802.3az D3.1
MAC transmit deferral during fast retrain

Matt Brown

IEEE 802.3 Plenary July 2010

Supported by

- Ilango Ganga
- David Chalupsky
- Hugh Barrass
- Brad Booth

Summary

- Several open comments point out the need for a mechanism to defer transmission of packets while a 10GBASE-T PHY is in fast retrain, without simultaneously triggering adverse behavior in protection and connection management.
- Related comments include
  - Draft 3.1: 83, 96, 97
  - Draft 3.0: 163,164, 170, and 359

Draft 3.1

- A 10GBASE-T PHY will send either Local Fault (LF) or IDLE to the RS when the PHY is in fast retrain.
  - selectable from MDIO 1.147.1.
- Local Fault
  - indicates there is a problem to the RS/MAC, but does not differentiate between a temporary fault (e.g., fast retrain) and a persistent fault, (e.g., normal retrain, pulled cable, etc.).
  - local fault might be usable to defer packets, but then it would not be immediately useful for failover and other consequent actions
- IDLE
  - ensures the RS/MAC does not react as though the link is down, but neither allows the RS/MAC to defer traffic, thus packets are unavoidably discarded during fast retrain

Proposal

- This presentation proposes:
  - a new <Link Unavailable> signal
  - provision for the PHY to optionally send the <Link Unavailable> signal to the RS during fast retrain
  - provision for the RS to detect the <Link Unavailable> signal
  - provision for the RS to set the CARRIER_DETECT parameter and TXD/TXC output in response to the <Link Unavailable> signal

Implementation

- Create a new optional signal to indicate <Link Unavailable> .
- Clause 45 “Fast Retrain Signal Type” register field will select signal from options IDLE, Local Fault, or <Link Unavailable>.
- Generate the <Link Unavailable> signal, when selected, by the 10GBASE-T 64B/65B RX state machine while in fast retrain.
- The XS TX and DTE RX PCS state machines will transparently pass the <Link Unavailable> signal.
- Rate adaptation may drop <Link Unavailable> signal as required.
- The RS RX detects the <Link Unavailable> signal and sets the CARRIER_DETECT parameter to CARRIER_ON in response.
- An Annex 4A MAC may use the CARRIER_DETECT state to defer transmission of packets.
- RS detection of and response to the <Link Unavailable> signal is optional.

<Link Unavailable> signal

- <Link Unavailable> may be signaled by use of a new sequence ordered set.
- A sequence ordered set has the advantages...
  - transparent to both standard XGXS and proprietary serial XGMII links.
  - compatible with rate adaptation as defined
    - 48.2.4.2.3 and 49.2.4.10 permit deletion of sequence ordered sets
  - No changes to clauses 48 and 49.
- Define a new signal ordered set in 46.3.4:
  - Lane 0/1/2/3 = Sequence/0x00/0x00/0x03/
  - Description = <Link Unavailable>
- The <Link Unavailable> signal may be detected by the RS link fault state machine (46.3.4.3) by adding a 3rd fault <Link Unavailable> to the list. See later slides.

10GBASE-T receive 64B/65B state diagram – Part I

- Make the following changes to Figure 55-16
- Delete state FR_RX_INIT
- Change entry RX_INIT entry criteria to

  *pcs_reset + hi_lfer + !block_lock + (fr_active * !pcs_data_mode)*
- Change first action in RX_INIT to if-else statement

  if !fr_active
  rx_raw = LBLOCK_R
  else
  rx_raw = fr_sigtype
  end

10GBASE-T receive 64B/65B state diagram – Part II

- Re-define lpi_fr_sigtype as follows:

  fr_sigtype [renamed from lpi_fr_sigtype]
  If fast re-train is supported, this variable is set based on the value in 1.147.2:1 as follows:
  00b IBLOCK_R
  01b LBLOCK_R
  10b UBLOCK_R (see next slide)
  11b Reserved

