Task Group	IEEE 802.1au
Revision	1.3
Author	Hugh Barrass (Cisco)

Definition for new PAUSE function

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	Date	Originator	Comment
1.0	5/30/2007	Hugh Barrass	Initial Submitted Version
1.1	1/10/2008	Hugh Barrass	Updates based on collaboration
1.2	1/23/2008	Joe Pelissier	Updates based on collaboration
1.3	1/30/2008	Joe Pelissier	Updates based on input from the January DCB Interim Meeting

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

Cluadio 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.

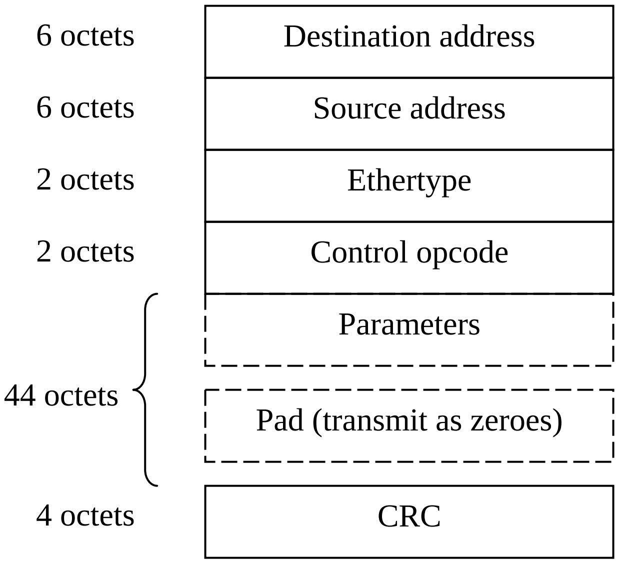
Editors Note: The frame semantics support a MAC control function similar to that defined in presentation "barrass_2_0505" (given to the 802.3ar Task Force during the May 2005 session).

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: TBD from the list of Multicast addresses that do not travel through a switch.
- Source address: sending station address
- Ethertype: TBD

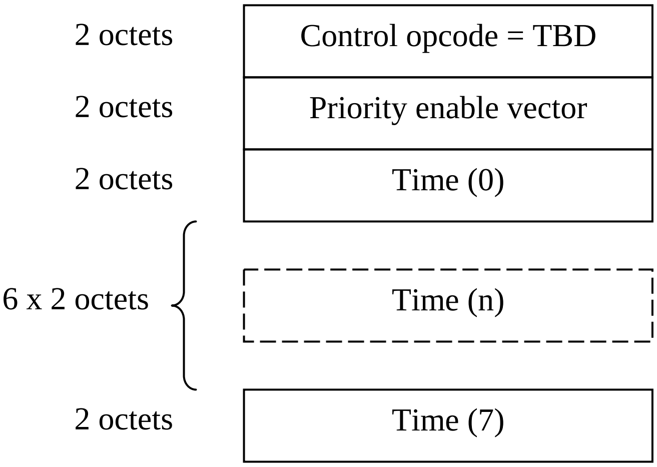
Note that MAC control frames are never tagged or envelope frames.

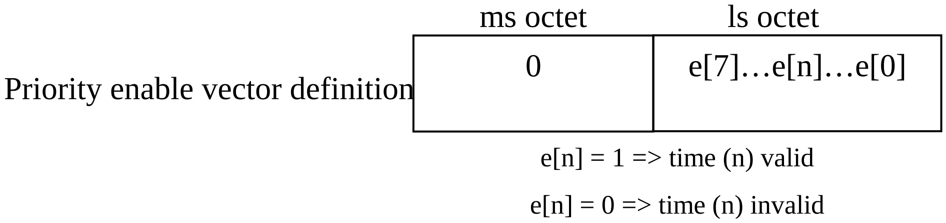
Editor's note: The value used for destination address and Ethertype will be determined *later in the project.*

3.2 New codeblock definition

The PFC PAUSE frame is defined with a unique opcode, the following semantics are used:

Editor's note: The value of the control opcode will be determined later in the project.




ms octet
ls octet
Priority enable vector definition
0
e[7]…e[n]…e[0]
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.

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.

3.4 PAUSE description

The priority based PAUSE function includes a request primitive specifying:

a) The globally assigned 48-bit multicast address TBD;
b) The PFC PAUSE opcode;
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)

3.5 Parameter semantics

The pause_time(n) operands are defined in an identical manner to the pause_time operand of IEEE 802.3 Annex 31B.2.

3.6 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.

3.7 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.

3.8 Status indication

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

3.9 Timing considerations

All of the timing considerations at the physical layer for PFC PAUSE are identical to those defined for basic PAUSE operation. Additional delays to the processing of incoming PFC PAUSE frames are allowed according to the following table. Following receipt of a PFC PAUSE frame the station shall not begin to transmit a new frame after the time prescribed in 802.3, 31B.3.7 plus the additional time shown.

Editor's note: These values will be brought forward into the new standard rather than referencing 802.3.

Link speed	Additional PFC PAUSE processing delay
Speed </= 100Mbps	TBD pause quanta
100Mbps < Speed </= 1Gbps	TBD pause quanta
1Gbps < Speed </= 10Gbps	TBD pause quanta
10Gbps < Speed	TBD pause quanta

.

Editor's note: we currently believe that the pause quanta above will be in the range of 1 to 5. Additional study is required.

Note that for whatever priorities are paused, the traffic for those priorities must stop within the allowed time.

Editor's note: Receiver implementations will need some absorption buffers to prevent packet loss in the time it takes to PAUSE the link partner's transmit function. The size of the absorption buffer is dictated by the PAUSE response time, the PHY latency and the media latency (as described for basic PAUSE). If each priority has an entirely separate buffer then each one needs to be large enough to absorb the slack; however, if the implementation uses perfectly shared buffers then the total shared buffer needs only to be large enough to absorb the slack. The shared buffer may therefore be up to 8 times smaller than the separate buffers. In both cases the slack is defined in terms of time at the egress transmit rate.

4. 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 PAUSE for any priority unless that priority has been enabled.

5. 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.