



ethernet alliance

Converge Data Center Applications Into a Single 10Gb/s Ethernet Network

*Explanation of Ethernet Alliance Demonstration
at SC10*

Contributing Companies:

Amphenol, Broadcom, Brocade,
CommScope, Cisco, Dell, Chelsio, Emulex,
Force 10, Fulcrum, Intel, Ixia, JDSU,
Mellanox, NetApp, Panduit, Spirent, Volex

November 2010



Table of Contents

| | |
|--|-------------------------------------|
| 1. Executive Summary..... | 3 |
| 2. Technologies in the demonstration | 3 |
| 3. Description of Demonstration Setup – Data Center Networking | Error! Bookmark not defined. |
| 4. Test Results | Error! Bookmark not defined. |
| 5. Conlusion | 9 |



1. Executive Summary

Introduction

Continuously demonstrating new technology trends in Ethernet is the top goal of Ethernet Alliance and its members. Since SC09, Ethernet technology continues evolving at a fast pace. Many new technologies have been emerging to make Ethernet faster, more reliable, power-efficient, and converged.

At this year SC10 Ethernet Alliance booth, we are building a large Ethernet network demonstrating 100Gb/s Ethernet, 10GBase-T, and a single unified 10Gb/s Ethernet network for datacenter applications. This white paper will focus on explaining the setup on the converged Ethernet network.

The next generation data center network is to a single unified network based on Ethernet converging LAN, SAN and IPC traffic. Together with server/storage virtualization and blade server technology, the future datacenter is promised to be greener and lower in Total Cost of Ownership (TCO). In the demonstration, various enabling technologies are introduced which are crucial for achieving the next generation data center goals. Among the technologies seen from the SC09 demo, EA is proud to showcase for the first time the newly ratified HPC standard - RDMA over Converged Ethernet (RoCE), a highly efficient RDMA technology that takes advantage of the Data Center Bridging (DCB) technology, into the converged Ethernet network.

Similar to last year, diverse 10Gb/s data center interconnect technologies such as 10GBASE-T and SFP+ 10Gb/s Ethernet Direct Attach Cables are embedded into the next generation data center demonstration.

This showcase highlights how the converged network with high speed Ethernet can deliver impressive and improved services compared to previous implementation in separate networks. The success of ecosystem build-up among leading network equipment vendors is a strong sign that the advancement to converged data center networks is really happening.

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 DCB protocols and are also referred to as Enhanced Ethernet (EE) and is 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 interprocessor 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 process 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.

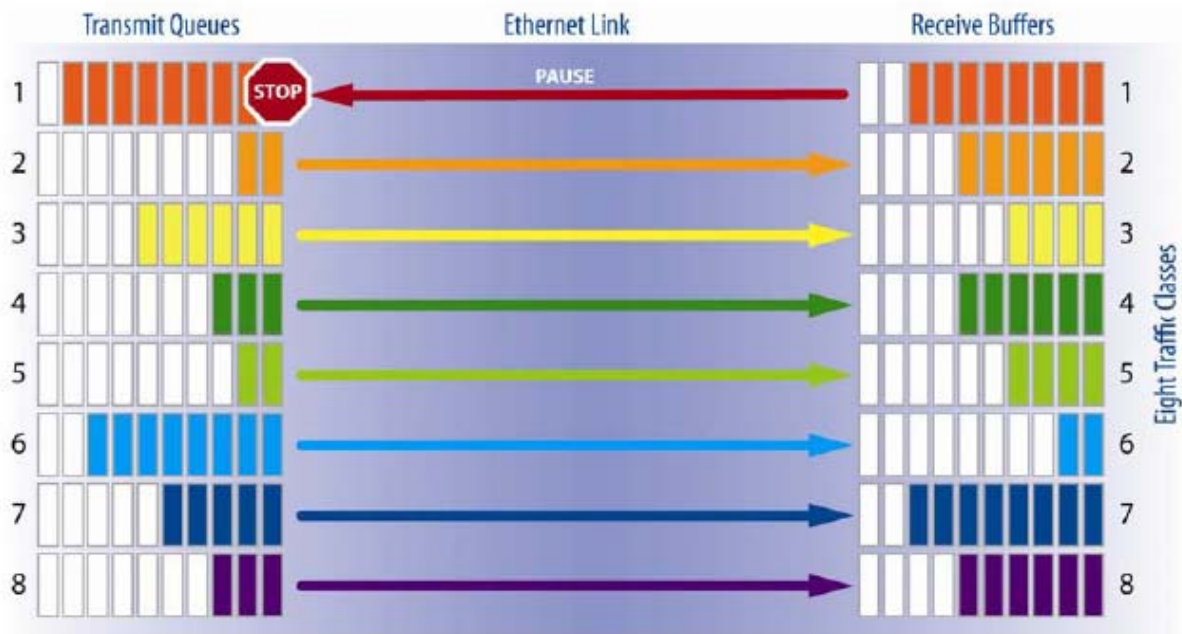
DCBX is defined in the same specification as ETS. It provides a mechanism for Ethernet devices (Bridges, end stations) to detect DCB capability of a peer device. It also allows configuration and distribution of ETS parameters from one node to another. This simplifies management of DCB nodes significantly, 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 FlowControl (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.

