Credit based Link Level Flow Control and Capability Exchange Using DCBX for CEE ports.

Keshav Kamble (kkamble@us.ibm.com)
Jeffery Lynch
Renato Recio
Casimer DeCusatis
Mitch Gusat
Cyriel Minkenberg
Vijoy Pandey
IBM Corporation
Credit Exchange Logic

• A new frame i.e. Credit Exchange Frame (CE Frame) with a MAC Control Ethernet Type 0x8808. OPCODE 0x0110/0x0111. (Slide 4)
• A new layer 2 tag named CE-Tag or Credit Exchange Tag. (Slide 5)
• To ensure loss-less delivery of frames where the frames are of different priorities ranging from 0 to 7, sender (MAC TX) optionally requests credits by sending a new CE frame to the peer port. This frame contains credit requests for multiple priorities per tenant slice (channel Id).
• To ensure loss-less delivery of frames of one specific priority, the sender (MAC TX) adds a CE-Tagged frame or piggy backs on the data frame in transmission logic. CE-Tag contains request for credits for single priority and has low overhead. Optional!
• The peer ports (receivers or MAC RX) receive the CE Frame or CE-Tagged frame and interprets the requests for corresponding priorities / priority.
• Calculated amount of credits are issued to the peer ports in a CE frame or adding a CE-Tag on next data frame being transmitted.
• Upon receipt of credits, sender sends frames for the appropriate / allowed priorities per tenant.
• Unit of credit exchange is 512 bits or each peer port can decide its own unit of credit as per it local TM / Buffer Mgr implementation.
• Suggest extension to the DCBX TLV to exchange capabilities of credit exchange and credit unit selection.
• Credit aging and timers. Consideration to avoid loss of credits.
Credit Exchange

- Credit Requested = F(size of egress queue or number of frame/s to be transmitted, port speed, tenant slice bandwidth, RTT delay, prediction factor, etc.).
- Priority Value = IEEE 802.1P value of the frame to be transmitted OR COS queue priority value.
- Channel Id = F(S-Tag or other multi-tenancy unit).
- Credits granted = F(credits requested, TX buffers available, credits available for xchange, port shaper settings, global prediction factor, etc.)
- Credit Units = Min 512 bits to max of MTU size.

- Algorithmic Steps:
  1. Exchange Capabilities (Port A, B).
  2. Request Credits (Optional).
  3. Grant Credits.
  4. Transmit data.
  5. Receive data.
  6. Accounting.
CE Frame Format

| 01:80:C2:00:00:01 | SMAC (48 bits) | 0x8808 | MAC Control OpCode | TC Selection Vector | TC Credit Vector | Pad | CRC |

- **Standard**
- **Credit Req = 0x0110 / Std**
- **Credit Resp = 0x0111 / Std.**
- **Reserved 8 bits + P[7],P[6],........P[0]**

- Response Credits for P0 or value = 0 in request packet.
- Response Credits for P1 or value = 0 in request packet.
- Response Credits for P2 or value = 0 in request packet.
- Response Credits for P3 or value = 0 in request packet.
- Response Credits for P4 or value = 0 in request packet.
- Response Credits for P5 or value = 0 in request packet.
- Response Credits for P6 or value = 0 in request packet.
- Response Credits for P7 or value = 0 in request packet.

- **16 bit Channel Number (Optional value, default = 0x0)**
Frame Format with CE-Tag

Credit Exchange Tag (64 bits)

DMAC | SMAC | CE Tag Eth Type | TC Selection Vector | TC Credit | Data | Pad | CRC

- Standard or Programmable for Req and Response. 16 bits
- Reserved 8 bits + P[7], P[6], ..., P[0]. Select only one P bit.
- Response Credits for the selected priority or value = 0 in request packet. 16 bits.
- 16 bit Channel Number (Optional value, default = 0x0)

Ethernet Frame Data
## Flow Control Methods

<table>
<thead>
<tr>
<th>Port Speed</th>
<th>IEEE 802.3X Flow Control</th>
<th>IEEE 802.1Qbb PFC</th>
<th>Credit based Flow Control (CFC)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>10Mbps-1Gbps</td>
<td>YES</td>
<td>-NA-</td>
<td>-NA-</td>
<td></td>
</tr>
<tr>
<td>10Gbps</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>Priorities 0-7. Programmable Flow Control method per priority.</td>
</tr>
<tr>
<td>40Gbps</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>Same as above.</td>
</tr>
<tr>
<td>100Gbps</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>Same as Above. Additionally, consideration to 16 bit tenant or bandwidth slice.</td>
</tr>
<tr>
<td>400Gbps</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>Same as above.</td>
</tr>
<tr>
<td>1000Gbps (or 1.6Tbps ?)</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>Same as above.</td>
</tr>
</tbody>
</table>
Exchange of CET Capabilities

• DCBX protocol extension to exchange:
  ▪ Credit Exchange Capabilities for flow control between peer ports.
  ▪ Unit of credit exchange. The unit value can be in number of chunks of bits where a chunk length can be 512 bits to MAX_CR_BLK.
  ▪ MAC_CR_BLK can be MTU size supported or 2548 bytes or any other number > 512 bits. 2548 bytes are adequate to accommodate FCOE frame with security encryption.
  ▪ Capability to exchange channeling capabilities which enables adding 16 bit channel Ids (e.g. like SVID) in credit exchange frame or CE-Tag.
Flow Control Capability Exchange

Device A

DCB MIB

Exchanged parameters
Operational parameters
Local parameters

Device B

DCB MIB

Exchanged parameters
Operational parameters
Local parameters

LLDP Messages

Ethernet Link
DCBX Extension / Brief Algorithm

- The DCBX protocol can be extended to advertise the Credit Exchange capabilities.
- New Application Protocol TLV, “CET” should be defined.
- This TLV should be originated by a physical port at a peer to peer basis.
- Switching devices down the line should exchange such TLV messages on all of their DCBX member ports to their peer devices. (OPTIONAL)
- Thus, after complete convergence, a complete path from source to the destination can understand the credit exchange capabilities of their peer ports.
BACKUP
Problem Statement

• Current CEE port based flow control works based on post queuing status on ingress.
• Ethernet needs more predictable and dynamically controllable flow control and frame acceptance mechanism with request-response mechanism.
• Interaction between peer ports before data exchange brings in more certainty and better resource utilization. Ethernet needs such a mechanism for purpose specific applications and convergence of Infiniband over Ethernet.
• Ethernet needs credit based flow control mechanism to bring in more certainty for transmission.
• Enable CEE and Metro Ethernet to have reliable I/O convergence over longer distances and simultaneously reduce buffering overheads.
• Capability to exercise PFC or CFC (Credit Based Flow Control) for select priorities.
E2E Exchange of CET Capabilities

DCBX Supported Cloud with CET capabilities

Bi directional DCBX Exchange

Non-CET capable Cloud