



ethernet alliance

Data Center Bridging Plugfest

November 2010



Table of Contents

1 Introduction & Background

Error! Bookmark not defined.

- 1.1 Introduction 4
- 1.2 DCB Plugfest Objectives and Participants 4
- 1.3 Brief description of the standard and technology 7
 - 1.3.1 The Problem 7
 - 1.3.2 DCB - The Solution?..... 7

2 Testing

Error! Bookmark not defined.

- 2.1 Test Methodology 8
- 2.2 What was tested 9
 - 2.2.1 Test Track 1: PFC Tests 9
 - 2.2.2 Test Track 2: Pairwise -- ETS Test 12
 - 2.2.3 Test Track 3: Peer-to-Peer DCBX Test 13
 - 2.2.4 Test Track 4: 10 GBE Interconnect Interoperability 14
 - 2.2.5 Test Track 5 (Optional): QCN with iSCSI over DCB Application 14

3 Results

Error! Bookmark not defined.

- 3.1 The Priority-based flow control (PFC) Test Results 15
- 3.2 Testing the Reaction to PFC Pause Frames 15
- 3.3 Testing the Generation of PFC Pause Frames 16
- 3.4 ETS Test Results 16
- 3.5 The DCBX Test Results 18
- 3.6 10 GBE Interconnect Test Results 20



Table of Figures and Tables

Table 1: DCB Vendors Participating 5

Table 2: Cable Vendors Participating 6

Table 3: Test Equipment Vendors Participating..... 6

Figure 1: PFC test configurations for FCoE application (with FCoE switch or FCF)..... 9

Figure 2: PFC test configuration for iSCSI over DCB application (with switch or direct connection) 10

Figure 3: PFC test configuration for RoCE application with DCB or FCoE switch 11

Figure 4: ETS test configuration with two traffic classes..... 12

Figure 5: ETS test configuration with three traffic classes..... 12

Figure 6: DCBX test configuration 13

Figure 7: QCN test configuration 14

Figure 8: An initiator stops transmitting when receiving a pause frame..... 15

Figure 9: Xgig Expert reports devices that don't react quickly enough to the Pause requests..... 16

Figure 10: The traffic throughput is plotted per priority to verify the ETS Test Results. 17

Figure 11: The trace shows DCBX parameter exchanges..... 19



Introduction & Background

Introduction

In May 2010, the Ethernet Alliance sponsored an interoperability event where 14 vendors including DCB switch, converged network adapter, simulation/monitor, and cable providers came together at the University of New Hampshire Interoperability Lab (UNH-IOL) for the second closed door interoperability test event, also known as a “plugfest”. The purpose for this event was to test the interoperability of several Data Center Bridging (DCB) technologies, which are a collection of various standards defined by IEEE 802.1™. The purpose of these extensions is to enhance the capabilities of Ethernet and is a part of the many efforts of the Ethernet Alliance Data Center Bridging subcommittee. The term DCB describes enhancements to the existing Ethernet technology that enables convergence of various protocols and applications in data centers (LAN, SAN, and HPC) onto a single interconnect technology. As a subcommittee within the Ethernet Alliance, the Data Center Bridging subcommittee focuses on education and testing of Ethernet products in the data center.

The objectives of the Data Center Bridging subcommittee include:

- Act as a reference and resource for the industry for both existing and emerging data center-focused Ethernet technologies
- Identify additional areas of technical work to ensure Ethernet leadership as the fabric in the data center
- Sponsor interoperability events and plugfests for various emerging technologies through third party test facilities
- Align with IEEE 802® task force members and T11 members to promote emerging Ethernet standards for data center networking

DCB Plugfest Objectives and Participants

The objective of this plugfest was to expedite debugging and vendor/product interoperability between the participating companies. The technologies to build next-gen Data Center Networking, the applications based on this infrastructure, and the various 10GbE physical interfaces to support DCB were tested. The technologies include Data Center Bridging Exchange (DCBX), Priority-based Flow Control (PFC), Enhanced Transmission Selection (ETS), and Quantized Congestion Notification (QCN). The applications include FCoE, iSCSI over DCB, and RDMA over Converged Enhanced Ethernet (RoCE). Finally 10GbE cables of various lengths with SFP+ interfaces were tested.

Vendors participating in this plugfest are shown in Table 1.



Table 1: DCB Vendors Participating

| Vendor | Model | Hardware Revision | Software/ Firmware | Technology Tested |
|---------------|--|-------------------|-----------------------|---------------------|
| Cisco Systems | Nexus 4000 Blade DCB | | | DCBX, PFC, ETS |
| | Nexus 5000 FCoE/iSCSI Switch | 1.2 | 4.1.3(n2)1a | DCBX, PFC, ETS |
| Dell | iSCSI storage | | | DCBX, PFC, ETS, QCN |
| Emulex | OneConnect UCNA Oce10102 10Gb NIC/FCoE/iSCSI Adapter | | | DCBX, PFC, ETS |
| Intel | X520 dual port 10GbE 82599 adapter | | Windows | DCBX,PFC, ETS |
| | | | Open Suse | |
| Marvell | MHPS48 | | 2.0.0.7 | DCBX, PFC, ETS, QCN |
| Mellanox | FCF ConnectX ®-2 EN 10 Gigabit Ethernet NIC with RoCE | | | DCBX, PFC, ETS |
| NetApp | FAS3170 Series FCoE Storage | | 8.0.1 x 16 | DCBX, PFC, ETS |
| | FAS3170 Series FC Storage | | 8.0.1 x 16 | |
| QLogic | QLE8142 Converged Network Adapter | | 5.02.01 | DCBX, PFC, ETS |
| | SB5800V 8 G FC Switch | | 2.0.0.7 | Fibre Channel |



Table 2: Cable Vendors Participating

| Vendor | Cable | Model |
|---------|--------------------|------------|
| Leviton | 7M copper SFP+ | SFP10-E07 |
| | 5m copper SFP+ | SFP10-E05 |
| Panduit | 5m Copper | PSF1PXD5M |
| | 7Mm Copper | PSF1PXD7M |
| | OptiCore 10Gig OM3 | FXE3-10M3Y |
| Volex | 15m/26 (a) | VAHS260164 |
| | 10m/30 (a) | VAHS300231 |
| | 5m/24 (p) | VAHS240183 |
| | 3m/28 (p) | VAHS280216 |
| | 7m/24 (p) | VAHS240184 |
| | 8m/24 (p) | VAHS240185 |
| | 3m/32 (p) | VAHS320039 |
| | 5m/28 (p) | VAHS280215 |
| | 5m/26 (p) | VAHS260159 |
| | 3m/30 (p) | VAHS300229 |

