# Data Center Bridging Tutorial

Mikkel Hagen

Research and Development

University of New Hampshire – InterOperability Laboratory

The convergence of technologies into a single fabric provides several benefits such as: less space, less heat, less electricity, less knowledge, less maintenance. All of these benefits amount to one major benefit that all site administrator can appreciate: less cost. Ethernet has become a ubiquitous technology. It is available in nearly every computer and almost everyone knows how to plug in a cable for it to "just work". It continues to scale in performance from 10 to 100 to 1000 Mbps and beyond to 10G and now 40/100G. As it continues to ramp up into new levels performance the cost continues to plummet, what once cost hundreds for GigE now is on motherboards and is in the "only a couple of dollars" range. 10GigE is quickly coming down from thousands of dollars per port to only hundreds now. All of this creates a great opportunity for other technologies to converge on Ethernet as a universal fabric.

Fibre Channel has long provided its own fabric, which has recently lagged Ethernet in performance. As 10GigE becomes more common Fibre Channel is moving towards 8Gig and as Ethernet develops the next generation of 40G/100G, Fibre Channel is developing 16G. For all of the benefits noted above, Fibre Channel over Ethernet has been introduced. This is essentially a new technology to tunnel Fibre Channel frames within Ethernet MAC frames. The major issue with this technology is that Fibre Channel is not designed to work on an unreliable network. Fibre Channel was developed with the idea of having a dedicated fabric that loses little to no packets. For this reason, Ethernet has developed a set of standards termed Data Center Bridging.

Data Center Bridging consists of four different technologies: Per-Priority PAUSE, Enhanced Transmission Selection, Congestion Notification and DCB Exchange[1]. Each technology works independently to provide enhanced Ethernet features and together they attempt to eliminate any packet loss due to congestion. Per-Priority PAUSE adds fields to the standard PAUSE frame that allow a device to inhibit transmission of frames on certain priorities as opposed to inhibiting all frame transmission. Enhanced Transmission Selection provides a means for network administrators to allocate link bandwidth to different priorities on a percentage of total bandwidth basis. Congestion Notification is a mechanism to transmit congestion information on an end-to-end basis per traffic flow. Finally, DCB Exchange is the mechanism in which peers can exchange capabilities to one another.

Per-Priority PAUSE, or Priority-based Flow Control (PFC), is defined in the 802.1Qbb standard[2]. It is only defined for full duplex MACs and allows flow control on a Per-Priority basis. It is invoked by clients of the MAC Control Sublayer through MAC Control PFC PAUSE primitives. PFC is used to inhibit transmission of data frames from one or more of the earlier defined eight priorities for a specified period of time. PFC cannot be used to inhibit MAC Control frames. Each PFC PAUSE frame contains an array of 8 fields containing a 2 octet priority_enable_vector field and a 2 octet time_vector field. The priority_enable_vector field indicates for each of the eight priorities which time_vector fields are valid and should be acted upon. The time_vector fields indicates an unsigned int length of time for which the transmission should be inhibited. The time value is measured in units of pause_quanta, equal to 512 bit times of the particular PHY layer. The range of valid pause times is 0-65535 pause_quanta.

Enhanced Transmission Selection (ETS) is defined in the 802.1Qaz standard[3]. ETS introduces a new 4bit fields called the Priority Group ID (PGID). There are 16 PGID values with 15

being a special “No Bandwidth Limit” value and 8-14 being reserved values. ETS allows one or more priority to be assigned to a PGID. Each PGID is allocated a percentage of available bandwidth on the link. Available bandwidth refers to the maximum percentage of available link bandwidth after priorities within PGID 15 are serviced. Once allocated, a PGID may only use available bandwidth up to the maximum percentage allocated.

Congestion Notification (CN) is defined in the 802.1Qau standard[4]. A consequence of link level pausing (i.e. 802.1Qbb) is "congestion spreading" - the domino effect of buffer congestion propagating upstream causing secondary bottlenecks. A layer two congestion control algorithm allows a primary bottleneck to directly reduce the rates of those sources whose packets pass through it, thereby preventing secondary bottlenecks.

