Thoughts following ITU-T SG15 Q13 October 2017 Liaison letter

(Titled “Impact on timing performance due to Ethernet PHY”)

Adee Ran

Intel

Purpose of this presentation

- Address the liaison letter
- Summarize information about possible sources of delay variability
- Identify gaps
- Recommend further activity

Liaison letter content summary

- Concerns:
  - Asymmetric delay between the two transmission directions
  - Periodical delay due to RS-FEC encoding (“every RS-FEC frame”)
  - Periodical delay due to codeword marker insertion (“every 1024 RS-FEC frame”)
  - Rate compensation
  - Alignment markers insertion/removal
- Goals/ideas
  - “it is important that phase error introduced by the PHY components is well controlled, possibly, according to strict standard rules, to within a few ns.”
  - The location of timestamping in Tx and Rx may affect the timing noise

    Proposal: “With reference to Figure 108-1 in ref.2, performing the timestamping in transmission above the RS-FEC layer in both transmit and reception, may remove some of this noise. Taking into account the delay added by the RS-FEC is another approach that could be considered (e.g.,. timestamping is done below the RS-FEC layer in both transmission and reception).”
  - “it is considered important that a consistent approach is followed among vendors in order to control the timestamping noise”
- Q13 asks
  - “advice concerning the specific aspects”
  - “updates concerning actions that may have been taken in order to guarantee that delays added by the Ethernet PHY components are properly controlled”
  - “consider if any action would be required to be initiated”

