Task Group Data Center Bridging
Revision bb-pelissier-pfc-proposal-0408v3.pdf
Author Hugh Barrass (Cisco), et al

Proposal for Priority Based Flow Control

Project Headline

Definition of a new PAUSE function that can halt traffic according to priority tag while allowing traffic at other priority levels to continue.

Modification History

Rev	Originator	Comment
0108	Hugh Barrass	Initial Submitted Version
0308	Joe Pelissier	Updated for March Plenary based on numerous inputs
bb-0308	Joe Pelissier	Updated Annex A to correct maximum frame size and add note about Energy-efficient Ethernet.
0408	Joe Pelissier	Clarified that Time[n] fields are present regardless of the value of the EN Bits.
0408v2	Joe Pelissier	Corrected Typos
0408v3	Joe Pelissier	Added note in 4.6 that flow control of one priority may halt other priorities within the same queue

Table of Contents

1. Authors

The following people, with company affiliations, have contributed to the preparation of this proposal:

Anoop Ghanwani - Brocade
Anjan - Cisco
Bruce Klemin - QLogic
Claudio DeSanti- Cisco
Craig W. Carlson - QLogic
Dan Eisenhauer - IBM
David Peterson - Brocade
Douglas Dreyer - IBM
Ed Bugnion - Nuova
Ed McGlaughlin - QLogic
Glenn - Brocade
Hemal Purohit - QLogic
Hugh Barrass - Cisco
Irv Robinson - Intel
J. R. Rivers - Nuova
Jeffrey Lynch - IBM
Jim Larsen - Intel
Joe Pelissier - Cisco
John Hufferd - Brocade
Manoj Wadekar - Intel
Mike Ko - IBM
Parag Bhide - Emulex
Pat Thaler - Broadcom
Ravi Shenoy - Emulex
Renato Recio - IBM
Robert Snively - Brocade
Roger Hathorn - IBM
Sanjaya Anand - QLogic
Silvano Gai - Nuova
Stuart Berman - Emulex
Suresh Vobbilisetty - Brocade
Taufik Ma - Emulex
Uri Elzur - Broadcom

2. Introduction

This document contains a proposal for a new MAC control frame for the purpose of a priority-based PAUSE.

The function is very closely related to the PAUSE function (802.3x) defined in IEEE 802.3 Clause 31, Annex 31A and Annex 31B.

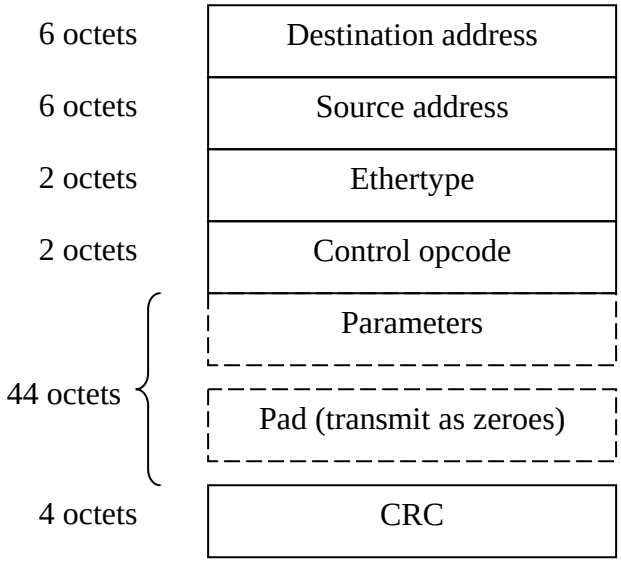
Editors Note: The frame semantics support a MAC control function similar to that defined in presentation "barrass_2_0505" (given to the 802.3ar Task Force during the May 2005 session).

3. Frame format definition

The basic format of a MAC control frame is defined in IEEE 802.3, Clause 31. The opcodes used are defined in Annex 31A and the format of an 802.3x PAUSE frame is defined in Annex 31B. The new PAUSE function is referred to as Priority Based Flow Control (PFC).

3.1 Basic frame format

The MAC control frame format is described in subclause 31.4.1 of IEEE 802.3 with the following diagram:



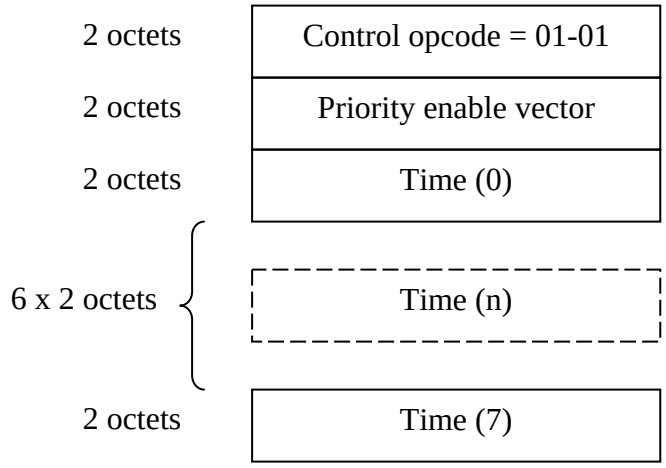
The fields contain the following values:

- Destination address: 01-80-c2-00-00-01.
- Source address: sending station address
- Ethertype: 88-08

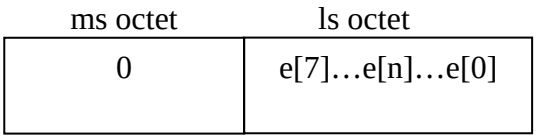
Note that the Destination address and Ethertype values may change as a result of the standardization process. Also note that MAC control frames are never tagged or envelope frames.