10GBASE-T receive 64B/65B state diagram – Part III

- Update 45.2.1 with modifed register field definition
  - Note that 1.147.1 was supposed to be defined in Draft 3.1.
    - See comments 3.1/23 and 3.1/99.
    - The changes below assume 1.147.1 was already defined.
  - Modify entry in Table 45-53a for 1.147.1 as follows:
    - Bits = “1.147.2:1”
    - Name = “Fast retrain signal type”
    - Description = “0 = Local Fault, 1 = Idle, 2 = <Link Unavailable>, 3 = Reserved”
    - R/W = “R/W”.
  - Modify 45.2.1.76a.2 Fast retrain signal type (1.147.2:1)
    - For PHYs that support fast retrain, this bit maps to lpi_fr_sigtype as defined in 55.4.5.1. When fast retrain signal type is set to one, the PMA sends IDLE characters on the receive path during fast retrain. When Fast retrain signal type is set to zero, the PMA sends local fault on the receive path during fast retrain. When fast retrain signal type is set to 3, the PMA sends <link unavailable> on the receive path during fast retrain.

Define UBLOCK_R XGMII block type

- Define a new receive block type constant in 55.3.5.2.1

  UBLOCK_R<71:0>

  72-bit vector to be sent to the XGMII containing two <Link Unavailable> ordered_sets. The <Link Unavailable> ordered set is defined in 46.3.4.

- It is not necessary to define a corresponding block for the transmit direction.

RS detection of <Link Unavailable> (part I)

- The <*Link Unavailable*> signal may be detected by the RS receive link fault monitor.
- In Section 46.3.4 add the following paragraph...

  For operation with links that may be temporarily unavailable, optional detection of a third fault condition, <Link Unavailable> , is provided. The <Link Unavailable> fault is indicated by the PHY receive function by continuously sending the <Link Unavailable> ordered_set. The <Link Unavailable> ordered set is specified in Table 46-5.

RS detection of <Link Unavailable> (part II)

- In Section 46.3.4.2 the following variables will be redefined:
- fault sequence
  - change end of definition to...
  - “indicating Local Fault, Remote Fault, or optionally <Link Unavailable> .
- link fault
  - Add new status...
  - “<Link Unavailable> signaled by the PHY (optional)”
- seq_type
  - Add new sequence value...
  - “<Link Unavailable> ; 0x00 in lane 1, 0x00 in lane 2; 0x03 in lane 3 (optional)”

RS detection of <Link Unavailable> (part III)

- In Section 46.3.4.3...
- No change to the state machine in Figure 46-9 is required.
- End of the second paragraph after Figure 46-9 is modified...
  - “containing a Remote Fault, Local Fault, or optional <Link Unavailable> sequence ordered_set.”
- In the TXC/TXD output definition, change the condition of the third criteria to:
  - “c) link_fault = {Remote Fault or <Link Unavailable> (optional) }”

Control of CARRIER_STATUS parameter (part I)

- In 46.3a.2.1 define...
  - LPI_CARRIER_STATUS: The LPI_CARRIER_STATUS variable indicates how the CARRIER_STATUS parameter is controlled by the LPI_REQUEST parameter. The LPI_CARRIER_STATUS is either TRUE or FALSE as determined by the Transmit LPI state diagram in Figure 46-10a.
- In Figure 46-10a...
  - Change “CARRIER_STATUS = ON” to “LPI_CARRIER_STATUS = TRUE”
  - Change “CARRIER_STATUS = OFF” to “LPI_CARRIER_STATUS = FALSE”

Control of CARRIER_STATUS parameter (part I)

- In 46.1.7.3 (802.3az/3.1).
  - Change the second paragraph as follows:
  - “For PHYs that support EEE, CARRIER_STATUS is set in response to LPI_INDICATION. For PHYs that support <Link Unavailable> fault detection, CARRIER_STATUS is set in response to link_fault. CARRIER_STATUS is set to CARRIER_ON if LPI_CARRIER_STATUS is TRUE or link_fault is <Link Unavailable> . CARRIER_STATUS is otherwise set to CARRIER_OFF.”

