Priority-Based Flow Control (PFC): Draft PAR & 5 Criteria

Data Center Bridging Task Group

Introduction

- Objective
  - Enable 802 full-duplex LANs within Data Centers to provide flow control characteristics similar to that of IPC and storage fabrics (e.g., Fibre Channel and InfiniBand)
    - Does not eliminate all frame loss
    - Eliminates frame loss due to temporary congestion
      - PFC and Congestion Notification eliminate the need for end nodes to deal with congestion at higher levels (back-off, slow restart, etc.)
  - Allow conventional traffic to co-exist on such LANs
    - Not all traffic is flow controlled
    - Segregated by priority code points (as is done with Congestion Notification)

Why is this needed?

- See the Data Center Needs for I/O Consolidation:
  - http://www.ieee802.org/802_tutorials/nov07/Data-Center-Bridging-Tutorial-Nov-2007-v2.pdf

Data Center Bridging Networks

- Data Center Bridging
  - Applies to end stations and bridges
  - Limited to short-range networks
    - Limited bandwidth-delay
    - Low hop-count
  - Includes:
    - Congestion Notification
    - Enhanced Transmission Selection
    - Priority-based Flow Control (PFC)
    - Discovery and Capability Exchange Protocol
- No intention to extend this to larger topologies (e.g. wide area)

Where is Priority-Based Flow Control used?

- Used only in Data Center Bridging Networks
- PFC and Congestion Spreading:
  - Priority Based – restricts to only relevant traffic classes
  - Where needed, Congestion Notification controls congestion spreading by slowing contributing sources

Title (2.1)

- Amendment to 802.1Q
- Standard for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks – Amendment 9: Priority-based Flow Control.

PAR Scope (5.2)

- This standard specifies protocols, procedures and managed objects that enable flow control per traffic class and augments Congestion Notification in Data Center Bridging networks. Data Center Bridging networks (bridges and end nodes) are characterized by limited bandwidth-delay product and limited hop-count. Traffic class is identified by the VLAN tag encoded priority values. Priority-based flow control is intended to eliminate loss due to congestion. This is achieved by a mechanism similar to the 802.3x PAUSE, but operating on individual priorities. This mechanism, in conjunction with other Data Center Bridging technologies, enables support for higher layer protocols that are highly loss sensitive while not affecting the operation of traditional LAN protocols utilizing other priorities.

PAR Scope (5.3)

- Is the completion of this document contingent upon the completion of another document?

  No. The functions described by this project are intended to operate in conjunction with P802.1Qau and P802.1Qaz; however, no document dependency is expected.

PAR Purpose (5.4)

- Data Center Bridging networks employ higher layer protocols that depend on the delivery of data frames with a much lower probability of frame loss than is typical of IEEE 802 VLAN bridging networks. These protocols were designed for an underlying transport that approaches lossless behavior and therefore do not include appropriate response to frame loss due to congestion (e.g. back-off, slow restart, etc.). This amendment enables multiple data center networks (e.g. IPC, storage, etc.) to be converged onto a single network.

Need for the project (5.5)

- There is significant customer interest and market opportunity for 802 LANs as a converged Layer 2 solution in high-speed short-range networks such as data centers, backplane fabrics, single and multi-chassis interconnects, computing clusters, and storage networks. These environments currently use Layer 2 networks that do not discard packets due to congestion (e.g., Fibre Channel, InfiniBand). This project will bring comparable frame loss characteristics to 802 LANs in Data Center Bridging environments. This in conjunction with the other Data Center Bridging technologies enable converged networks. Use of a converged network will realize operational and equipment cost benefits.

Stakeholders for the Standard (5.6)

- Developers and users of networking for data center environments including networking IC developers, switch and NIC vendors, and users.