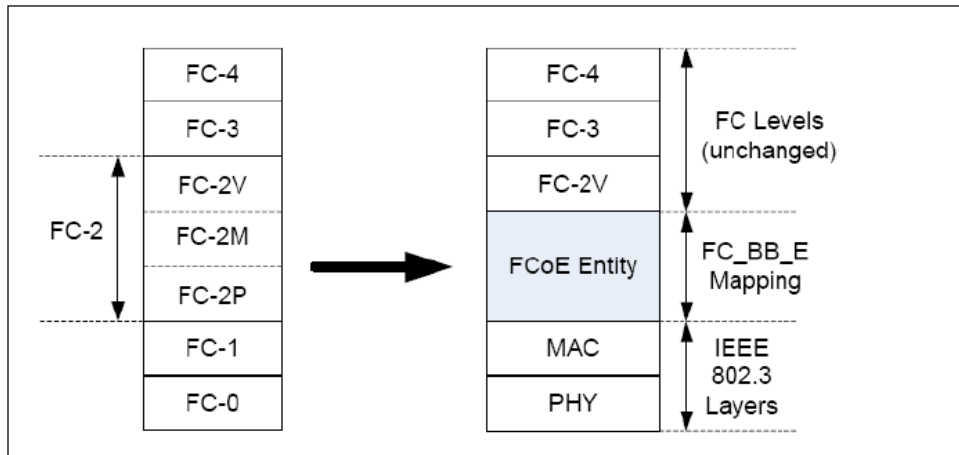
Figure 1: Priority Flow Control



FCoE

FCoE is an ANSI T11 standard for the encapsulation of a complete Fibre Channel (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 providing 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

Internet Wide Area RDMA Protocol (iWARP) 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 delivers improved performance by a) eliminating intermediate buffer copies, b) delivering a kernel-bypass solution, and c) accelerated TCP/IP transport processing.

