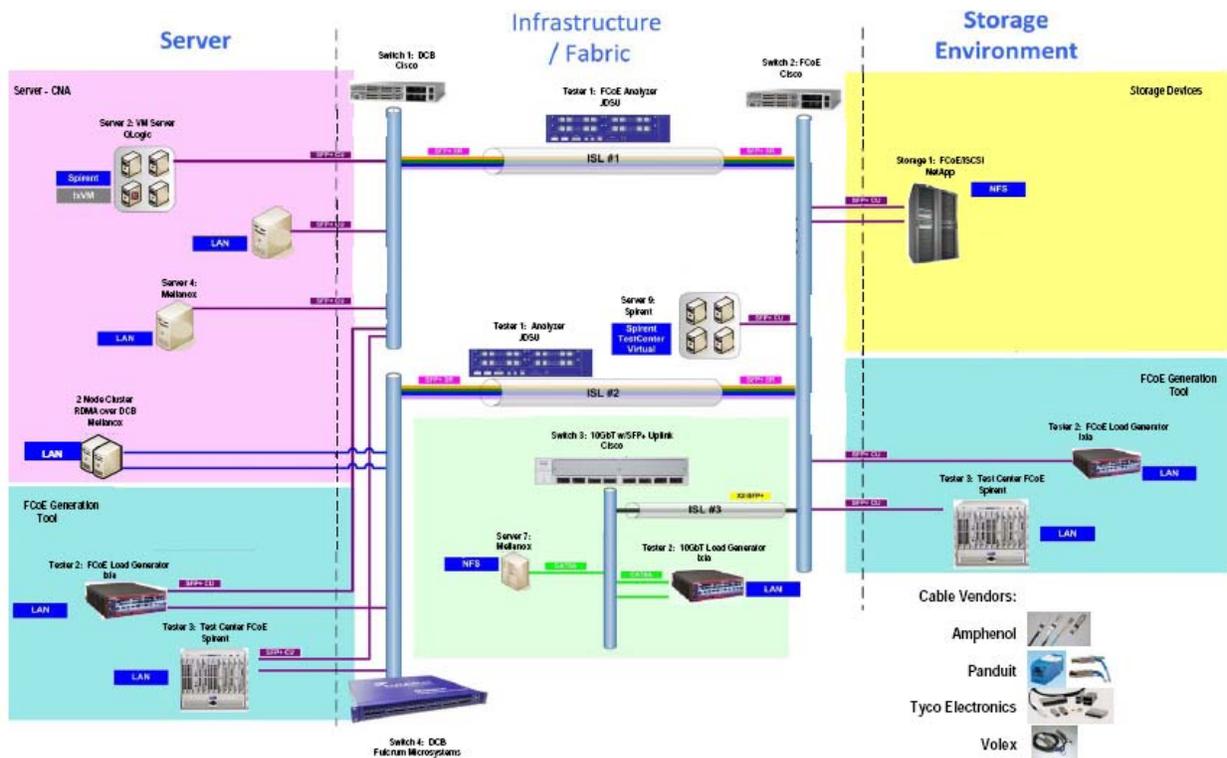


Figure 7: TCP LAN Network

In addition to the traffic types sharing the 10GbE trunk link bandwidth, we also demonstrate iWARP traffic with dedicated 10GbE bandwidth in order to provide the lowest latency and highest performance on these time critical CPU communications. The iWARP cluster computing data is running through switch #1 and #2. Figure 8 shows the network configuration for the iWARP demonstration.