Table 3: Test Equipment Vendors Participating

| Vendor | Brand and Version Information | Technology Tested |
|-------------------------|---|-------------------|
| Ixia | IxSAN (software : 1.20) IxNetwork (software : 5.40) LSM10GXM8 FCoE enabled 10GbE ports | DCBX,PFC, ETS |
| JDSU (Formerly Finisar) | Xgig 8G FC and 10GbE test platform (Protocol Analyzer, Jammer, Load Tester) Medusa Labs Testing Suite (MLTT) I/O application | DCBX,PFC, ETS |

| | | |
|---------|--|----------------|
| | test software | |
| Spirent | TestCenter Traffic Generator 10GbE (8 ports) | DCBX, PFC, ETS |

Brief description of the standard and technology

The Problem

With data centers becoming bigger and more complex, managing different interconnect technologies for traffic from each application is becoming cost and resource intensive. Data centers deploy different networks based on distinct interconnect technologies to transport different traffic from different applications; for example, storage traffic is transported over TCP/IP-based iSCSI SAN, Fibre Channel-based SAN, or InfiniBand. Client-server application traffic is handled by an Ethernet-based LAN, while server-to-server IPC may be supported over one of various interconnects such as InfiniBand or Myrinet. A typical server in a high-performance data center has multiple interfaces (Ethernet, FC, and InfiniBand) to allow it to communicate with the various disparate networks.

IT managers today realize that 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 types onto one consolidated network.

Traditional standards-based Ethernet networks do not meet the requirements of storage and high-performance computing applications. Three significant limitations of traditional Ethernet are:

1. No mechanism to distinguish between classes of traffic and provide quality of service by traffic class. The common QoS methodologies in Ethernet are typically at the management layer and impose packet overhead. This might cause performance degradation that may not be acceptable to other protocols such as Fibre Channel and InfiniBand.
2. Fibre Channel (FC) was designed with buffer based flow control to never drop a packet. Traditional Ethernet does not accommodate “lossless” operation.
3. When the amount of data entering the network exceeds network capacity, Ethernet networks become what are known as “over-subscribed” and will “drop” data frames in certain circumstances. This causes inconsistent response times which are not tolerable by time critical applications such as storage and High Performance Computing (HPC) data.

DCB - The Solution?

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 traffic.

Converged Ethernet Network contains TCP controlled LAN, lossless SAN, low jitter, and low latency IPC traffic. Each one of these traffic types has a different set of requirements that in many cases conflict with other traffic types. IEEE 802.1Q Data Center Bridging (DCB) workgroup introduced a set of new features to Ethernet enabling a unified Ethernet network to manage these disparate traffic types and support optimized performance for each individual traffic type. The protocols under construction are:

- IEEE 802.1Qbb- Priority-based flow control (PFC)
- IEEE 802.1Qaz - Enhanced Transmission Selection (ETS) and DCBX

- IEEE 802.1Qau - Quantized Congestion Notification (QCN) (optional test)

The configuration management of these features is controlled in an extension to LLDP called DCBX.

Testing

Test Methodology

Data Center Bridging introduces two low-level features to Ethernet that enable a single Ethernet network to support these disparate traffic types:

- Priority-based Flow Control
- Enhanced Transmission Selection

The configuration management of these features is defined in DCBX.

The goal of Ethernet Alliance's testing is to demonstrate and verify these three features in a multivendor environment.

The baseline version of the standards that all vendors agreed to use can be found here:

<http://www.ieee802.org/1/files/public/docs2008/dcb-baseline-contributions-1108-v1.01.pdf>

The individual documents for each technology can be found at the following links:

- IEEE P802.1Qbb: Priority-based Flow Control:
<http://www.ieee802.org/1/files/public/docs2008/bb-pelissier-pfc-proposal-0508.pdf>
- IEEE P802.1Qaz: Enhanced Transmission Selection (aka Priority Groups):
<http://www.ieee802.org/1/files/public/docs2008/az-wadkar-ets-proposal-0608-v1.01.pdf>
- IEEE P802.1Qaz: DCB Capability Exchange Protocol (DCBX):
<http://www.ieee802.org/1/files/public/docs2008/az-wadkar-dcbx-capability-exchange-discovery-protocol-1108-v1.01.pdf>

Another feature of Data Center Bridging called Quantized Congestion Notification (QCN) was attempted for the first time in product-support DCB plugfest event. QCN is an end-to-end congestion management feature that enables throttling of traffic at the end nodes of a network.

The individual document for this technology can be found here:

[IEEE P802.1Qau: Quantized Congestion Notification \(QCN\)](#)

A Priority Group is group of data traffic bound together by management for the purpose of bandwidth allocation. All data traffic in a single group has similar traffic-handling requirements with respect to latency and loss. There are eight traffic priorities (0 - 7).

A Priority Group ID of 15 has a special meaning that priorities mapped to this PGID will not be subjected to bandwidth limits. Priority group 15 is serviced before other priority groups are serviced. The usage of priority group 15 was tested for the first time at this DCB Plugfest.

What was tested?

