Proposal for Message Timestamp Point

Richard Tse, Microchip Technology
David Law, Hewlett Packard Enterprise

IEEE 802.3cx Teleconference, Nov 17, 2020

Supporters:

- Andras de Koos, Microchip Technology
- Clark Carty, Cisco Systems
- David Law, Hewlett Packard Enterprise
- Denny Wong, Xilinx
- Dino Pozzebon, Microchip Technology
- Marek Hajduczenia, Charter Communications
- Mark Bordogna, Intel
- Richard Tse, Microchip Technology
- Sriram Natarajan, Cisco Systems
- Steve Carlson, High Speed Design Inc.
- Ulf Parkholm X, Ericsson

Message Timestamp Point Discrepancy

Subclause 90.7 of IEEE 802.3 states

- “The transmit path data delay is measured from the input of the beginning of the SFD at the xMII to its presentation by the PHY to the MDI. The receive path data delay is measured from the input of the beginning of the SFD at the MDI to its presentation by the PHY to the xMII.”

however…

Subclause 7.3.4.1 of IEEE 1588-2008 and subclause 11.3.9 of IEEE 802.1AS define the message timestamp point as follow:

- “the message timestamp point for an event message shall be the beginning of the first symbol after the Start of Frame (SOF) delimiter”
- “the message timestamp point for a PTP event message shall be the beginning of the first symbol following the start of frame delimiter”

Previous Discussions

- tse_itsa_02_0120.pdf
  - Introduced the issue
  - Shows the text in IEEE 802.3 that conflicts with the specifications defined in IEEE 1588 and IEEE 802.1AS
  - Shows the principals behind the generation of time error if different message timestamp points are used
- parkholm_itsa_01_0120.pdf
  - Discussed timestamp error due to different message timestamp points for 25GE with RS-FEC when using 32-bit xMII per clause 106 of IEEE 802.3
- tse_3cx_01_0420.pdf
  - Proposed text for changing the message timestamp point to match that of IEEE 1588 and IEEE 802.1AS
  - Suggested adding informative text describing the effect of this change for implementations that do not use the IEEE 1588 and IEEE 802.1AS message timestamp points and methods that could be used to compensate for the change

Making a Decision

Considerations for making a decision:

- Because of the conflicting specifications on the message timestamp point, any existing implementation could have chosen to be compliant to either one
- Regardless of which one we choose, there will be existing implementations that might have their performance degraded (see later slide on “Annex XX”)
- We should allow existing implementations to remain compliant to IEEE 802.3
- To enable high accuracy, we should recommend just one message timestamp point for IEEE 802.3cx

Proposal:

- For IEEE 802.3cx, recommend using the message timestamp point that matches that of the applications (IEEE 1588, IEEE 802.1AS) this function was meant to service

Proposal (1/2)

- Modify clause 90.7 as follows:
- Insertions are highlighted in blue and deletions are highlighted in ~~red~~.

The TimeSync capability requires measurement of data delay in the transmit and receive paths, as shown in Figure 90–3. The message timestamp point shall be either the beginning of the start of frame delimiter (SFD) or the beginning of the first symbol after the SFD.

Note: It is recommended that the beginning of the first symbol after the SFD is used as the message timestamp point since this matches the measurement points specified in IEEE Std 802.1AS IEEE Standard for Local and Metropolitan Area Networks--Timing and Synchronization for Time-Sensitive Applications and IEEE Std 1588 IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems. The use of the beginning of the SFD as the message timestamp point may impact the accuracy that can be achieved by a time synchronization protocol using this TSSI. Further information can be found in Annex XX.

Proposal (2/2)

- **Modify text in clause 90.7 as follows:**

  The transmit path data delay is measured from the ~~beginning of the SFD~~message timestamp point at the xMII input to the ~~beginning of the SFD~~message timestamp point at the MDI output. The receive path data delay is measured from the ~~beginning of the SFD~~message timestamp point at the MDI input to the ~~beginning of the SFD~~message timestamp point at the xMII output.