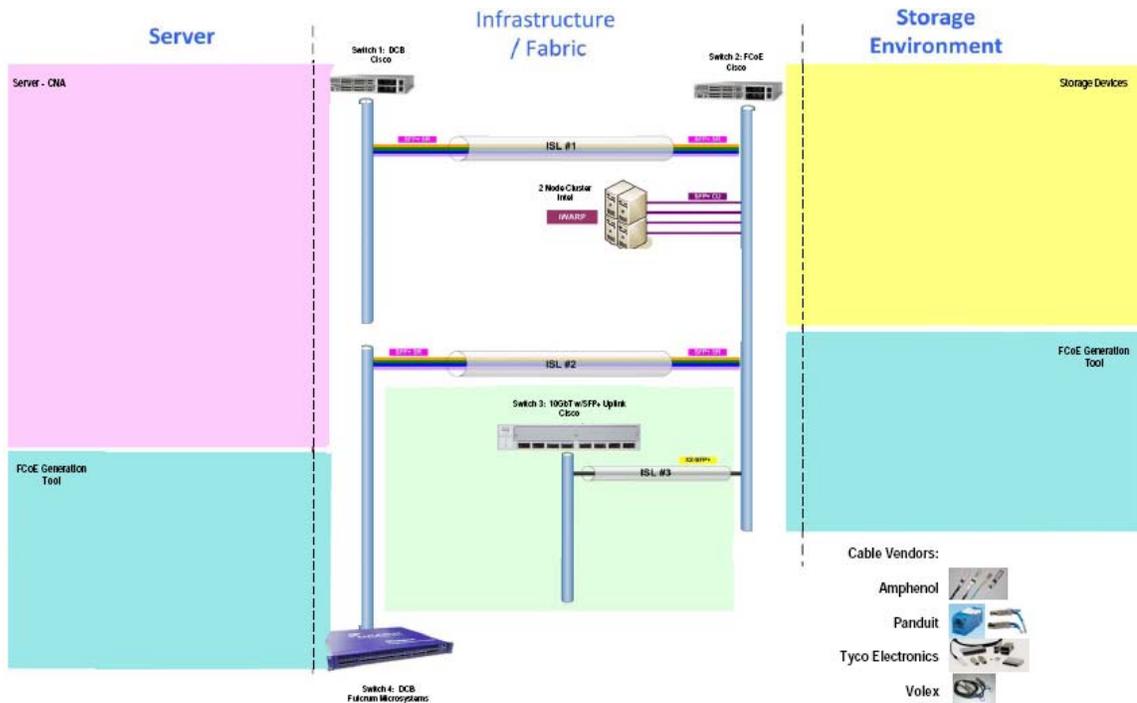


Figure 8: iWARP Network

The following 10GbE physical interfaces are used to create the converged Ethernet demonstration:

- 10GbE SFP+ Direct Attached Copper Cables with passive and active connector interfaces provide the top of rack interconnect from servers and storage to the switching fabric
- 10GBASE-SR SFP+ optical transceivers connected over laser optimized OM3 multi-mode fiber provide inter-switch links
- 10GBASE-T over a CAT6A structured cabling system connects servers to switch ports in the 10GBASE-T sub-network



4. Introduction to products in the demonstration

Amphenol: Amphenol is demonstrating a full line of high performance 10GbE and 8GbFC SFP+ interconnect products. The SFP+ product line includes Direct Attach Copper cables both, passive and active, 10GbE SR transceivers, and OM3 optical cables.

Cisco: Cisco is providing high performance 10GbE switch solutions for the Ethernet Alliance converged Ethernet fabric. The Cisco Nexus 5000 Unified Fabric solution is running iSCSI, FCoE and LAN traffic over a single 10GbE wire. Advanced Data Center Bridging features such as Priority Flow Control and Bandwidth Management will be featured. The Cisco Catalyst 4900M series features 10GBASE-T and SFP+ interfaces running iSCSI and LAN traffic to the larger fabric.

Dell: Dell is supporting the Ethernet Alliance booth at SC09 with a Data Center Bridging iSCSI solution. This includes a Dell EqualLogic PS Series iSCSI storage array featuring 10GbE, SFP+, Data Center Bridging, Priority Flow Control, DCBx protocol, and Enhanced Transmission Selection.

Fulcrum Microsystems: Fulcrum Microsystems is demonstrating a Monaco reference platform that contains the FM4224, a FocalPoint 24-port 10GbE switch-router chip. The FM4224 contains all the features required for converged data center fabrics including support for FCoE. The Monaco platform will be used to demonstrate some of these features including Priority Flow Control and Enhanced Transmission Selection.