Test Track 1: PFC Tests

Test Track 1 Case 1: PFC Tests with FCoE Application

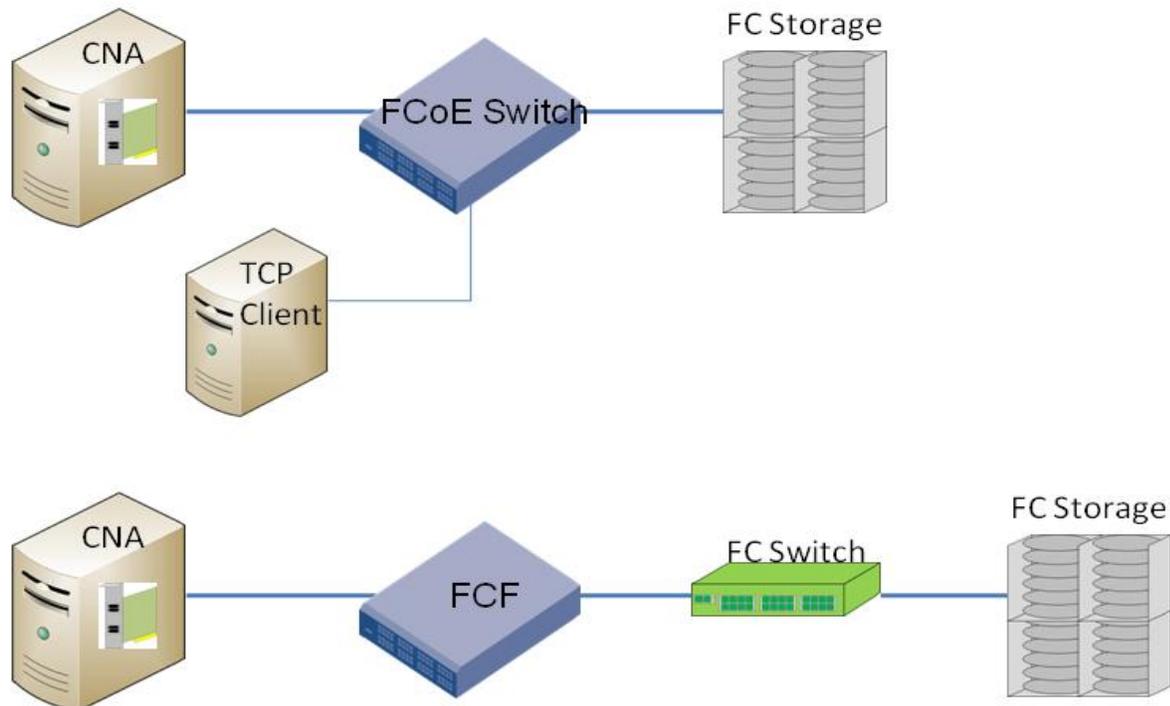


Figure 1: PFC test configurations for FCoE application (with FCoE switch or FCF)

Procedures:

- i. Configure the devices to enable VLAN on FCoE traffic. Configure VLAN header for FCoE traffic to be: VLAN ID=101, PRIORITYGROUP=3.
- ii. Configure the CNA and FCoE switch to enable PFC on FCoE traffic class.
- iii. Do not specify bandwidth allocation per priority group defined on FCoE switch. The default switch priority (i.e.; 50/50) will be enabled (this is true for all PFC-only tests). Since FCF only supports DCBX, PFC, and FCoE to FC forwarding and vice versa, there is no need to configure the priority group on FCF.
- iv. Disconnect/reconnect initiator/CNA.
- v. From the initiator, launch FCoE write-only traffic to the FC targets, while sending TCP traffic to remote TCP client. Make sure the FC link bandwidth is saturated (either by adjusting IO application blocks/queue depth or utilizing Jhammer to eliminate BB-credit).
- vi. (Optional) Plug-in Analyzer to monitor the link between CNA and FCoE switch and setup the capture trigger to be DCBX or PFC frames or ABTS (to detect potentially dropped frames).
- vii. Verify the results of following:
 - a. The PFC parameters in the DCBX frames from both initiator and switch follow the setup definition in "i".

- b. (optional) PFC Pauses expire or are released from switch, and PFC response time is within the spec.
 - c. No drop frames from lossless traffic classes are detected.
 - d. Verify when FCoE traffic is paused that TCP is not.
 - e. Measure the sustained FCoE MB/s at the Initiator.
- viii. Go through power cycle with initiator and FCoE switch and repeat steps "i" to "v" recording the results described in step "vi".

Observable results: PFC requests issued by the switch and CNA responses to the PFC requests. The throughput of FCoE write traffic from CNA is reduced. Expect to see no dropped FCoE frames during PFC pause stage.

Test Track 1 Case 2: PFC Tests with iSCSI over DCB Application

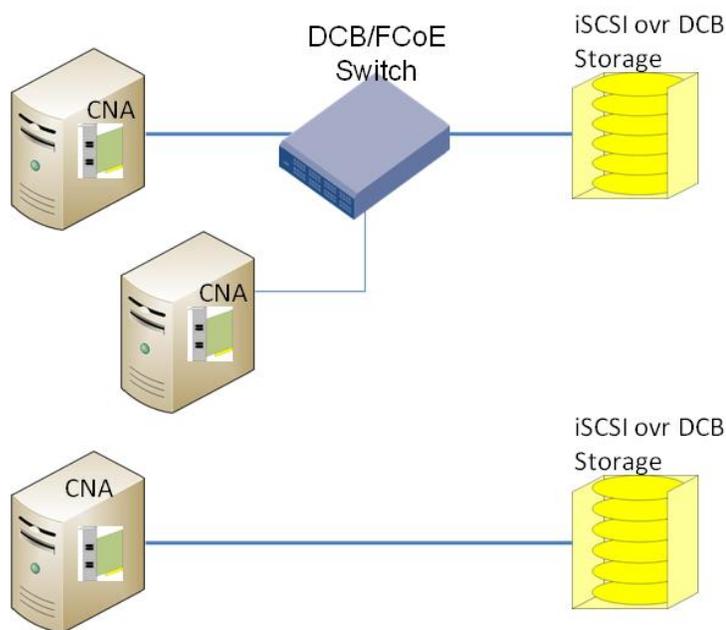


Figure 2: PFC test configuration for iSCSI over DCB application (with switch or direct connection)

Procedures:

- i. Configure 2 or more CNAs (more CNAs could be used in order to saturate 10GbE iSCSI link).
- ii. Configure the devices to enable VLAN on iSCSI-over-DCB traffic. Configure VLAN header for iSCSI traffic to be: VLAN ID=102, PRIORITY GROUP=4.
- iii. Configure the CNA and DCB or FCoE switch to enable PFC on iSCSI-over-DCB application.
- iv. Disconnect/reconnect target.
- v. Launch from the initiator servers iSCSI 100% write application to the iSCSI over DCB target. At the same time launch TCP application between servers following the same traffic direction as iSCSI traffic. Make sure the iSCSI storage link bandwidth is saturated (by justifying IO block size and queue depth).
- vi. (Optional) Plug-in Analyzer to monitor the links between CNAs/iSCSI storage and DCB/FCoE switch setting up the capture trigger to be DCBX or PFC frames.