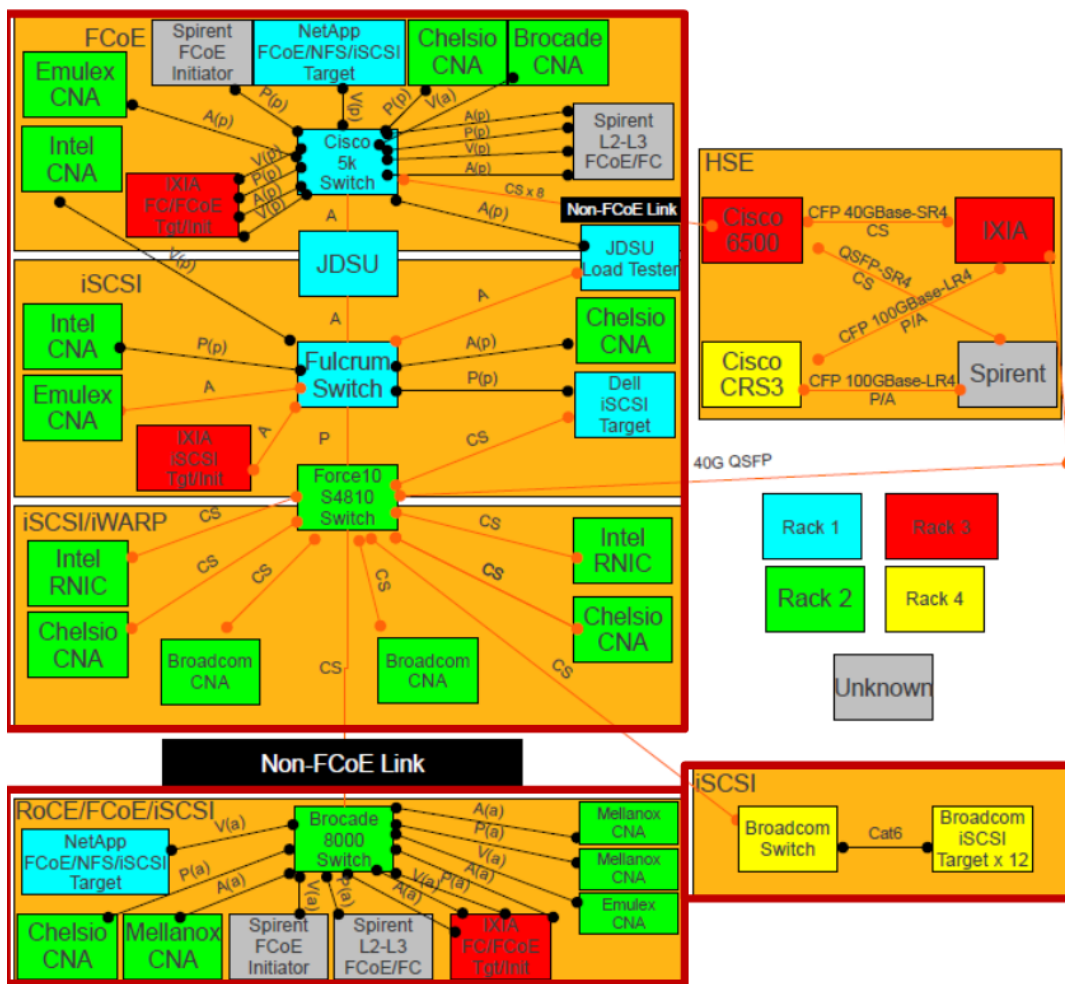
RoCE

RDMA over Converged Ethernet (RoCE (and pronounced "Rocky")), the new standard allows a mechanism to provide this efficient data transfer with very low latencies on DCB technology-enabled Ethernet networks. RoCE refers the new protocol layer which is a verbs compliant Infiniband (IB) transport running directly over the DCB enabled Ethernet layer to carry RDMA applications. RoCE is a proven and efficient RDMA transport to provide HPC technology in mainstream data center application at 10GigE and 40GigE link-speeds.



3. Description of Demonstration Setup - Data Center Networking

Figure 3: Data Center Ethernet
(Data center network outlined in red)



Cable Vendor Legend:

A=Amphenol CS=CommScope P=Panduit V=Volex (a)=active copper (p)=passive copper default=optical

Participating Companies:

Amphenol, Broadcom, Brocade, CommScope, Cisco, Dell, Chelsio, Emulex, Force 10, Fulcrum, Intel, Ixia, JDSU, Mellanox, NetApp, Panduit, Spirent, Volex

The Data Center Ethernet related demonstration contains:

- DCB capable 10GbE switches including switch with Fibre Channel Forwarder (FCF) capability for FCoE applications serving as the foundation of the converged Ethernet network
- 10GbE servers installed with Converged Network Adapter (CNA), a single I/O adapter that supports two or more of the transportation protocols - FCoE, iSCSI over DCB, iWARP, TCP/IP, and RoCE. Some servers are running applications on multiple virtual machines managed by VMware hypervisor
- Unified storage systems running one or more storage traffic kinds - NFS, iSCSI, iSCSI over DCB, and FCoE protocols simultaneously within a single array to further demonstrate storage system convergence
- Load balancing clusters running one or more high performance computing traffic kinds - iWARP and RoCE in the converged network
- Various cable kinds that connect the 10GbE links: SFP+ Directed Attached Copper, CAT6A, and multimode optical cables

Traffic on Data Center Ethernet segments:

There are two types of traffic loads filling bandwidth on various pipes of the 10GbE Data Center Ethernet segments; first is lossless FCoE/iSCSI and the second is standard lossy Ethernet loads.

Given the diverse configuration, to properly orchestrate the task of driving traffic on more than ten separate initiator servers, JDSU's Medusa Labs Test Tool is leveraged as a single management pane to execute traffic tests of lossless block-level FCoE and iSCSI while also addressing lossy file-based iSCSI storage.