Congestion notification is broken up into two algorithms: CP and RP. CP, Switch or Congestion Point Dynamics is the mechanism in which a switch buffer samples incoming packets and generates a feedback message addressed to the source of the sampled packets with the extent of the congestion. RP, Rate Limiter or Reaction Point Dynamics is the mechanism by which a Rate Limiter (RL) decreases its sending rate based on feedback and increases its rate voluntarily to recover lost bandwidth and probe for available bandwidth.

CP computes a congestion measure and with a probability depending on the severity of congestion. For example, as congestion gets higher there is a higher likelihood of randomly sampling the buffer and sending a congestion notification message (CNM) to the source of the sampled packet indicating congestion level.

RP will decrease rate proportional to the degree of congestion reported in the CNM received. Since Ethernet does not contain acknowledgements there is no feedback mechanism in which to increase rate once limited, so a timer is implemented. The rate increases in two phases: Fast Recovery and Active Increase. In Fast Recovery (FR) the RL will increase its rate by 1/2(Current Rate + Target Rate) every 150KBytes transmitted at the reduced rate if no more CNM arrive. This will occur for 5 cycles. Active Increase (AI) phase begins after 5 successful cycles of FR. During AI, the RL will probe for additional bandwidth by updating the Target Rate and Current Rate in 50 packet cycles.

Devices in a Virtual Bridged Network that are configured to support a Congestion Notification Priority form what is called a Congestion Notification Domain (CND). Congestion Notification Priority (CNP) consists of one value of the priority parameter such that all devices in a CND are configured to assign frames at that value to the same CP and/or RP. Different priorities coincide with different applications or even single applications. Frames with the same priority value and all assigned to a single flow queue and RP in the originating end station form a Congestion Controlled Flow (CCF). Every frame in a CCF carries a CN-tag. The CN-tag contains a FlowID. The FlowID and destination address are the only means in which to identify a target of a CNM. The FlowID is not to be interpreted, it is purely to identify a flow within a RP. End Stations add the CN-tag including the FlowID using a mechanism beyond the scope of the standard. As both FlowID and Destination Address are used to identify a unique flow, different end stations can use the same FlowID without a problem.

Data Center Bridging Exchange (DCBX) protocol is defined in the 802.1Qaz standard[3]. DCBX is used to exchange configuration information with the directly connected peer. The protocol may also be used for misconfiguration detection and for configuring the peer. DCB exchanged parameters are packaged into Organizationally Specific TLVs transmitted via the LLDP protocol. Exchanged parameters are broken up into two different groups: Administered and Operational. Administered parameters are those that are configured. Operational parameters are those that are the operational state of the administered parameter, which may or may not be the same. They can change due to exchanges with the peer and are only present for administered parameters that can be changed by the peer. DCBX is expected to operate only over a point to point link and if multiple peers are discovered, the peer's TLVs should be ignored until the multiple peers condition is resolved. DCBX currently only has TLVs defined for Priority Groups, PFC and Applications.

## References


1. IEEE 802.1 Data Center Bridging Task Group. http://www.ieee802.org/1/pages/dcbridges.html
2. IEEE 802.1Qbb/D1.0. Draft Standard for Local and Metropolitan Area Networks – Virtual Bridged Local Area Networks – Amendment XX: Priority-based Flow Control. February 9, 2009.
3. IEEE 802.1Qaz/D0.2. Draft Standard for Local and Metropolitan Area Networks – Virtual Bridged Local Area Networks – Amendment XX: Enhanced Transmission Selection for Bandwidth Sharing Between Traffic Classes. November 24, 2008.
4. IEEE 802.1Qau/D1.4. Draft Standard for Local and Metropolitan Area Networks – Virtual Bridged Area Networks – Amendment 7: Congestion Notification. February 12, 2009.