Intel: Intel is supporting converged Ethernet with 10GbE cards. The 10GbE SFP+ NIC will demonstrate FCoE and iSCSI with industry standard Data Center Bridging enabling lossless delivery. The 10GBASE-T connection will show native iSCSI acceleration to storage devices; and Intel's 4 node cluster is using NetEffect Ethernet Server Cluster Adapters showing RDMA over Ethernet traffic using its low latency iWARP technology.

Ixia: In the Ethernet Alliance demonstration, Ixia has a complete set of high density 10GbE load modules and network test software solutions that offer a complete end-to-end Data Center network test system on a unified L2-7 platform. The platform supports critical infrastructure protocols such as PFC, DCBX, FCoE, CEE and FIP that deliver lossless Ethernet; new test applications have added high scale virtualization,

full software-based traffic generation supporting end-user performance measurements and real SCSI initiator/target emulation test capabilities in a live Data Center.

JDSU: JDSU is demonstrating the multi-protocol Xgig platform that provides the complete test solutions for FCoE and DCB networks. In addition, JDSU is also demonstrating its application based speed benchmarking test tools that drive high IO generations and allow to validate quality products in real converged network environments

Mellanox Technologies: Mellanox will demonstrate a variety of ConnectX-2 single-chip 10GbE solutions. A Low Latency Ethernet cluster will demonstrate application latency as low as 3 μ s. FCoE with hardware offloads running on a Data Center Bridging network demonstrates high-performance I/O consolidation. These advanced capabilities are delivered over 10GBASE-T, 10GBASE-SR, and SFP+ direct attached copper cables.

NetApp: In the Ethernet Alliance demonstration at SC09, NetApp brings convergence-ready storage for an end-to-end 10GbE infrastructure based on the FCoE standard. Moving to a unified 10GbE infrastructure in the data center enables an organization to efficiently migrate all storage traffic to Ethernet to achieve capital and operational efficiencies. The NetApp unified storage demo also includes FC and NFS traffic running on the same storage system.

Panduit: Panduit is showing High Speed Data Transport capabilities within the Ethernet Alliance Converged Network demonstration. This demonstration showcases a multi-vendor, multi-technology 10GbE network within an ecosystem of fully operating complimentary active equipment. Panduit is exhibiting our High Speed Data Transport media of 10GigTM OM3 fiber, 10GigTM SFP+ Direct Attached Copper Cable Assemblies, and TX6ATM 10GigTM UTP Copper Cabling systems.

QLogic: QLogic is demonstrating their single-chip 8100 Series Converged Networking Adapter - QLogic's 2nd generation CNA. The 10GbE CNA will demonstrate FCoE storage networking with full hardware offload for superior SAN performance. The initiator will be shown operating in a virtual and non-virtual environment running over a converged Data Center Bridging Ethernet network.

Spirent: At the Ethernet Alliance booth at SC09, Spirent is demonstrating support for Data Center Bridging, FCoE, FIP, SCSI and RDMA traffic patterns and Data Center Benchmarking. Spirent will also highlight Layer 2-7 virtualization/cloud computing testing that measures end user QoE/QoS.



Tyco Electronics: Tyco Electronics is demonstrating passive and active SFP+ copper cable assemblies in support of a converged Ethernet network infrastructure. These high speed SFP+ direct attach copper cable assemblies are compliant with SFF industry standards and fully support multiple protocols, including 10GbE, 8G Fibre Channel and FCoE. Tyco Electronics will also be exhibiting other Ethernet connections such as RJ45, MRJ21, and QSFP.

Volex: Volex is demonstrating SFP+ passive and active copper interconnect solutions for the Ethernet Converged Network infrastructure.

5. Testing Methodologies and Testing Results

Converged Ethernet is enabled through DCB enhancements that provide high performance networking for storage and cluster computing data over 10GbE networks. In this demonstration, we are focusing on showing the following two features:

- PFC
- ETS

To activate priority flow control and bandwidth QoS, requires high bandwidth applications to saturate converged links ISL#1 and ISL#2. As shown in Figure 4 and 5, each server will run high I/O traffic driven by the I/O generation tools. The goal is to demonstrate the high I/O performance capability of a converged network. Load generation tools send the oversubscribed FCoE, iSCSI and IP traffic to guarantee congestion scenarios at the ISL links. The measurement tools will verify and demonstrate PFC and ETS features by displaying throughput information on the exhibit monitors. The following diagram illustrates the location of the monitors in the converged network demonstration area.

