8802-3/802.3 REVISION REQUEST 1197

DATE: 5th Sep, 2008

NAME: Pat Thaler
COMPANY/AFFILIATION:Broadcom
E-MAIL:pthaler@broadcom.com

REQUESTED REVISION:
STANDARD:802.3
CLAUSE NUMBER:31B.3.7
CLAUSE TITLE:Timing considerations for PAUSE operation

PROPOSED REVISION TEXT:
I don't have exact proposed text. There is more than one option.
A) Increase the round trip delay allowed for 10 Gig ports to 74.

Note that this assumes that WIS is not allowed over LRM. I couldn't find any place where the standard explicitly said the combination was not allowed but it doesn't include a port type for it (i.e. there is no 10GBASE-LRMW). Possibly an explict statement should be added.

If WIS over LRM was allowed, the maximum round trip is 77.

B) State that the maximum delay is 74 PHYs 10 Gig PHY types that support maximum link lengths of 100 m.

C) State that maximum delay is 74 for 10GBASE-T and 10GBASE-KR plus FEC. Maximum delay for all other 10 Gig PHYs is 61 pause_quantum bit times.

RATIONALE FOR REVISION:
There is an inconsistancy in 802.3. The PAUSE reaction delay in 31B.3.7 is not sufficient for all PHYs. A corrigenda should be considered for fixing this because it is fairly common for people allocating the buffers needed for non-drop operation with PAUSE to rely on the value they read in 31B.3.7

31B.3.7 Timing considerations for PAUSE operation allows 60 pause_quantum bit times for the reaction to the reception of PAUSE:

"At operating speeds of 10 Gb/s and above, a station shall not begin to transmit a (new) frame more than sixty pause_quantum bit times after the reception of a valid PAUSE frame that contains a non-zero value of pause_time, as measured at the MDI."

That was the maximum when we completed the first 10 Gig addition to 802.3. That reflected the delays in Table 44-2 for a MAC plus the longest delay PHY at that time: a pair of XGXS sublayers, 10GBASE-R PCS, WIS and PMD.

But since then we have added other PHYs and PMDs and some have a higher delay.

The 10GBASE-T PHY is allowed a delay of 50. Add 16 for the MAC, RS and MAC Control and 8 for XGXS and the delay is 74.

10GBASE-KR without FEC also slightly breaks the 60 pause quanta - its PMD allows 2 pause_quantum bit times of round trip delay. Given how short the medium delay is for this PHY, the extra pause_quantum isn't going to break any upper layer implementation that is built for 100 m links over other PHYs.

FEC adds another 12 pause quanta (6144 bit times) so 10GBASE-KR with FEC can have 73 pause_quantum bit times of delay - almost the same as 10GBASE-T.

Another instance occurs if one runs WIS over LRM. In that case one has a PMD with 18 pause quanta of delay replacing one with 1 so the totoal delay would be 77, but I think this can be ignored as outside the standard. There isn't a WIS PMD type for LRM and LRM specifies only the signaling rate for 10GBASE-R without a WIS.

Option A is the smallest modification of the text and it is consistant with the approach taken for the slower speed PHYs - one delay per speed.

Option B rationale is that the two PHYs with longest delay (assuming that WIS plus LRM is not a valid PHY) are used with short media. Specifying the extra delay as only applying to PHYs that support physical media up to 100 m allows a PHY independent upper layer implementation can use the same buffering to cover slower PHYs with shorter link delays and faster PHYs with longer link delays. If we increase the delay for all PHYs an implementation independent upper layer design might need to assume that it can have that greater PHY delay plus long link delays.

Option C is more specific about which two are slower so those upper layer implementors not supporting those lower PHYs would know they could use the shorter number. But the list would have to be maintained if any new PHYs with delay over 61 are added. The change from 60 to 61 is to support the 10GBASE-KR delay without FEC.

IMPACT ON EXISTING NETWORKS:
There may be some existing networks that use PAUSE with too small a buffer because they relied on the 60 pause_quant bit times. But since the higher delay PHYs are already in the standard, such implementations already may not have enough buffering for no drop. This doesn’t change that situation. The change will help those doing future configuration to choose the right buffering levels. In many cases, the buffer thresholds are configurable so existing networks may be able to change configuration to allow for the additional delay.

Please attach supporting material, if any	
Submit to:- David Law, Chair IEEE 802.3	
E-Mail: David_Law@ieee.org	
and copy:- Wael William Diab, Vice-Chair IEEE 802.3	
E-Mail: wdiab@broadcom.com	

For official 802.3 use	
REV REQ NUMBER: 1197	
DATE RECEIVED: 5th Sep, 2008	
EDITORIAL/TECHNICAL	
ACCEPTED/DENIED	
BALLOT REQ'D YES/NO	
COMMENTS: XX-Xxx-XX Ver: D1.0 Status: R	

For information about this Revision Request see -

http://www.ieee802.org/3/maint/requests/revision_history.html#REQ1197