Five Criteria

Broad Market Potential

a) Broad sets of applicability
   - Mechanisms to avoid frame loss due to congestion are essential to support the highly loss sensitive higher layer protocols used in Data Center Bridging networks for data storage, clustering, and backplane fabrics. Back-end data storage networks, clustering networks and backplane fabrics with limited number of hops are amenable to a flow control mechanism that operates hop-by-hop.
   - The data traffic to be controlled by the proposed flow control mechanism will be segregated using priority code points encoded in the VLAN tag, ensuring that traffic types that are not amenable to hop-by-hop flow control may co-exist with those that are.

b) Multiple vendors and numerous users
   - Multiple equipment vendors, as well as INCITS T11 Technical Committee, have expressed interest in the proposed project. In addition, multiple vendors have announced product supporting similar technologies in a proprietary way. There is strong and continued user interest in combining separate existing networks into a converged infrastructure, based on international standards, resulting in the realization of operational and equipment cost savings.

c) Balanced costs (LAN versus attached stations)
   - The introduction of this flow control mechanism is not expected to materially alter the balance of costs between end stations and bridges. Significant equipment and operational costs savings are expected as compared to the use of separate networks for traditional LAN connectivity and for loss/latency sensitive applications.

Compatibility

- The proposed standard will be an amendment to 802.1Q, and will interoperate and coexist with all prior revisions and amendments of the 802.1Q standard.
- The data traffic to be controlled by the proposed flow control mechanism will be segregated using priority code points encoded in the VLAN tag, thus ensuring that traffic types already supported by VLAN Bridges are not affected.
- The proposed amendment will contain MIB modules, or additions to existing MIB modules, to provide management operations for any configuration required together with performance monitoring for both end stations and bridges.
- The proposed standard will contain managed objects that will enable its use in conjunction with P802.1Qau and P802.1Qaz.

Distinct Identity

a) Substantially different from other IEEE 802 standards.

IEEE Std 802.1Q is the authoritative specification for priority aware Bridges and their participation in LAN protocols. No other IEEE 802 standard addresses priority-based flow control by bridges.

b) One unique solution per problem (not two solutions to a problem)

IEEE 802.3x defines a link flow control that pauses traffic on the whole link. The need to subject certain classes of traffic to flow control mechanisms, while allowing others to operate without flow control, has not been anticipated by any other IEEE 802 specification. Consequently, this proposal is the only solution to the problem of allowing a coexistence of such traffic types.

c) Easy for the document reader to select the relevant specification.

IEEE Std 802.1Q is the natural reference for priority based handling of traffic flows, which will make the capabilities added by this amendment easy to locate. The amendment will clearly state where its use is appropriate.

Technical Feasibility

a) Demonstrated system feasibility.

Similar techniques are widely deployed in other networking technologies of similar extent, such as Fibre Channel and InfiniBand, as well as in proprietary enhancements to 802.1Q bridging. These deployments have demonstrated that the proposed techniques are preferable to discarding packets during congestion for certain traffic types in networks of limited extent.

b) Proven technology, reasonable testing.

These and similar techniques have been proven in real world deployments of Fibre Channel, InfiniBand, and other networking technologies of similar extent. These techniques have been shown to be reasonably testable.

c) Confidence in reliability.

These and similar techniques have been proven reliable in real-world deployments of Fibre Channel, InfiniBand, and other networking technologies of similar extent.

d) Coexistence of 802 wireless standards specifying devices for unlicensed operation.

Not applicable.

Economic Feasibility

a) Known cost factors, reliable data.

The proposed amendment will retain existing cost characteristics of bridges including simplicity of queue structures and will not require maintenance of additional queues beyond the existing per traffic class (priority) queues for conformance to either its mandatory or optional provisions. In particular per flow queuing will not be required.

b) Reasonable cost for performance.

The proposed technology will reduce overall costs where separate networks are currently required by enabling the use of a converged network. The proposed solution allows a network to avoid frame loss due to congestion without significant throughput reduction.

c) Consideration of installation costs.

Installation costs of VLAN Bridges or end stations are not expected to be significantly affected; any increase in network costs is expected to be more than offset by a reduction in the number of separate networks required to be installed and managed.