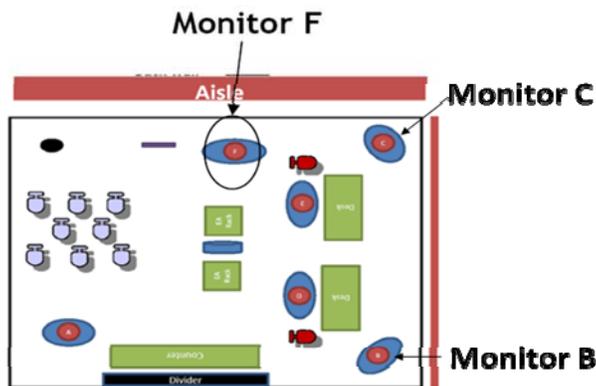


Figure 9: Monitor Maps for Viewing the Test Results

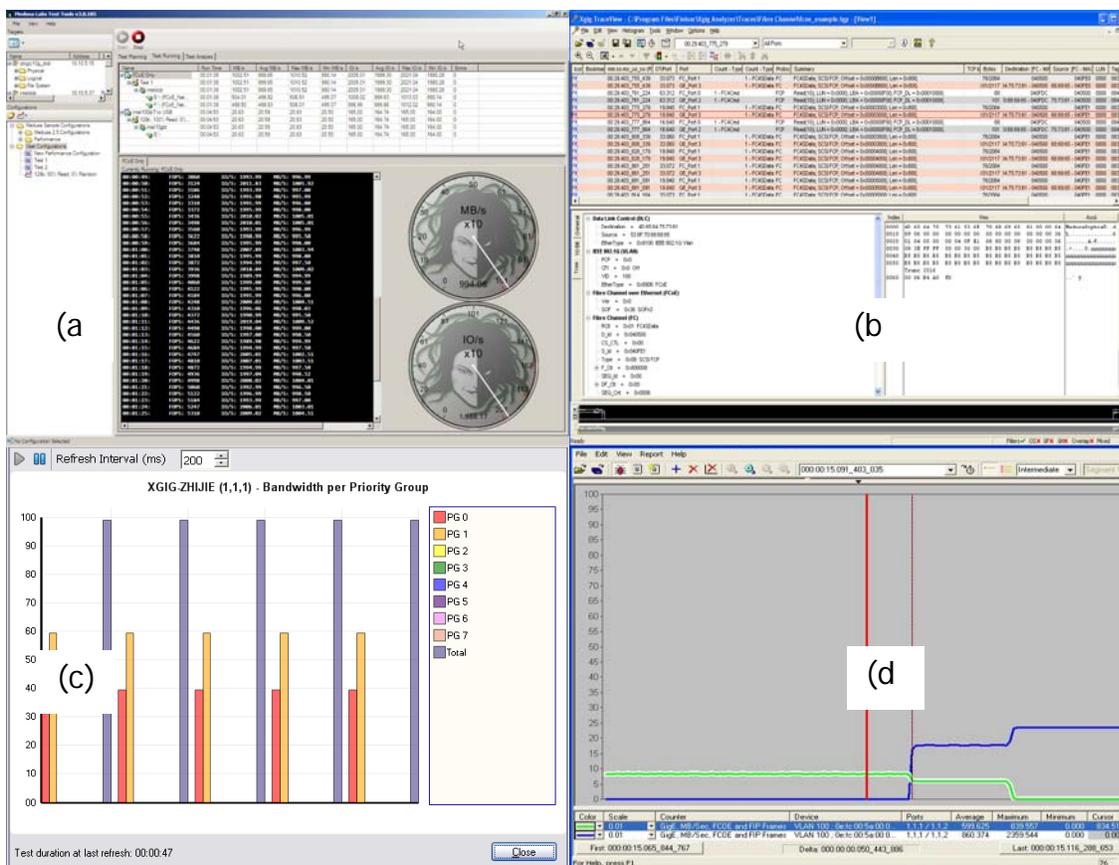


Figure 10: Testing results displayed from monitor F

- (a) MLTT displays the I/O performance
- (b) Xgig Traceview shows the PFC event in the trace
- (c) Xgig Load Tester demonstrates throughput variation per ETS management
- (d) Xgig Expert evaluate ETS results against DCBX setup