Other considerations

- A MAC/RS that does not support the <link unavailable> detection and deferral, e.g., a legacy device, will react the same to both the <link unavailable> signal and IDLE.
  - Consider simply replacing IDLE with <link unavailable> instead of adding <link unavailable> as third option.
  - The MDIO field remains 1.147.1 (instead of expanding it to 1.147.2:1) and IDLE is replaced with <link unavailable> in the state machines and variables.
  - Reduces the choice of signals to two clear alternatives...
    - LF and LU, instead of LF, IDLE, and LU.

Supplemental material

Related comments

Comment 3.1/83 (new), 3.1/96 (new)

Cl **55** *SC* **55.4.2.5.14** *P* **216** *L* **49** # 83

Brown, Matthew — Applied Micro (AMCC)

Comment Type **TR** *Comment Status* **D**

The is a pile-on comment for Draft 3.0 comment #359. The response to comment #359 addresses incorrectly detecting a failed link by optionally replacing the local fault signal with the idle signal during fast retrain. The reponse did not address loss of data during a fast retrain. To prevent loss of data, a mechansm is required which informs the MAC to defer transmission; while not indicating a link failure, avoiding adverse effects on MAC clients.

SuggestedRemedy

Provide a mechanism to signal from the PHY to the RS a temporary interruption during fast retrain. Provide a mechanism in the RS to cause the MAC to defer transmission of packets while fast retrain is active, particular for a MAC which is connected to a PHY through a XAUI interface. To accomplish this create a new character, similar to /LI/, call tentatively /CRS/ (carrier sense). Send /CRS/ continuous to the RX XGMII while fast retrain is active. In the RS, while receiver /CRS/ from the RX XGMII set PLS_CARRIER.indication(CARRIER_STATUS) to CARRIER_ON.

Proposed Response *Response Status* **W**

PROPOSED REJECT.

For discussion by the task force.

See also #100.

This is out of scope for clause 55.

Cl **46** *SC* **46.1.7.3** *P* **140** *L* **37** # 96

Ganga, Ilango — Intel Corporation

Comment Type **TR** *Comment Status* **D**

The spirit of the EEE objectives is not to drop or corrupt frames; however fast retrain mechanism, as defined, has the potential to drop frames. Some of the upper layer protocols expect no packet drop characterestics and certain reliability at link level. Fast retrain condition may cause frame loss up to several ms. So implement a mechanism that has ability to defer frame transmission during fast retrain.

SuggestedRemedy

Set the PLS_CARRIER.inidication primitive when the PMA indicates fr_active (PMA_FR_ACTIVE.indication) to defer transmission during fast retrain. This will ensure no packet drop during fast retrain.

Proposed Response *Response Status* **W**

PROPOSED REJECT.

The subject of deferral was discussed during the resolution of comment #164 and #361 on draft 3.0. The decision was taken to use Local Fault as the sole means to signal from the PHY to the RS that fast retrain is in progress. The proposed remedy of this comment would add an additional signal to the XGMII to convey the state of the proposed new primitive.

See also comment #100, #97, #83

Comment 3.1/97 (new),3.0/163 (unresolved)

Cl **55** *SC* **55.3.2.2.9** *P* **195** *L* **10** # 97

Ganga, Ilango — Intel Corporation

Comment Type **TR** *Comment Status* **D**

As per D3.1, either IDLE or Local Fault is generated during fast retrain. Currently local fault may be used to trigger link failure condition to the higher layers. At a system level such link failure conditions may be used to initiate link failover mechanisms for high availability. Asserting local fault does not unambiguously indicate if the local fault is due to link failure or fast retrain. Any timeout mechanisms to delay signaling link failure to higher layers may delay the highavailability/failover features to take effect. So it is best to define a separate control code to indicate fr_active (PMA_FR_ACTIVE.indication) to the RS sublayer. This could be used to signal a fast retrain condition.

