Congestion Isolation PAR & CSD

Paul Congdon

Background

- This is the text that was edited on the DCB phone conference – Jan 9th, 2018. It is now represented in the following documents:
  - new-dcb-congdon-draft-congestion-isolation-PAR-0118-v02.pdf
  - new-dcb-congdon-draft-congestion-isolation-CSD-0118-v01.doc
- Only new or major blocks of text are included. Refer to the above documents for the complete PAR and CSD.

5.2.b Scope of the project:

- This amendment specifies protocols, procedures and managed objects that support the isolation of congested data flows within networks of limited bandwidth delay product. This is achieved by enabling bridges to individually identify flows creating congestion, adjust transmission selection for packets of those flows and signal to the upstream peer. This mechanism avoids head-of-line blocking for uncongested flows sharing a traffic class in lossless networks. Congestion Isolation is intended to be used with higher layer protocols that utilize end-to-end congestion control in order to reduce packet loss and latency.

5.3 Is the completion of this standard dependent upon the completion of another standard: Yes

- This amendment will specify a new Link Layer Discovery Protocol (LLDP) Type-Length-Value (TLV) and its associated YANG model. Project IEEE 802.1ABcu is currently specifying the YANG model for IEEE Std 802.1AB which must be completed in order for this amendment to specify its extension.

5.5 Need for the Project:

- There is significant customer interest and market opportunity for large scale, low-latency, lossless Ethernet data centers to support high-performance computing and distributed storage applications. Congestion is the primary cause of loss and delay in these environments. These environments currently use higher layer end-to-end congestion control coupled with priority-based flow control at Layer 2 to avoid performance degradation from packet loss due to congestion. As the Ethernet data center network scales in size, speed and number of concurrent flows, the current environment creates head-of-line blocking for flows sharing the same traffic class. Isolating flows that cause congestion reduces latency for flows not causing congestion and improves the scale and performance of the Ethernet data center network. This amendment will support the identification and isolation of the higher layer protocol flows creating congestion and will interoperate with existing IEEE 802 and higher-layer congestion management capabilities. Use of a consolidated Ethernet data center network will realize operational and equipment cost benefits.

5C requirements
1.2.1 Broad market potential

a) Broad sets of applicability.

- Congestion is the primary reason for loss in data center networks with a limited network bandwidth-delay product. Higher layer congestion control protocols are widely deployed in those networks to reduce performance degradation due to loss. The higher layer protocols are limited in their ability to fully mitigate loss as data center networks scale in size and expand with higher-speed links. To eliminate loss in data center networks, the higher layer congestion control protocols can be combined with priority based flow control; however, negative consequences of head-of-line blocking and congesting spreading have been observed. Congestion Isolation improves the effectiveness of widely used higher layer congestion control protocols by isolating the flows that are causing congestion and providing additional time for the end-to-end protocols to react. Congestion Isolation can be applied to all current data center environments as well as future converged high-performance computing environments.

5C requirements
1.2.1 Broad market potential

b) Multiple vendors and numerous users.

- Multiple equipment and integrated circuit vendors have expressed interest in the proposed project. There is strong and continued interest from data center network operator in converging specialized high-performance networks to Ethernet and in the realization of operational and equipment cost savings through use of a consolidated network. Furthermore, numerous vendors are building new high-speed solid-state data storage access solutions over Ethernet networks, presuming that they can be realized with familiar technology and a consolidated network.

5C requirements
1.2.3 Distinct Identity

- IEEE Std 802.1Q is the sole and authoritative specification for VLAN-aware Bridges and their participation in LAN protocols. The existing congestion notification in IEEE Std. 802.1Q is distinctly different in that it signals across the L2 network to a source reaction point and does not perform any flow isolation at the congestion point. Congestion Isolation performs flow isolation locally within the bridge, limits signaling to the next hop neighbor and does not define an equivalent reaction point. In order to support re-use and ease implementation efforts, congestion isolation incorporates and derives applicable procedures and protocols from congestion notification. The proposed amendment is intended to handle the short term effects of congestion while higher layer congestion control protocols, such as those based on IETF's explicit congestion notification (ECN), moderate the sources of congested traffic. No other IEEE 802 standard addresses congestion isolation by bridges.

5C requirements
1.2.4 Technical Feasibility

a) Demonstrated system feasibility.

- The proposed amendment incorporates techniques for flow identification and traffic scheduling that are currently available in many production data center switches.

b) Proven similar technology via testing, modeling, simulation, etc.

- Performance improvements and a reduction in head-of-line blocking have been demonstrated through simulation and analysis of methods consistent with the proposed project. The proposed project can identify flows that are causing congestion by using techniques similar to the existing IEEE Std. 802.1Q congestion notification and IETF's explicit congestion notification (ECN). Flow identifying information can be maintained for congested flows. This information is similar to the information need to support existing features such as Access Control Lists (ACLs).

5C requirements
1.2.5 Economic Feasibility

a) Balanced costs (infrastructure versus attached stations).

- The proposed amendment does not significantly change the cost characteristics of bridges and does not require additional traffic classes. Implementations of the proposed amendment may need to include a flow table that can recognize flows that have been identified as causing congestion. Simulation data has shown that size of this table can be small and limited because the congested flows are only a small fraction of the overall flows in transit. The proposed amendment operates independently of attached stations and does not impact station implementation.

b) Known cost factors.

- The proposed solution can reduce overall cost of data center networks by improving scaling and consolidating dedicated high-performance computing networks on to a common Ethernet fabric. This allows data center operators to consolidate their storage and computing networks, and to run traffic in a lossless environment that helps minimize flow completion time.

5C requirements
1.2.5 Economic Feasibility

c) Consideration of installation costs.

- Installation costs of data center bridges are not expected to be significantly affected.

d) Consideration of operational costs (e.g., energy consumption).

- The proposed amendment uniquely incorporates existing technologies and as a consequence is not expected to significantly affect the operational cost of data center networks. A small amount of additional configuration is required, but fewer bridges and links need to be configured because the proposed amendment reduces the need to overprovision the data center network to achieve similar performance.

8.1 Explanatory

- 5.3: IEEE Std 802.1AB refers to IEEE 802.1AB-2016 - IEEE Standard for Local and metropolitan area networks - Station and Media Access Control Connectivity Discovery
- 5.3: Project IEEE 802.1ABcu refers to P802.1ABcu - Standard for Local and Metropolitan Area Networks - Station and Media Access Control Connectivity Discovery Amendment: YANG Data Model