3.2 New codeblock definition

The PFC PAUSE frame is defined with a unique opcode, the following semantics are used (note that the Control opcode may change as part of the standardization process):



Priority enable vector definition



.

e[n] = 1 => time (n) valid

e[n] = 0 => time (n) invalid

Time (n) is defined as the pause timer for priority n,
defined in the same manner as in subclause 31B.2 of 802.3

3.3 Interpretation of 802.3x PAUSE frames

After the use of PFC has been negotiated, there shall be no use of 802.3x format PAUSE frames. If an 802.3x format PAUSE frame is received, the receiving MAC should ignore the frame.

4. Priority based PAUSE operation

The priority based PAUSE function is similar to the MAC control PAUSE function defined in IEEE 802.3, Annex 31B. This section describes the additions to support the priority based PAUSE function.

4.1 PAUSE description

The priority based PAUSE function includes a request primitive specifying:

a) The globally assigned 48-bit multicast address 01-80-c2-00-00-01;

b) The PFC PAUSE opcode 01-01;

c) A request_operand indicating the set of priorities addressed and lengths of time for which it wishes to inhibit data frame transmission of the corresponding priorities. (See section 3)

4.2 Parameter semantics

The Time(n) operands are defined in an identical manner to the pause_time operand of IEEE 802.3 Annex 31B.2. Note that all eight Time(n) values are present regardless of the value of the corresponding e[n] bit. For each e[n] bit set to one, the corresponding Time(n) value is valid. For each e[n] bit set to zero, the corresponding Time(n) bit is reserved.

4.3 Transmit operation

The response to the request is similar to the basic PAUSE function, with the appropriate set of format of operands for PFC PAUSE.

The MAC control sublayer does not interfere with the transmission of data frames (initiated by request primitive) as part off the PFC PAUSE function.

4.4 Receive operation

Upon receipt of a valid PFC PAUSE frame, the MAC control sublayer starts 1 to 8 separate counters (depending on the priority enable vector). These timers operate in an identical manner to the single pause timer of IEEE 802.3 Annex 31B.3.3.

4.5 Status indication

The indication primitive contains a vector of “paused / not paused” indications corresponding to the state for all 8 priorities.

4.6 Timing considerations

Editor's note: This section essentially adds 28 pause_quanta to the delay of the PHY to specify an upper bound on the response time for PFC PAUSE. 10GBASE-T was used as the worse case for the budget as follows: 8 pause_quanta for the XGXS, 16 pause_quanta for everything between the Reconciliation Sublayer and the MAC Control inclusive, and four pause_quanta for everything above the MAC Control layer.

On a full duplex link it is possible to receive PFC PAUSE frames asynchronously with respect to the transmission of Data frames. For effective flow control, it is necessary to place an upper bound on the length of time that a device can transmit Data frames after receiving a valid PFC PAUSE frame with a non-zero Time(n) request_operand.

Reception of a PFC PAUSE frame shall not affect the transmission of a frame that has been submitted by the MAC Control sublayer to the underlying MAC (i.e., the TransmitFrame function is synchronous, and is never interrupted).

A station shall not submit a new frame to the PHY layer for transmission from a corresponding priority queue more than 14 336 bit times after the reception of a PFC PAUSE frame from the PHY that contains a nonzero value of Time(n) and a corresponding e[n] bit set to one.

If an implementation utilizes a single queue for multiple priorities, then reception of a PFC PAUSE for one of those priorities may result in head of line blocking for frames of different priorities within the same queue.

Note that in addition to the above delays, system designers should take into account the delay of the PHY and of the link segment when designing devices that implement the PFC PAUSE operation to ensure frames are not lost due to congestion (see Annex A for additional discussion on this topic).

5. Management

Editors Note: This section TBD. Note that management control will allow enabling of PFC PAUSE independently for each priority. A station shall not assert PFC PAUSE for any priority unless that priority has been enabled.

6. Higher Layer function

The PFC PAUSE function is supported by a modification to the scheduling of frames for transmission defined in clause 7 of 802.1D. Data frames from each priority of traffic may be scheduled for transmission if, and only if the indication status for that priority of traffic is "not_paused" for the destination port.

