Enhancements to Traffic Scheduling and DCBX

Anoop Ghanwani, Dell

Joe White, Dell

Acronyms

DCBX	Data Center Bridging eXchange
ETS	Enhanced Transmission Selection
RoCE	RDMA over Converged Ethernet
SP	Strict Priority
TC	Traffic Class
TSA	Traffic Selection Algorithm

Overview

- Review of traffic scheduling and DCBX in 802.1Q-2012
- Use cases that benefit from strict priority
- Limitations of 802.1Q-2012 TSAs
- Proposed enhancements to TSAs and DCBX

Review of 802.1Q-2012

- A TC is assigned a TSA
  - Strict priority
  - Credit-based shaper
  - ETS
  - Vendor-specific

Review of 802.1Q-2012 (2)

- ETS, Clause 37.3
  - TCs configured for ETS receive bandwidth in proportion to a configured weights for bandwidth that is available to ETS classes
    - After SP and credit shaper classes have received service
  - If one of the TCs is not using all of its assigned bandwidth, excess bandwidth is used by other TCs
  - How excess bandwidth is shared is NOT specified

Review of 802.1Q-2012 (3)

- DCBX is used for capabilities advertisement and configuration
  - Mapping of priority to TC
  - Specify TSA for each TC
  - For ETS classes, specify Bandwidth % for each TC

Use Cases for Strict Priority

- SP is important for low latency applications
  - High frequency trading
  - RoCE applications, e.g. Microsoft SMB Direct
  - Control traffic, e.g. heartbeats, sync messages
- In many deployments using SP with more than one queue is desirable

Limitations of 802.1Q-2012 TSAs

- Strict priority
  - No way to limit the bandwidth consumed by a TC
    - May be required by SLA
  - Starvation of lower TCs is possible
- Credit-based shaper
  - Not configurable via DCBX
  - Limits the bandwidth that can be consumed by a TC
- ETS
  - Latency properties
    - Different TCs will likely experience similar latency
  - Excess bandwidth distribution
    - Excess bandwidth distributed is not specified
    - An implementation using WRR would assign it in proportion to bandwidth %

Addressing the Limitations of TSAs in 802.1Q-2012

- Two new controls
  - Minimum bandwidth guarantee (MinBG)
  - Maximum bandwidth limit (MaxBL)
- These can be applied to TCs with SP or ETS

Minimum Bandwidth Guarantee

- Each queue in the system will first receive access to its MinBG
  - In order of priority
  - This allows a lower priority queue to receive service up to a certain bandwidth, once all higher priority queues have received their MinBG
- Once the MinBG is satisfied for all TCs, the system reverts to “normal” operation
  - TCs are serviced in the order determined by the TSAs

Maximum Bandwidth Limit

- Any queue that has achieved its MaxBL is removed from service
  - Stops receiving more service even if there is no other traffic in the system
  - Addressed by credit-based shaper in 802.1Qav

Example #1

TC/Queue	TSA	MinBG	MaxBL	Offered load	Output
3	SP	-	60%	100%	60%
2	SP	10%	40%	100%	20%
1	ETS	10%	-	100%	10%
0	ETS	10%	-	100%	10%

- First, TC 2, 1, 0 each receive their MinBG in that order
- Next, TC 3 is serviced till it reaches MaxBL
- Finally, TC 2 receives the remainder
  - Does not exceed its MaxBL
- The order of service is important as it impacts latency

Example #2

TC/Queue	TSA	MinBG	MaxBL	Offered load	Output
3	SP	60%	60%	40%	40%
2	SP	10%	40%	30%	30%
1	ETS	10%	-	0%	0%
0	ETS	10%	-	100%	30%

- First, TC 3 receives 40%, within MinBG & MaxBL
- Next, TC 2 receives 10%, its MinBG
- Next, TC 0 receives 10%, its MinBG
- Next, TC 2 receives 20%, offered load within MaxBL
- Next, TC 0 receives 20%, the remainder

Proposed Enhancements to TSAs and DCBX

- TSAs
  - Define the behavior of MinBG and MaxBL as they apply to SP and ETS
- DCBX
  - Define new TLVs in DCBX to configure MinBG and MaxBL for each TC

Next Steps

- Motion to create a PAR to enhance 802.1Q?

Earlier work in IEEE 802.1

- Similar concepts have been discussed before
  - http://www.ieee802.org/1/files/public/docs2005/new-congdon-improved-queuing-0505.pdf
  - http://www.ieee802.org/1/files/public/docs2005/new-congdon-improved-queuing-0705.pdf

THANK YOU