SuggestedRemedy

1. Define a seprate control code to indicate fast retrain condition to the higher layers (RS sublayer). Providing fr active signal allows systems flexibility to implement failover/lossless characteristics. 2. For the PHYs that support fast retrain, specify an option to assert PLS_CARRIER.indication during fast retrain active that allows tx deferral.

Proposed Response *Response Status* **W**

PROPOSED REJECT.

This was discussed at the previous meeting and the taskforce could not reach agreement on making this change.

For further discussion by the taskforce.

Cl **46** *SC* **46.1.7** *P* **135** *L* **24** # 163

Brown, Matthew — Applied Micro (AMCC)

Comment Type **GR** *Comment Status* **R** *frdata*

Receipt of local fault also causes override of transmitted signal. Receipt of local or remote fault should also result in asserting carrier_sense.

SuggestedRemedy

Append to last sentence of paragraph "or link is in a fault state."

Response *Response Status* **U**

REJECT.

Carrier deferral for loss of data during fast retrain is not being implemented - see response to comment #164.

Comments 3.0/359 (unresolved), 3.0/369 (closed)

Cl **55** *SC* **55.4.2.5.15** *P* **209** *L* **42** # 359

Ganga, Ilango — Intel Corporation

Comment Type **TR** *Comment Status* **A** *frdata*

The effect Clause 55 Fast Retrain on the Reconciliation Sublayer & MAC is unclear. Fast Retrain mechanism should be specified in a such a way that it does not indicate link down/link failure to the higher layers and also does not cause any data loss (that may cause packet drops). When the PHY Control State Diagram exits the PCS Data state to enter PMA_INIT_FR, it is unclear what action the PHY will take with respect to the XGMII path to the MAC. If PHY sends Local Fault up to the XGMII (i.e., if block_lock is lost, forcing the Local Fault ordered set) then the MAC will see this as a loss of link and this will be very disruptive to the System. The Fast Retrain mechanism is 'fast' enough to allow for recovery without sending alarms to higher functions. However, if the fast retrain is not signaled to the MAC, then the MAC may continue to send data that will be lost. It is also undesirable to drop 30msec of data without notification.

SuggestedRemedy

Fast Retrain mechanism should be specified in such a way that it does not cause a Local Fault (or signal link down to higher layers). The mechanism should also prevent the MAC from transmitting data during the retrain period to avoid any data loss or packet drops.

Response *Response Status* **W**

ACCEPT IN PRINCIPLE.

See motion in diab_01_0510.pdf

Also make the following changes to Clause 45:

Define a new register bit:

1.147.1 : Fast retrain signal type : 1 = send IDLE during fast retrain, 0 = send local fault during fast retrain

Insert 45.2.1.76a.2 Fast retrain signal type (1.147.1)

For PHYs that support fast retrain, this bit maps to lpi_fr_sigtype as defined in 55.4.5.1.

When Fast retrain signal type is set to one, the PMA sends IDLE characters on the receive path during fast retrain. When Fast retrain signal type is set to zero, the PMA sends local fault on the receive path during fast retrain.

See parnaby_03_0510.pdf for the changes to clause 55

Also see response to comment #164 for data loss or packet drops

Cl **55** *SC* **55** *P* **209** *L* # 369

Bennett, Michael — Lawrence Berkeley Na

Comment Type **T** *Comment Status* **A** *frdata*

Submitted on behalf of Paul Langner Paul.Langner@aquantia.com Currently the IEEE fast-retrain mechanism being proposed does not implement a mechanism to inform the MAC that the link is temporarily unavailable. As a result, the MAC will continue to send data during a fast-retrain (for up to 30 ms). This data will all be lost. In order to prevent this from occurring, a mechanism is needed to inform the MAC that the link is temporarily unavailable, so that the data will not be lost, and can be buffered until the link is available.

SuggestedRemedy

Create a control code (similar to Local Fault) that indicates that the link is temporarily unavailable, and this control code would be sent continuously to the MAC until the retrain is completed.

Response *Response Status* **C**

ACCEPT IN PRINCIPLE.

See responses to comment #359 and #164