  For a PHY that includes an FEC function, the transmit and receive path data delays may show significant variation depending upon the position of the ~~SFD~~message timestamp point within the FEC block. However, since the variation due to this effect in the transmit path is expected to be compensated by the inverse variation in the receive path, it is recommended that the transmit and receive path data delays be reported as if the ~~SFD~~message timestamp point is at the start of the FEC block.

  The receiver of a multi-lane PHY is expected to include a buffer to compensate for skew between the lanes. This buffer selectively delays each lane such that the lanes are aligned at the buffer output. The earliest arriving lane experiences the most delay through the buffer and the latest arriving lane experiences the least delay through the buffer. The receive path data delay for a multi-lane PHY is reported as if the ~~beginning of the SFD~~message timestamp point arrived at the MDI input on the lane with the smallest buffer delay.

For Annex XX (Informative, Compensation Options)

- What can be done for implementations that use a message timestamp point at “the beginning of SFD” to adapt them to using a message timestamp point at “the beginning of the symbol after the SFD”?
  - Single-lane interfaces (without FEC)
    - Add time offset of one byte time to the timestamp on a transmit interface
    - Add time offset of one byte time to the timestamp on a receive interface
    - See subsequent slides for additional insights on multiple lanes and FEC
  - Alignment markers (also codeword markers)
    - Timestamp errors resulting from alignment markers will occur infrequently (only when the SFD byte and the DA byte of a frame that is timestamped are separated by alignment markers) and would typically be filtered out by PTP ToD recovery algorithms
      - AM event occurs once every 16384 blocks
      - Probability of timestamp event is once per Ethernet frame + IFG (i.e., once every 10 blocks for min sized frame and small IFG, less for larger frames or larger IFGs)
      - Thus, probability of AM corrupting the timestamp of *any* frame is no greater than:

        $P \leq (1/16384) * (1/10) \cong 6E\text{-}6$
  - 25GE with RS FEC (from parkholm_itsa_01_0120.pdf)
    - If SFD and the first DA byte are located in different FEC blocks, their timestamps will differ by the time of one FEC block (5280 UI)
    - A receiver talking to a transmitter using the other message timestamp point will regularly see timestamp jumps equal to the time of one FEC block. This behavior could be detected by the receiver and compensated for.
    - See subsequent slide for additional insights on this topic…

Additional Insights (1/2)

- Regarding 25GE with RS FEC Issue:
  - Denny Wong noted that a difference in message timestamp point (between SFD and symbol after SFD) should not cause a timestamp error of one FEC block (5280 UI), as indicated in slide 5 of parkholm_itsa_01_0120.pdf
    - Clause 90.7 states only that “the transmit and receive path data delays be reported as if the ~~SFD~~message timestamp point is at the start of the FEC block”, it does not say the message timestamp point is moved to the start of the FEC block
    - Given the FEC path data delay is a constant value (and equal to that of the symbol at the start of the FEC block) and the timestamps for the SFD byte and for the symbol after SFD byte at the xMII are only separated by one byte time, the timestamp difference between the two message timestamp points would only be one byte time

Additional Insights (2/2)

- Regarding Multi-PCS lane distribution and the message timestamp point
  - Andras de Koos noted that using Option C + Method 2 for multi-PCS lane distribution delay minimizes the timestamp error range between the two message timestamp point options
  - Option C + Method 2 solution
    - Given:
      - "transmit and receive path data delays be reported as if the ~~SFD~~message timestamp point is at the start of the FEC block and multilane distribution sequence", the multi-PCS lane path data delay is a constant value
      - The timestamps for the SFD byte and for the symbol after SFD byte at the xMII are separated by one byte time
    - The timestamp difference between the two message timestamp points would only be one byte time (0.08ns for 100GE)
  - Option B + Method 1 solution
    - Each set of N outputs on a Tx multi-PCS lane function has the same timestamp (where N = # of lanes)
    - Timestamp difference between successive sets of Tx multi-PCS lanes is equal to N blocks
    - If SFD byte is in one set of multi-PCS lanes and symbol after SFD is in next set of multi-PCS lanes, the timestamp difference between the two will be N blocks (12.8ns for 100GE's 20 lanes)

Thank you!

Any questions or comments?