From the monitor F, visitors can observe the following testing results

- I/O performance of each server
- PFC frames in the captured trace
- The bandwidth per each traffic class read from Xgig Expert: FCoE, iSCSI and TCP/IP to demonstrate the ETS concept
- Throughput variations and PFC statistics read from the Load Tester to demonstrate the ETS bandwidth management over the time
- Lossless property of the converged Ethernet network shown by the Load Tester

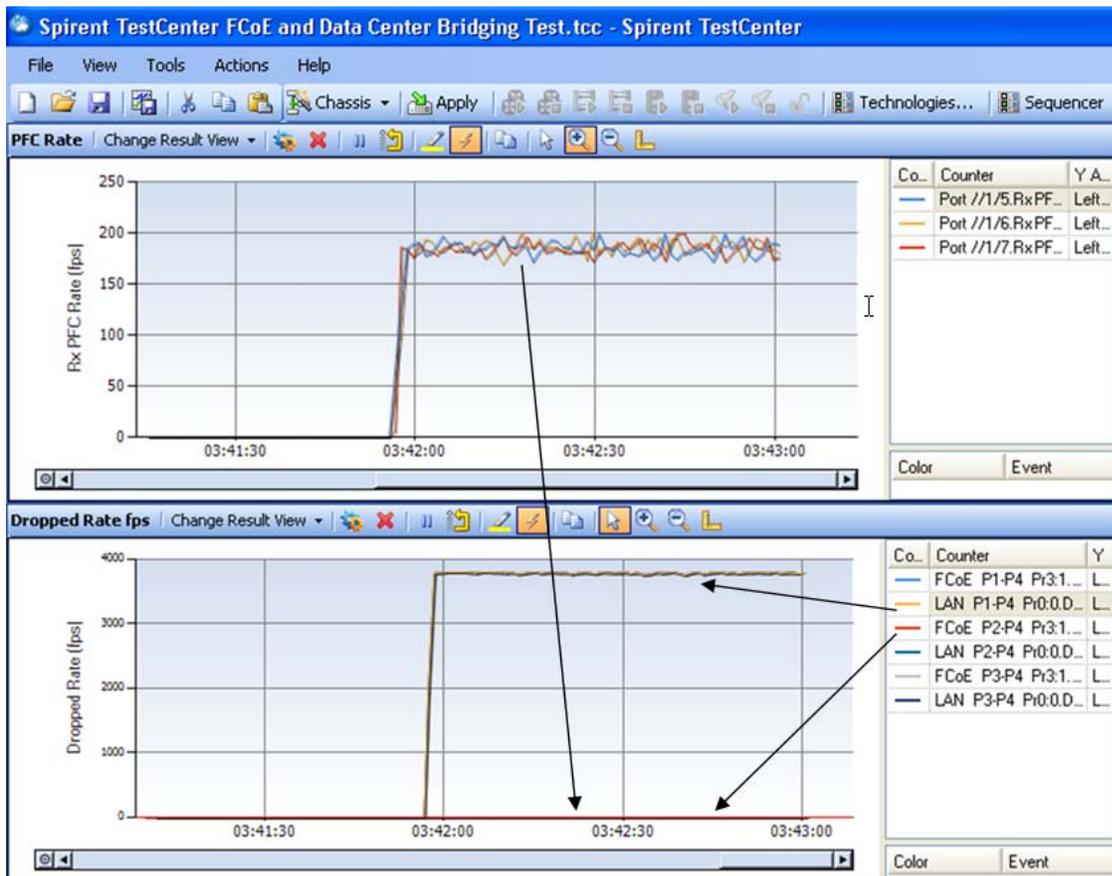


Figure 11: Testing results displayed from monitor B

From monitor B, visitors can observe the following testing results

- The priority flow control over the time due to the ETS bandwidth management
- Throughput and latency performance of the switches
- Virtualization test demonstration