- vii. Verify the results of the following:
 - a. The PFC parameters in the DCBX frames from both initiators and switch follow the setup definition in "ii".
 - b. (Optional) PFC Pauses expire or are released from the storage and switch and PFC response time is within the spec.
 - c. No drop frames from lossless traffic classes are detected.
 - d. Verify when iSCSI traffic is paused that TCP is not.
 - e. Measure the sustained iSCSI MB/s at the Initiators.
- viii. Go through power cycle with initiators, target, and switch and repeat step "i" to "vi" and record the results described in "vii".

Observable results: PFC requests issued by the switch and CNA response to the PFC request. The throughput of iSCSI write traffic from CNA is reduced. Expect to see no dropped iSCSI frames during PFC pause stage.

Test Track 1 Case 3: PFC Test with RoCE Application

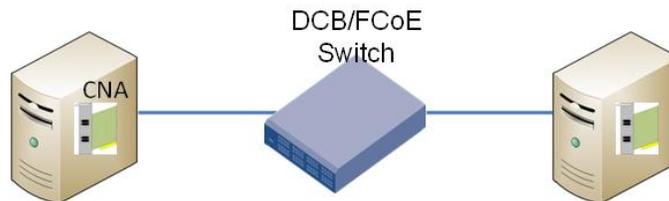


Figure 3: PFC test configuration for RoCE application with DCB or FCoE switch

Procedures:

- i. Configure the devices to enable VLAN on RoCE traffic. Configure VLAN header for RoCE traffic to be: VLAN ID=103.
- ii. Configure the CNA and DCB or FCoE switch to enable PFC on RoCE traffic class
- iii. Disconnect/reconnect HPC server/CNA.
- iv. Launch RoCE traffic between the two HPC servers through the switch; put Jammer in-line of traffic and insert PFC frames to each HPC server.
- v. (Optional) Plug-in Analyzer to monitor the link between CNA and DCB or FCoE switch and setup the capture trigger to be DCBX or PFC frames.
- vi. Verify the results of followings:
 - a. The PFC parameters in the DCBX frames from the servers and switch follow the setup definition in "i".
 - b. (optional) PFC response time is within the spec.
 - c. No dropped RoCE frames are detected.
 - d. Verify when RoCE traffic is paused.
- vii. Go through power cycle with servers and DCB or FCoE switch and repeat step "i" to "v" and record the results described in "vi".

Test Track 2: Pairwise -- ETS Test

For two classes:

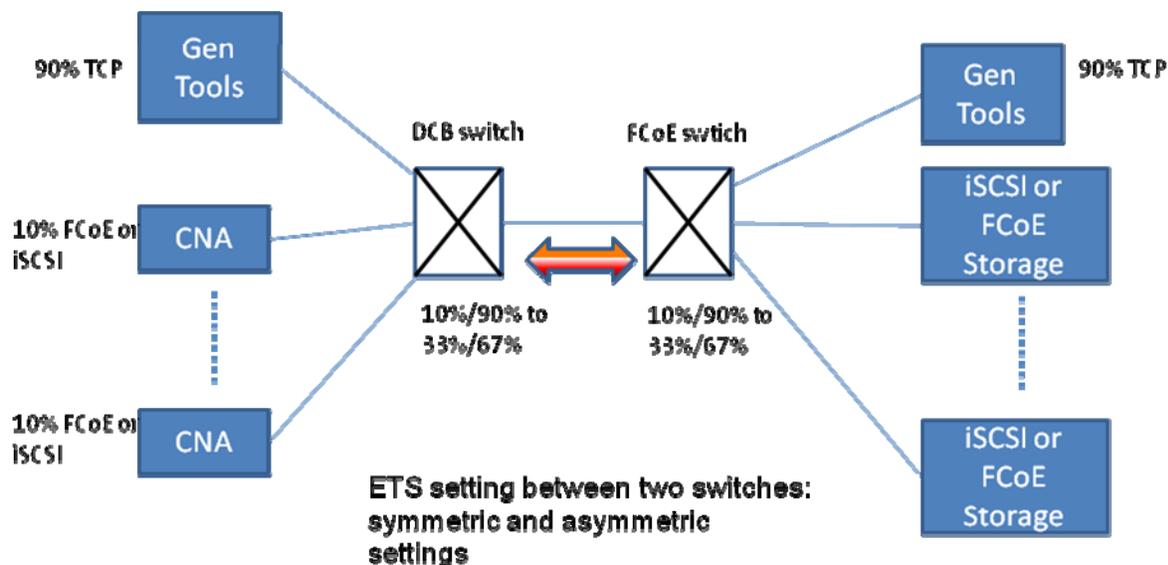


Figure 4: ETS test configuration with two traffic classes

For three classes:

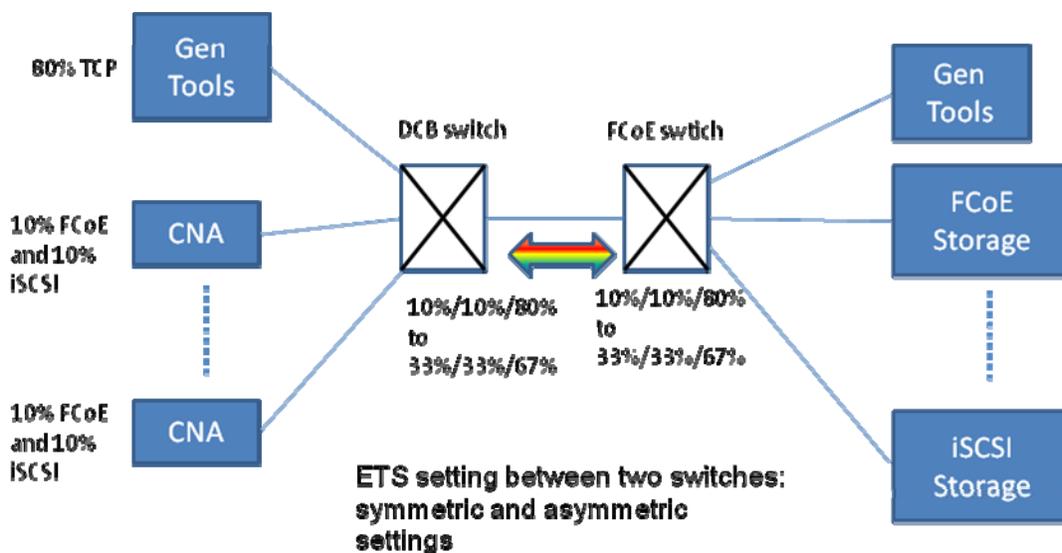


Figure 5: ETS test configuration with three traffic classes

Procedures:

- i. Configure the two generation ports to create two or three (optional) priority traffic, VLAN ID=101, PRIORITY GROUP=3 and VLAN ID=102, PRIORITY GROUP=4, and/or VLAN=1, PRIORITY GROUP=7 to oversubscribed the last generation port.
- ii. Configure ETS parameters of Priority 3 or Priority 4 to be 10% and the rest to 90%. The ETS parameters of two switches set to be symmetric.
- iii. (Optional) For three traffic class test, configure ETS parameters to 10% each of Priority 3 and Priority 4 and the rest to 80%.
- iv. (Optional) Plug-in Analyzer to monitor the inter-switch link; setup the capture trigger to be DCBX or PFC frames.
- v. Start generating write-only FCoE/iSCSI traffic. The traffic should consume as much of available bandwidth as possible. Start the TCP traffic on the same direction and should observe the write traffic drop.
- vi. Verify the results of following:
 - a. The PFC and ETS parameters in the DCBX frames from the DCB/FCoE switch follow the setup definition in "ii", and "iii".
 - b. PFC Pauses are released from the switch to the connected CNA initiators.
 - c. The write traffic drops to 10% as configured in ETS when the TCP traffic starts.
 - d. No drop frames from lossless traffic classes are detected.
- vii. Stop TCP traffic and should observe the write traffic resume back to normal bandwidth.
- viii. Repeat steps "v" to "viii" changing the traffic to read-only. Should observe the similar results except the PFC pause is issued from the switch to the connected storage.
- ix. Change one switch ETS setting to 33% for Priority 3 or 4 and the rest to 67% for asymmetric setup; (Optional) for 3 classes test set ETS to be 33% each for Priority 3 and 4, and the rest to 34%.
- x. Repeat the write only and read only tests.
- xi. Observe the ETS result differences from read and write tests.
- xii. Verify the following:
 - a. The bandwidth allocation should follow the configurations of DCB switch for write-only and FCoE switch for read-only.

Observable results: for symmetric ETS settings, the bandwidth allocation at congestion follows the ETS setup; for asymmetric ETS setting, the bandwidth allocation at congestion follows different ETS setup dependent on the direction of traffic.

Test Track 3: Peer-to-Peer DCBX Test

This test verifies that during FCoE and iSCSI over DCB applications, different states of DCBX parameters can successfully negotiate or gracefully fail on PFC and ETS.

Either pre-IEEE 1.01 or IEEE 1.1 (11/25) version can be used in this test track.



Figure 6: DCBX test configuration

Procedures:

- i. Connect two peers - CNA to switch, storage to switch, switch to switch, or CNA to storage.
- ii. Observe the DCBX exchange process either through device management console or trace capture.
- iii. Verify the followings:
 1. DCBX version matches.

2. PFC Type-Length-Value (TLV) settings are compatible between the two peers.
 3. The willing peer shall have the same settings (in addition to PFC) of unwilling peer.
 4. Error bit of DCBX TLV shall be 0.
 5. (Optional) Application TLV should be compatible between peers.
 6. The DCBX LLDP frames transmitted from each peer periodically; record the period.
- iv. Change the bandwidth allocation with new percentage value from the unwilling peer.
 - v. Verify the following:
 - i. Unwilling peer immediately sends the LLDP frame out with new settings.
 - ii. Willing peer will change the setting accordingly.
 - iii. Error bit of DCBX TLV shall be 0.

Observable results: the willing peer shall follow the settings of unwilling peer and PFC setting for lossless traffic (FCoE, iSCSI-over-DCB, and RoCE) shall be compatible peer-to-peer throughout the whole network.

Test Track 4: 10 GBE Interconnect Interoperability

This test is to verify the copper SFP+ cable (twinax) interoperability among different vendors at same interconnect interface. Vendors need to disable EEPROM checking.

Procedures:

- i. Using the test configuration shown in Test Track 2 #1 or #2 option, connect the server-to-switch and switch-to-storage with same interconnect option but different length cables (5m, 7m, and 20m for DAC).
- ii. Verify that the initiator detects storage targets through the switch.
- iii. Run PFC tests with one lossless traffic following the procedures in Test Track 2 for 10 minutes.
- iv. From the application or generation tools verify the data integrity and make sure there is no link breakdown; record the frame errors captured during the tests (CRC errors).
- v. Record session time-out counts as reference result of link performance.
- vi. Repeat "i" to "iv" with various interconnect options.

Observable results: check that application throughput is not degraded with copper connection.

Test Track 5 (Optional): QCN with iSCSI over DCB Application



Figure 7: QCN test configuration

Procedures:

- i. Configure two DCB capable peers to enable QCN protocol.
- ii. Run iSCSI traffic and verify the CN-Tag in each frame by capturing the trace using Analyzer.
- iii. Send CNM frames from the Congestion Point (CP) to the Reaction Point (RP).
- iv. Measure the sustained iSCSI MB/s and monitor the throughput variation over the time.

Observable results: Every iSCSI frame has CN-Tag with proper configuration according to the spec; RP throttle down the TX traffic while receiving CNM frames.

Results

There was significant progress at this plugfest as the maturity of development continued for each vendor's equipment from switches to initiator and target interconnects. From an interoperability perspective, vendors took the lessons learned from previous plugfest participation and there was a decrease in the amount of interoperability issues.

It is important for end points, both initiators and targets, to be lossless. In order to test this, full bandwidth must be achieved by filling the pipe. This was accomplished utilizing both traffic generation tools and ETS; the switch sent a pause frame and the traffic was managed successfully with no lost frames.

A significant difference from the last plugfest and the May 2010 plugfest was the ability to extensively test interoperability of a multihop, switch to switch, and DCBX networks.

The Priority-based flow control (PFC) Test Results

The PFC capability of each vendor was tested individually. This was accomplished in two steps:

- 1 Prove that the vendor stops transmitting traffic when it receives a PFC Pause frame.
- 2 Prove that the vendor generates a PFC Pause frame when it is unable to keep up with the line rate or because it has been configured to limit the bandwidth for a specific priority.