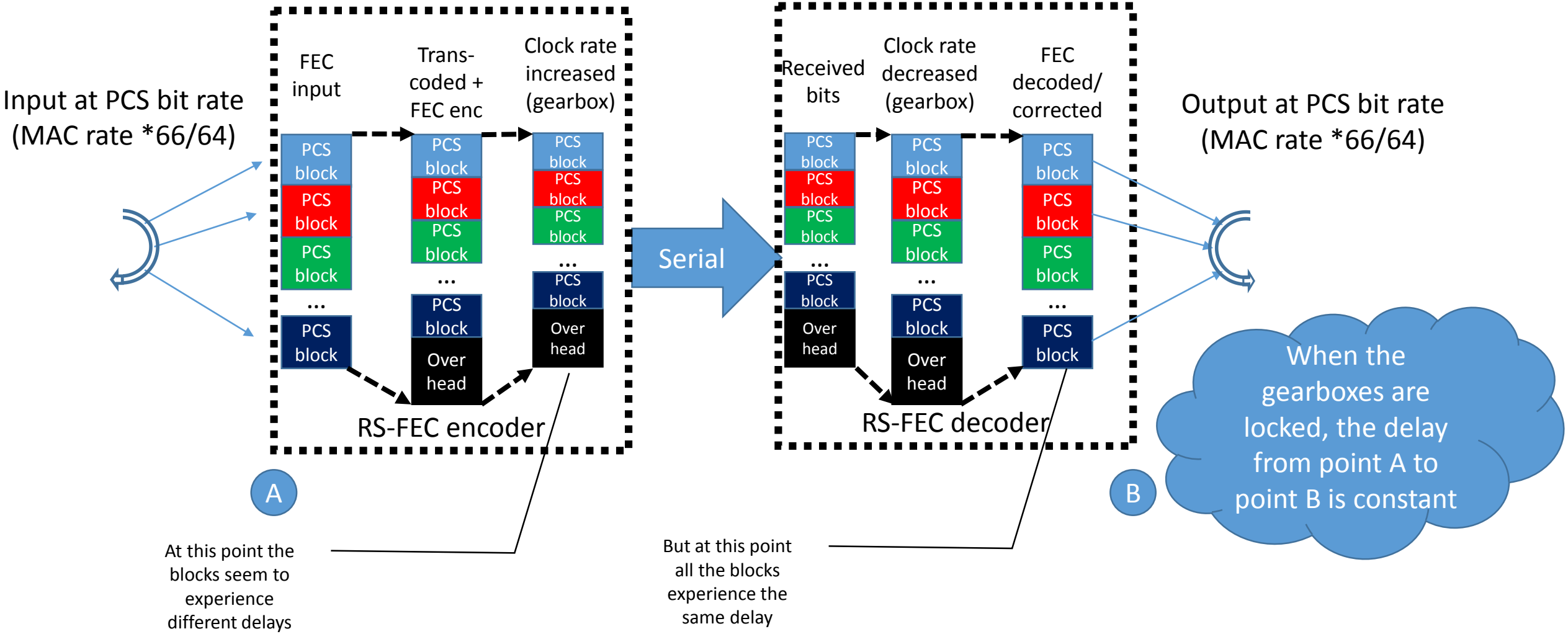
Addressing concern 1 – asymmetric delay

- All practical FEC schemes have longer delay on the receive path than on the transmit path.
  - For example, clause 108 RS-FEC, the FEC encoding algorithm is simple and can be implemented in a small number of data-path clocks (up to a few ns). At the receiver, a full FEC codeword must be stored, and then corrected, and then forwarded to the next sublayer – this creates a delay of at least a full codeword (100's of ns).
  - The Tx and Rx encoding delays may be implementation-specific, but are known for any specific design.
- In addition to the encoding and decoding delays, some implementations may use gearboxes for conversion between different clock rates. This may add delays that in both TX and RX.
  - These delays may have some randomness, but are fixed once the FEC is locked.
  - Gearbox delays may be measured (in implementation-specific manner).
- **Known asymmetry can be accounted for such that time synchronization accuracy is not affected. Reporting the delays in each direction is what clause 90 is about.**

Addressing concern 2 – RS-FEC “parity” insertion

- Clause 108 FEC, like most practical FEC schemes, is systematic - a parity block is inserted on the transmit path after a “payload” of data blocks. This creates occasional gaps between data blocks.
  - Room for the parity block is created by transcoding (in other FEC clauses this involves increasing the signaling rate and/or special modulation).
  - The parity block is removed from the data path during decoding at the Rx.
  - These operations do not insert or delete any PCS data (e.g. idle characters) and therefore are fully synchronous.
- FEC encoding may seem to cause variable delay…
  - A packet fed into the FEC encoder may be “advanced in time” due to the compression effect of the transcoding; this advancement depends on the location of the packet within the codeword.
  - However, the apparent Tx advancement is fully compensated by the de-compression of the reverse transcoding in the Rx.
  - The FEC throughput is fixed, so the **Tx and Rx delays sum to a fixed value.**
- **To avoid an unreal difference between min/max values, the reported Tx and Rx delays should ignore this compression effect.**
- Preferably, the FEC delays should be defined as
  - In the Tx direction: delay of an SFD from the encoder input to the encoder output, assuming the SFD is aligned with the start of the FEC payload
  - In the Rx direction: delay of an SFD from the decoder input to the decoder output, assuming the SFD is aligned with the start of the FEC codeword.
- We should consider adding these definitions to the standard.
- Similar concerns may exist in several other clauses that implement FEC functionality. The recommendation should be generic.

FEC encode + decode create a fixed delay



Addressing concern 3 – Periodical delay due to codeword marker insertion

- In clause 108, codeword marker is inserted once every 1024 RS-FEC codewords
  - The term “codeword marker” is unique to this clause, but the concept is similar to the alignment markers used in multi-lane PHYs.
- According to the standard text, the RS-FEC sublayer includes the following functionality:
  - At the TX, some idle characters (which always appear between packets) at the RS-FEC input are removed as necessary to create room for a 257-UI long CWM.
  - At the Rx, the CWM is removed and replaced with 32 idle characters.
  - The exact methods of removing idles in the Tx, and inserting idles at the RX, are not specified.
- Implementation using the layer separation and service interfaces described in the standard may require elastic buffers
  - Unlike gearboxes, elastic buffers are asynchronous and hold a number of bits that varies over time, to account for markers. This creates variable delays between packets.
  - If the Tx and the Rx are not using exactly the same method for insertion and deletion, the gaps between packets at both ends would not be the same.
  - Delay variations between data blocks (packet jitter) may be up to 32*8=256 bit times = 10.24 ns.
- **There are other ways to implement the same behavior without introducing timing uncertainty.**
  - The standard does not list all possible implementations, and optimized implementations exist.
- See also concern 5.