During heavy storage IO load in the system, a load generator will be flooding the trunks with lossy UDP traffic type, triggering Priority Flow Control (PFC) and Enhanced Transmission Selection (ETS) on the 10GbE links. This will result in a completely full 10GbE pipe servicing guaranteed lossless FCoE, iSCSI, and lossy Ethernet traffic.

4. Test Results

The DCB demonstration cycles through the following traffic state changes in the network:

State Zero

FCoE/iSCSI read achieve over 50% of line rate at each initiator. Both FCoE and iSCSI are lossless, supporting PFC.

State One

Lossy payload will be blasted across inter-switch trunks with JDSU Load Tester
Observe PFC and rates to committed ETS bandwidth

State Two

Throttle down FCoE/iSCSI, observe deltas

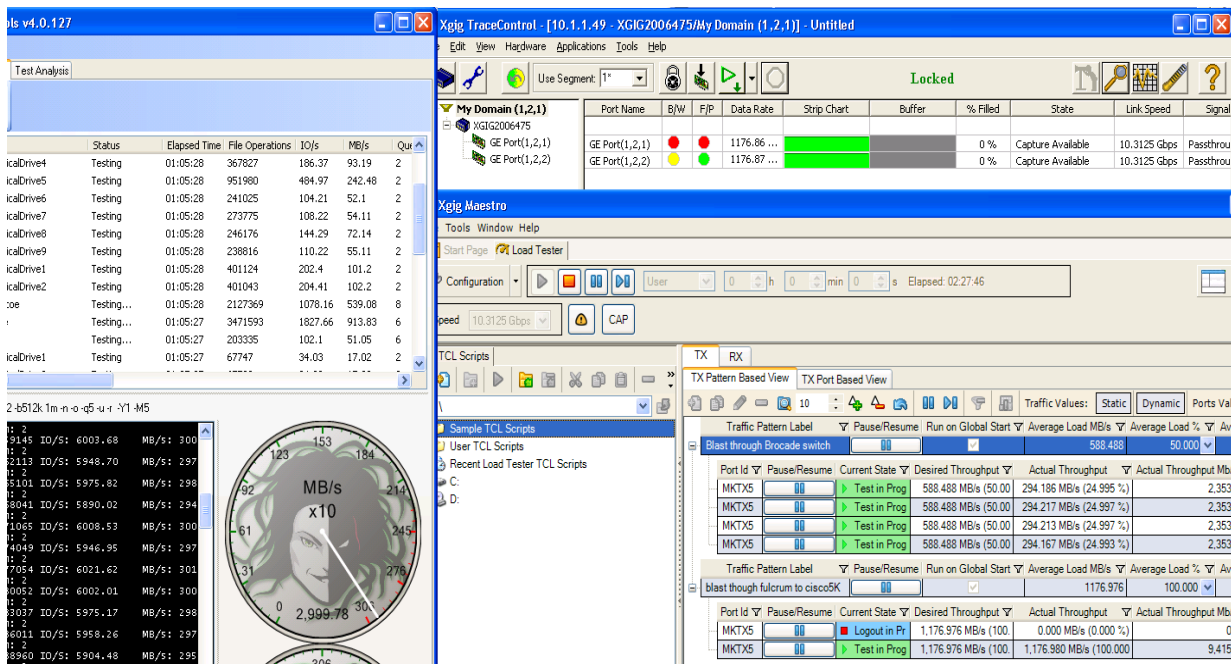
State Three

Resume FCoE/iSCSI and stop lossy payload

By executing the test procedures described above, storage traffic is seen delivering guaranteed bandwidth levels and properly adjusting to changing network load conditions.

The screen shot below displays the aggregate bandwidth of some initiators as seen from the centralized test tool. Also captured is the Xgig analyzer reporting of traffic load between the inter-switch trunk connecting the top two 10Gb switches. Thirdly, the comprehensive management display of the load tester is displayed showing the throughput of injected lossy Ethernet load.

Figure 4: Data Center Ethernet Live Traffic Profiles



5. Conclusion

Ethernet is everywhere and pervasive from end-to-end. The broadening list of Ethernet applications is advancing at a fast pace with fruitful enhancement efforts by many pioneers in the industry. The



Ethernet Alliance, with eighteen 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, RoCE, 10GbE low cost physical interfaces, SFP+ Directed Attached Copper Cables, and 10GBASE-T.

This large scale network integration demonstrates the dramatic progresses made towards converging the network infrastructure of the next generation data center: consolidation for lower Total Cost of Ownership.

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.