Testing the Reaction to PFC Pause Frames

To detect PFC Pause frames, the JDSU Xgig protocol analyzer was configured to capture and trigger on a PFC Pause frame. Once a trace with PFC pause frames was captured, the PFC pause frames were located, which priorities were being paused was identified, and all frames that did not belong to these priorities were filtered out. In Figure 7 below, the first PFC frame is pausing priority 3 for 3355 us. A second PFC frame is sent to extend the pausing state, and a third PFC frame with a zero Pause time is sent to release the pause state. The traffic from the initiator did stop at the first pause and resumed after the pause release.

| mm:ss.ms_us_ns (R) | Delta Tim | Port | Protoc | Summary |
|--------------------|-----------|-----------|--------|---|
| 00:00.000_478_005 | | Initiator | FC | FC4SDData; SCSI FCP; Offset = 0x00062000; Len = 0x0800; |
| 00:00.000_479_765 | 1.760 | Initiator | FC | FC4SDData; SCSI FCP; Offset = 0x0007C000; Len = 0x0800; |
| 00:00.000_481_526 | 1.760 | Initiator | FC | FC4SDData; SCSI FCP; Offset = 0x00077800; Len = 0x0800; |
| 00:00.000_483_286 | 1.760 | Initiator | FC | FC4SDData; SCSI FCP; Offset = 0x00077000; Len = 0x0800; |
| 00:00.000_483_927 | 0.642 | Switch | MPCP | PFC; Pause(3) = 3355 us; |
| 00:00.000_520_125 | 36.197 | Switch | MPCP | PFC; Pause(3) = 3355 us; |
| 00:00.000_535_403 | 15.278 | Switch | MPCP | PFC; |
| 00:00.000_536_248 | 0.845 | Initiator | FC | FC4SDData; SCSI FCP; Offset = 0x00011800; Len = 0x0800; |
| 00:00.000_538_008 | 1.760 | Initiator | FC | FC4SDData; SCSI FCP; Offset = 0x00062800; Len = 0x0800; |

The initiator pauses the transmission when receiving the pause frame, and resumes when the pause is released

Figure 8: An initiator stops transmitting when receiving a pause frame.

After manually verifying the proper reaction to PFC Pause frames, Xgig Expert software was used to detect all pause issues automatically on all traces. Using this technique, it was proved that each vendor device was actually pausing when requested; however, different devices took more or less time

to stop transmitting frames after receiving the pause. Some vendors even exceeded the time allowed in the specification.

| | Description | Timestamp | Source | Destination | Ports | Type | Total ... | Value 1 |
|---|---|---------------------|------------------------------|-----------------------------|---------------|------|-----------|-----------------------------|
| 3 | Bad TCP Retransmission (already ACK'ed) | 000:02:41.035_244_; | 10.1.1.100 ; 0000 | 10.1.1.60 ; 0000 | 1,3,1 / 1,3,2 | TCP | 539553 | Seq Num 0 |
| 3 | Frame Received while PFC Class Paused | 000:02:41.039_600_; | VLAN 101 ; 0e:fc:00:d5:00:02 | VLAN 101 ; 00:d5:00:D9:AE:8 | 1,1,1 / 1,1,2 | GigE | 124 | Time since Pause 3.661 (us) |

Figure 9: Xgig Expert reports devices that don't react quickly enough to the Pause requests.

Testing the Generation of PFC Pause Frames

Generation of PFC may be tested with:

- iSCSI and FCoE Initiator/Target devices
- both DCB switches and FCoE Switches

iSCSI Initiator/Targets do not automatically generate PFC Pause frames since they usually rely on the TCP stack ability to recover frame losses. Fibre Channel over Ethernet relies on a lossless fabric and requires a mechanism that resembles the buffer crediting mechanism used in native Fibre Channel fabrics.

Initiator/Target devices will only generate pause frames when they are overwhelmed by the traffic. Since some Initiators and Targets can handle very high traffic throughputs, it can be challenging to create overwhelming traffic streams; however, it was easier to test a switch's ability to generate PFC Pause frames since they could be configured to limit the storage traffic to a certain bandwidth using the ETS features. When ETS is enabled, the switch controls incoming traffic throughput with PFC Pause frames.

ETS Test Results

ETS is the mechanism used to allocate bandwidth per traffic priority group. This requires each frame to contain a VLAN header with the 3-bit priority field (called PCP) set to a proper value.

In ETS, there are 8 possible traffic priorities, 0 through 7, and a minimum bandwidth as percentage of full line rate is allocated to each priority. ETS utilizes the concept of Priority Group to group similar data traffic priorities together; the following discussion assumes 1 priority.

During ETS testing, there were several different allocations of bandwidth configured. We tested with either 1 or 2 storage traffic classes (FCoE and/or iSCSI) configured as no drop services and an additional IP traffic class to create conflicting bandwidth levels triggering flow control commands to protect the lossless data classes.

In one scenario, the ETS feature of a switch was configured as follows:

- Priority 3: Minimum of 10% full line rate bandwidth
- Priority 4: Minimum of 10% full line rate bandwidth
- Other Priorities: Minimum of 80% full line rate bandwidth

With 10 Gbps traffic, this translates to:

- Priority 3: Minimum of ~120 MB/s bandwidth
- Priority 4: Minimum of ~120 MB/s bandwidth
- Other Priorities: Minimum of ~960 MB/s bandwidth

For the ETS feature to become active, we needed to flood the link between the initiator and the switch with non-Priority 3 and Priority 4 traffic. Ixia, JDSU, and Spirent all provided traffic generators to send full line rate IP traffic through the switch to the initiator. At that point, the switch started to enforce the ETS feature. The switch dropped frames on the IP traffic and sent PFC Pause frames to the Priority 3 FCoE target as well as the Priority 4 iSCSI target. The traffic generator software was used to graph traffic throughput per priority. The following illustration shows that the switch enforced its ETS settings when the load tester ramped up traffic to full line rate. In fact, when the link became saturated with load tester traffic, the FCoE traffic on priority 3 was forced to 10% of the line rate and the iSCSI traffic on priority 4 was also forced to 10% of the line rate.