Note that the implementation of fewer priorities of traffic may be supported by combining two or more priorities into a single queue. In this case data frames may only be scheduled for transmission if, and only if the indication status for all of the priorities of traffic sharing the queue is "not_paused" for the destination port. In this case, an implementation may be optimized to contain fewer than 8 pause timers.

A. Calculation of Buffer Requirements *(Informative)*

In order to assure that frames are not lost due to lack of receive buffer space, receivers must ensure that a PFC PAUSE frame is sent while there is remains sufficient receive buffer to absorb the data that may continue to be received while the system is responding to the PFC

PAUSE. The precise calculation of this buffer requirement is highly implementation dependent; however, this Annex attempts to provide an example of how it might be calculated based on a hypothetical implementation.

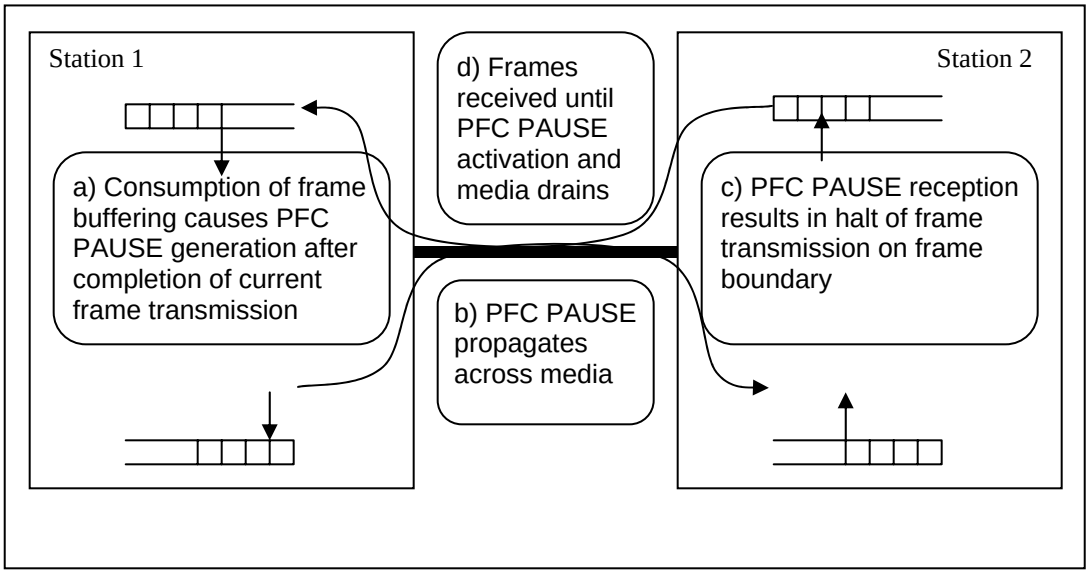


Figure A.1 PFC PAUSE Delays

Figure A.1 illustrates the various delays that must be considered which include:

a) Processing and queuing delay of the PFC PAUSE
b) Propagation delay across the media
c) Response time to the PFC PAUSE frame at the far end
d) Propagation delay across the media on the return path

The processing and queuing delay refers to the time required for a station to detect that it is low on receive buffer, queue the appropriate PFC PAUSE, finish transmitting any frame currently being transmitted, and then transmit the PFC PAUSE. In general, the time to detect the need to transmit the PFC PAUSE and queue it is negligible (again, this is implementation dependent). However, this may occur just as the transmitter is beginning to transmit a maximum length frame. Assuming a maximum length frame of 2000 octets, and a PFC PAUSE frame of 64 octets, the total worst case delay would be 16 512 bit times. This value would need to be increased appropriately if larger frame sizes are supported or if additional processing time is required within the implementation.

Next the propagation delay across the media must be considered. The propagation delay across twisted pair is approximately 0.66 x C where C is the speed of light (3×10^8 m/s). Thus, for 10G 802.3 links, the propagation delay works out to 10^{10}/0.66C bit times / m. Assuming a fiber length of 100m, 5051 bit times results.

The response time is specified in 4.6 of this specification for 10GBASE-T is 14 336 bit times plus the PHY delay of 25 600 bit times (see clause 55.11 of IEEE Std 802.3an™-2006) for a total of 39 936 bit times. In addition, it is possible that a maximum length frame has just begun transmission thus adding 16 000 bit times for a total of 55 936 bit times.

Finally, the return propagation delay (which accounts for data that is already in transit when the PFC PAUSE takes affect), accounts for an additional 5051 bit times.

This results in a grand total of 82 550 bits (approximately 10.1 KB) of buffering required for a 100m 10Gb/s link. As stated previously, more or less buffering may be required to account for implementation specific characteristics such as larger frame sizes, variances in the processing time of generating the PFC PAUSE frame, granularity of buffer allocation and possible sharing of buffers, among others factors. However, in general, the buffer requirements are approximately 2 x (media delay + maximum frame length) + length of PFC PAUSE + the responding end response time.

In addition, designers are cautioned to consider evolving standards such as P802.3az, Energy-efficient Ethernet, which may add additional delays.