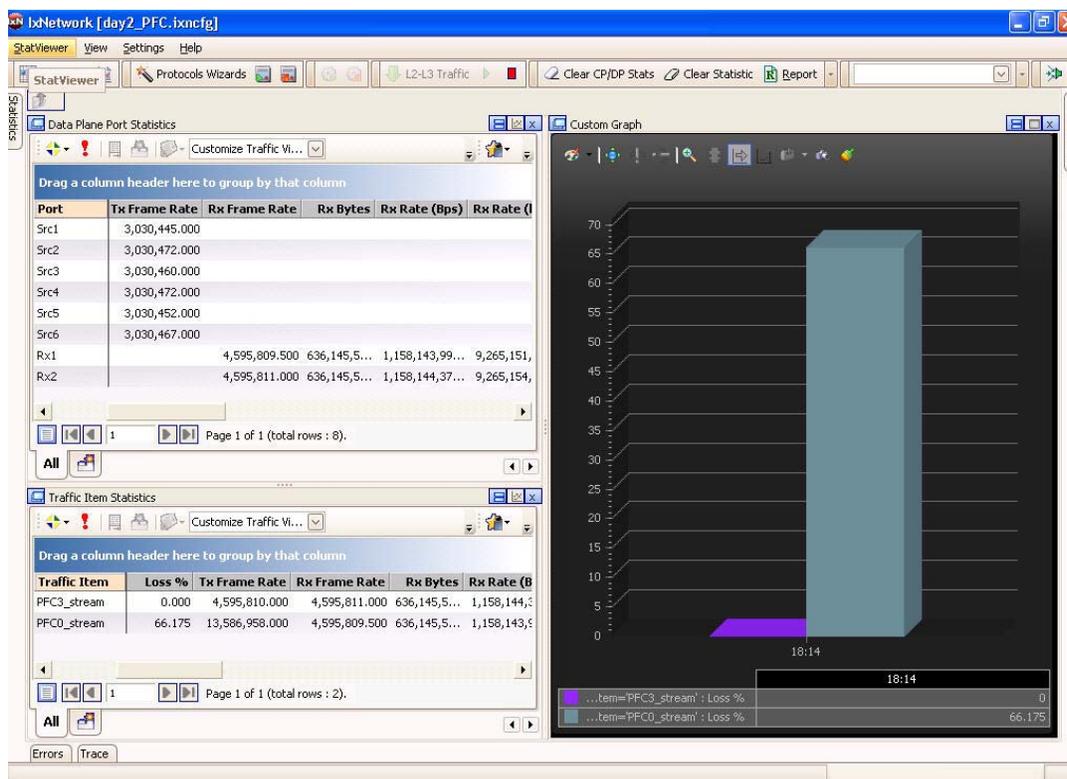


Figure 12: Testing results displayed from monitor C

From monitor C, visitors can observe the following testing results

- 10GBASE-T performance
- Converged FCoE and LAN traffic performance
- Impact of PFC on high priority storage transactions vs. best-effort traffic
- SCSI Initiator and Target emulation: stateful I/O with real servers and drives
- Server virtualization: measuring performance between internal and external VMs

6. Conclusions

The Ethernet Alliance with 14 industry leading Ethernet solution providers is demonstrating a highly consolidated network that carries SAN, LAN and IPC traffic in a single 10GbE network. Key enabling technologies demonstrated here include Data Center Bridging (DCB), Fibre Channel over Ethernet (FCoE), iSCSI over DCB, iWARP, 10GbE low cost physical interfaces: SFP+ Directed Attached Copper Cables and 10GBASE-T.



This industry first large scale network integration gives the IT manager a preview of the next generation data center infrastructure: both consolidated and virtualized with low power consumption and low Total Cost of Ownership (TCO).