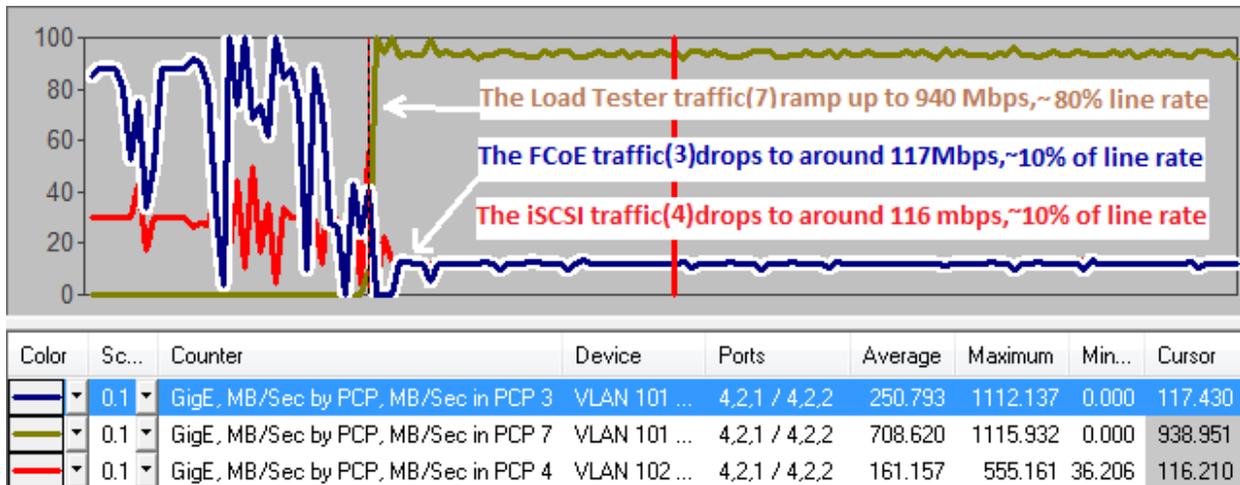


Figure 10: The traffic throughput is plotted per priority to verify the ETS Test Results.

The traffic generator tools all provided methods to ensure that no frames were dropped on the FCoE target side. The analyzer software ensured that all the frames on one side of the switch were also found on the other side and they were in the same order. The analyzer software did not detect any issue on the FCoE target side. This means that the switch did not drop frames while implementing flow control on the traffic.

In another test, we configured the switches with different ETS settings in a multi-hop DCB environment. In this scenario we configured the test bed as follows:

DCB switch configured with:

- Priority 3: Minimum of 10% full line rate bandwidth
- Priority 4: Minimum of 10% full line rate bandwidth
- Other Priorities: Minimum of 80% full line rate bandwidth

Upstream DCB switch / FCF configured with the following values:

- Priority 3: Minimum of 33% full line rate bandwidth
- Priority 4: Minimum of 33% full line rate bandwidth

- Other Priorities: Minimum of 67% full line rate bandwidth

We observed in this test that the traffic shaping of ETS was dependant on the switch a device was connected to. Initiators writing traffic from the DCB switch to targets on the Upstream DCB/FCF switch followed the ETS values of 10%, 10%, 80%, and the initiators that were on the FCF switch sending traffic to the targets on the DCB switch were engaged at 33%, 33%, 67%.

Both switches were able to generate PFC Pause frames and implement the ETS feature were tested. This test proved that the switches (both switches in the multi-hop environments tested) were enforcing ETS and generating PFC Pause frames when needed.

The DCBX Test Results

The Data Center Bridging capability eXchange Protocol (DCBX) adds new TLVs in a LLDP frame. It is used to announce ETS and PFC configurations on a link-by-link basis. Each switch and end device announces its ETS and PFC configurations through DCBX messages to its peers. Each peer has a choice of accepting or ignoring that configuration. The DCBX announcements occur several times when an end device connects to a switch and then approximately every 30 seconds. The DCBX protocol defines how link capabilities are discovered and what happens in the case of conflicting configurations on a per-feature basis.

To test DCBX, we connected end devices and switches and captured the DCBX exchanges. The exchanges were captured and examined using a DCBX-specific configuration of the Xgig TraceView software. TraceView allows observation of the important DCBX settings for each frame on a single line making it easy to follow state transitions from one frame to the other. The following figure illustrates several DCBX exchanges between the CNA end device and the switch.