Addressing concern 4 – Rate compensation

- The Tx and the Rx may have different clock frequencies at the MAC
  - Each up to ±100 PPM from the nominal frequency – the maximum difference is ±200 PPM
- To compensate, the Rx can change the IPG length by inserting or removing idle characters
  - Architecturally this is a function of the **Reconciliation Sublayer (RS)**
  - Time of arrival at the xMII is not affected by this function
- Rate compensation seems not to be a concern for time synchronization

Addressing concern 5 – Alignment markers

- Alignment markers were first introduced in 802.3ba, which predates 802.3bf
- The concern is essentially the same as codeword markers (concern 3)
- The alignment markers are inserted as a group, once every 2^14 blocks on each PCS lane
  - In 40G there are 4 lanes, this creates a block of 4*8=32 characters=256 bit times=6.4 ns
  - In 100G there are 20 lanes, this creates a block of 20*8=160 characters=1280 bit times=12.8 ns
- A PHY implemented with the sublayer separation and service interfaces described in the standard may have variable delay between the xMII and PCS output.
- As in the case of codeword markers, this variation can be avoided in implementation-specific ways.

Clause 90 TSSI

- TimeSync Service Interface (TSSI) is defined at the Reconciliation Sublayer (RS) or based on time of detection at the xMII.
- The data delays are defined in 90.7 as
  - Tx delay: “from the input of the beginning of the SFD at the xMII to its presentation by the PHY to the MDI”
  - Rx delay: “from the input of the beginning of the SFD at the MDI to its presentation by the PHY to the xMII”
- Timing indication based strictly on SFD detection at the xMII may be noisy in both directions.
  - This is inherent from the architectural position and service interfaces of TSSI.
- Changing the reference location to another sublayer (such as the RS-FEC service interface) would have severe implications…
  - It would affect existing compliant implementations
  - It would apply only to some PHYs (not a generic solution)
  - Is overly prescriptive and may not be suitable for all implementations.
- Implementations can improve accuracy using methods beyond the scope of the standard.
- Note that there are delay registers for the PCS, and for the PMA/PMD, but not for the RS-FEC when it is a separate sublayer.
  - This may create confusion when the RS-FEC is not co-located with either the PMD or the PCS.
  - Recommendation: use either PCS or PMA/PMD registers according to implementation convenience (e.g. placement of the RS-FEC function). Add clarification in clause 90.

Summary of recommendations

- To create a common language for reporting FEC delay for clause 90 applications, TimeSync PHY transmit data delay and PHY receive data delay min/max registers:
  - Should include any possible variability caused by the AM or CWM insertion/removal that impacts the TSSI.
  - Should not include the delay variability caused by the FEC encoding and decoding functions, as this variability cancels out on the combined Tx+Rx.
  - FEC delays should be reported and be defined as: packet delay through the FEC block in the respective direction, assuming the packet starts at the first block of the FEC codeword.
- RS-FEC delays should be included either the PCS delay registers or the PMA/PMD delay registers, but not both, according to implementation convenience.
- Study applicability of the above to other PHYs that include FEC functionality.

Review of concerns, ideas, and requests

- Goals/ideas
  - “it is important that phase error introduced by the PHY components is well controlled, possibly, according to strict standard rules, to within a few ns.”
    - Current standard rules do not enforce limits on delay variations.
  - The location of timestamping in Tx and Rx may affect the timing noise
    - This is true. Implementations that disregard this can cause variations of ~10 ns in existing PHYs (25G and 100G). There are ways to reduce the noise to within a few ns.
  - “it is considered important that a consistent approach is followed among vendors in order to control the timestamping noise”
    - For FEC encoding/decoding delay, we recommend a consistent approach as detailed above.
    - For location of timestamping, we don’t think interoperability is impacted and thus there is no need for specifying a single solution.
- Q13 asks
  - “advice concerning the specific aspects”
    - ?
  - “updates concerning actions that may have been taken in order to guarantee that delays added by the Ethernet PHY components are properly controlled”
    - We are discussing possible additions to clause 90 to clarify/recommend ways to reduce timing error
  - “consider if any action would be required to be initiated”
    - An ad hoc was formed to address the concern and recommend actions.

Ad hoc output:

- Presentation with proposed changes if any.
  - Attempt to get the changes into the P802.3cj revision project
- Liaison response (should be approved by the EC)
  - Planned at January 2018 interim meeting
- Report for the November 2017 closing plenary.

IEEE 802.3 ITU-T SG15 Q13 liaison letter ad hoc Minutes, November 8 2017

- Preliminary meeting 7:30 am
  - Marek Hajduczenia, Pete Anslow, David Ofelt, Arthur Marris, Adee Ran
- Ad hoc meeting 9:00 am
  - Arthur Marris, Steve Carlson, Adrian Butter, Kai Yang