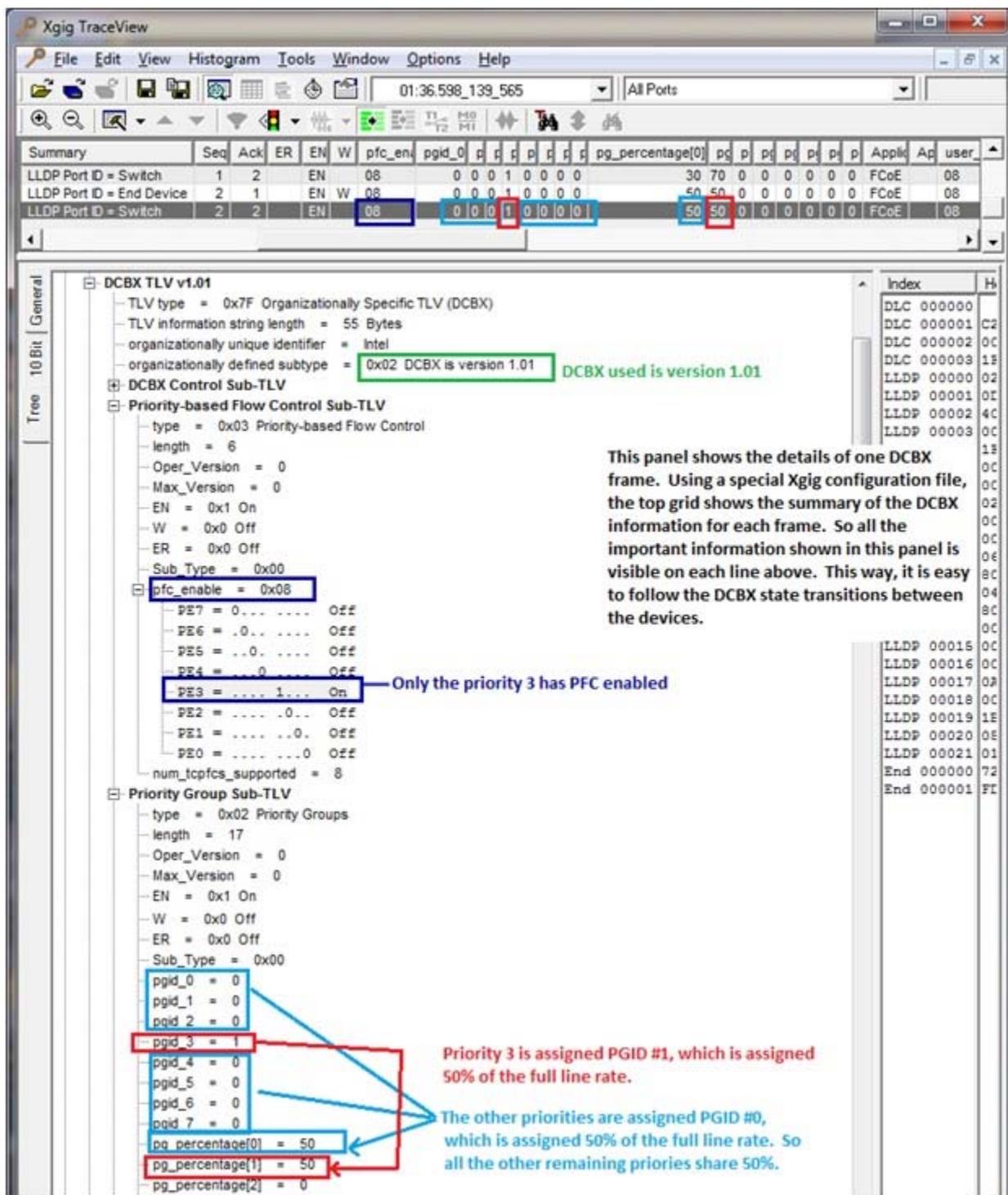


Figure 11: The trace shows DCBX parameter exchanges.

10GbE Interconnect Test Results

For the first time during this DCB Plugfest, we tested compatibility of participating SFP+ cable (twinax) vendors. On the switch side, we disabled any EEPROM checking so that no matter the cable connected, the link would be brought up on the interconnect. We tested cables at 3, 5, 7, and 10 meters and checked application throughput during each of tests. Baseline connectivity was done using fiber as the media to validate interconnect compatibility.

For each test, 10 minutes of FCoE traffic was generated between each of the Vendors initiators, targets and interconnect switches. For the most part, traffic was seen to pass across the links with no degradation in performance or errors on the links.

Glossary of Terms

Converged Network Adapter (CNA): a Virtual bridged Local Area Network host adapter that is capable of driving TCP and/or iSCSI-over-DCB, and/or FCoE, and/or iWARP traffic applications.

Data Center Bridging (DCB): an 802.1Q Bridge which implements protocols and capabilities for use in the data center.

Data Center Bridging capability eXchange Protocol (DCBX): a protocol that is used for conveying capabilities and configuration of DCB features between link neighbors to ensure consistent configuration across the network. This protocol leverages functionality provided by 802.1AB (LLDP).

DCB switch: a Virtual-bridged Local Area Network (VLAN) switch that supports DCB functionalities but requires external handling of FCoE encapsulation and FCF functions. A DCB switch does not report the DCBx version 1.01 TLV required for FCoE.

Enhanced Transmission Selection (ETS): IEEE 802.1Qaz task force has defined Enhanced Transmission Selection (ETS) to provide a common management framework for assignment of bandwidth to traffic classes.

FCoE switch: a Virtual bridged Local Area Network (VLAN) switch that supports both DCB and FCF functionalities. An FCoE switch reports the DCBx version 1.01 TLV required for FCoE.

Fibre Channel (FC): a network technology primarily used in a storage area network (SAN).

Fibre Channel Forwarder (FCF): a Virtual bridged Local Area Network (VLAN) connection that is able to transport FC frames (via encapsulation into FCoE frames) over Ethernet through a DCB switch and act as a Fibre Channel forwarder as defined in INCITS T11/FC-BB-5 for FIP identification and encapsulation in jumbo Ethernet frames.

Fibre Channel over Ethernet (FCoE): the proposed mapping of [Fibre Channel](#) frames over full duplex [IEEE 802.3](#) networks.

Fibre Channel Protocol (FCP): a transport protocol (similar to TCP used in IP networks) which predominantly transports SCSI commands over Fibre Channel networks.

Internet Small Computer System Interface (iSCSI): an [Internet Protocol](#) (IP)-based storage networking standard for linking data storage facilities. By carrying SCSI commands over IP networks, iSCSI is used to facilitate data transfers over intranets and to manage storage over long distances.

iSCSI-over-DCB storage: an iSCSI storage device that can support an iSCSI I/O application over a DCB network that uses PFC, ETS and QCN to manage iSCSI flows.

Internet Wide-Area RDMA Protocol (iWARP): iWARP (Internet Wide-Area RDMA Protocol) provides a low-latency / high-bandwidth solution for standards-based Ethernet by defining how the RDMA (Remote Direct Memory Access) protocol runs over TCP/IP. iWARP delivers improved performance over Ethernet by avoiding application context switching, eliminating intermediate buffer copies, and accelerating TCP/IP (Transport) compute.

RDMA over CE (RoCE): a light-weighted transport protocol defines how RDMA layered directly over Ethernet L2 and transport on the enhanced Ethernet which is based on DCB technologies.

Link Layer Discovery Protocol (LLDP): a vendor-neutral Layer 2 protocol that allows a network device to advertise its identity and capabilities on the local network. The protocol is formally defined as [IEEE standard 802.1AB-2005](#).

Priority-based Flow Control (PFC): IEEE 802.1Qbb is standardizing Priority-based Flow Control (PFC) which 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.

Quantized Congestion Notification (QCN): IEEE 802.1Qau is a Layer 2 congestion control protocol in which a congested network device can control the rates of Layer 2 sources whose packets are passing through it.

Type-Length-Value (TLV): within [data communication protocols](#), optional information may be encoded as a type-length-value or TLV element inside of the protocol.

The type and length fields are fixed in size (typically 1-4 bytes), and the value field is of variable size. These fields are used as follows:

- Type:** A numeric code which indicates the kind of field that this part of the message represents.
- Length:** The size of the value field (typically in bytes).
- Value:** Variable sized set of bytes which contains data for this part of the message.

Quality of Service (QoS): In the field of computer networking and other packet-switched telecommunication networks, the traffic engineering term quality of service (QoS) refers to resource reservation control mechanisms rather than the achieved service quality. Quality of service is the ability to provide different priority to different applications, users, or data flows, or to guarantee a certain level of performance to a data flow. For example, a required bit rate, delay, jitter, packet dropping probability, and/or bit error rate may be guaranteed.



About Ethernet Alliance

The Ethernet Alliance is a global 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.