RS-FEC synchronization for 25 Gb/s and beyond

Adee Ran, Intel Corp.

Kent Lusted, Intel Corp.

Contributions

- Assaf Benhamou, Intel
- Yoni Landau, Intel
- Mark Gustlin, Xilinx

Thanks!

Goals

- Explore RS-FEC codeword synchronization methods for 25G Ethernet (P802.3by)
- Consider communality with 100G (clauses 82 and 91)
- Consider applicability to 400G (the obvious 16x25 use case) and possible future 50G (single lane or 2x25)
  - Speculative – nice to have
- Examine PCS requirements
- Lay out options for consensus building

RS-FEC encoding and 257-bit transcoding scheme from clause 91 are assumed

Alignment markers

- Purpose of alignment markers:
  - Identify and de-skew physical lanes (for multi-lane distribution)
  - Error monitoring using BIP – unnecessary when FEC is used
  - Synchronize codeword boundaries (in clause 91)
  - For EEE purposes: identify no-signal condition, quick synchronization using RAMs
- AMs are used by both the PCS and the RS-FEC – different form, same function
  - Some elastic buffer functionality is assumed at the PCS – exchanging idles and groups of AMs.
  - In clause 91 AMs are removed and re-inserted, maintaining a constant throughput.
    - An alternative interpretation is that AMs are *transcoded* into a bit pattern that is distributed over the RS-FEC lanes, such that the resulting output of the transmitter lanes allows easy alignment at the receiver (different per lane with a common prefix).
    - This bit pattern appears as the PCS AM payloads when viewed at the output of each lane, plus a 5-bit pad – but it's an arbitrary choice.
    - For this presentation, we will refer to this bit pattern as a **transcoded alignment marker** (**TAM**).
- For 25G with RS-FEC, the main interest is codeword synchronization.
  - The no-FEC option is not addressed in this presentation.

Desired appearance of TAMs at RS-FEC output

- There are two main requirements:
  1. For easy integration with transcoding (making TAMs fit into whole transcoded blocks), the TAM size should be a multiple of 257 bits (equivalently, the total number of AMs inserted by the PCS should be a multiple of 4).
  2. For codeword synchronization, TAM period (in PCS blocks across all lanes) should be a multiple of 80 (the number of PCS blocks in a codeword payload).
- In 802.3bj, both requirements were met by having 20 PCS lanes at 5 Gb/s: a TAM is a block of 20 AMs that appears every $20*2^{14}$ PCS blocks.
- This doesn’t happen naturally in 25G, nor in 50G…
- Let’s consider some alternatives

Option #1: use 257-bit TAMs

- A similar idea is described in [1], slide 14
- TAM is equivalent to four PCS blocks
  - For 25G, a single-lane PCS periodically inserts a group of four AMs
  - A possible future 50G can use a two-lane PCS, and periodically inserts a group of two AMs on each lane.
- TAM period must be a multiple of 20 66-bit blocks
  - [1] suggests $5*2^{14}$; We should also address separation of RAMs for EEE
- Keep most RS-FEC logic
  - Input lane alignment and output lane distribution are modified
  - AM re-insertion is different (equivalently, a new TAM format is required)
- TAM appearance on the RS-FEC lanes:
  - For 25G, TAM is a single 257-bit block on a single lane; can be any pattern; receiving RS-FEC replaces it with the AMs.
  - For 50G, TAM consumes 13 full symbols on lane 0, and 12 symbols + 7-bit pad on lane 1 (13*10+12*10+7=257).
    - TAM should be constructed such that a common prefix appears on both lanes.
- This would not work for 400G (25x16): PCS needs 16 lanes to align fiber skew.

[1] “25G Ethernet Layering and Gaps”, baden_25GE_01_1114

Option #2: keep the 20 lanes

- A similar idea is included in [2]
- Keep all the clause 82 PCS definitions but run at a different bit rate
    - PCS lane bit rate varies per Ethernet speed: 25G at 1.25 Gb/s, 400G at 20 Gb/s, 50G at 2.5 Gb/s, to maintain 20 PCS lanes.
- Keep all RS-FEC definitions (with a possible exception of output lane distribution)
    - TAM is 5*257=1285 bits long
    - If RS-FEC lane distribution is still in 10-bit symbols
        - For 25G and 50G, TAM insertion is 20 payloads on a single lane, or 10 payloads per lane, respectively + 5-bit pad (based on clause 91).
        - For 400G over 16 lanes, the TAM includes 8 full symbols per lane + 5-bit pad (8*10*16+5=1285).
    - In [2] a bit-muxing scheme (clause 83 PMA) is suggested instead of symbol-muxing
        - This keeps the output lane distribution of clause 91 unchanged too
        - However, it would not work for 400G; in addition, error propagation may somewhat weaken the code.
- RS-FEC logic design essentially unchanged between 25G and 100G
- For 25G and 50G, large TAM causes larger buffers, longer lock times and higher latency variation than necessary.
- For 400G, if PCS and FEC are not co-located, additional gearboxing is required to map the 20 PCS lanes onto 16 CDAUI lanes and back.

2. “Sub-Layering for 25GbE”, lo_25GE_01_1114

Option 3: keep the 5G per PCS lane

- Re-use clause 82 PCS, but with a variable number of 5 Gb/s lanes
  - 5 lanes for 25G, 10 lanes for 50G, 80 lanes for 400G
  - New AM encodings (a lot) required for the PCS in the 400G case
- Several RS-FEC changes compared to clause 91:
  - Input from PCS: Lane sync, alignment, and reordering have a different number of lanes.
  - Transcoding:
    - For 25G and 50G, a TAM occupies a non-integer number of 257-bit blocks. The final portion of the TAM should be transcoded alongside regular blocks. This part can be treated as a data block for transcoding purposes (so it occupies 64 or 128 bits respectively).
    - For 400G, a TAM spans 20 full 257-bit blocks (5140 bits) – exactly a codeword payload.
  - TAMs appearance on output lanes:
    - 25G: an arbitrary bit pattern, should be transcoded back into 5 AMs
    - 50G: a bit pattern defined such that it creates unique patterns per lane with a common prefix (not a simple re-insertion)
    - 400G: can be similar to clause 91 – 5 AM payloads on each lane + 20-bit pad
- AM separation from clause 82 kept unchanged – TAM will align with RS-FEC codewords
- For 400G, huge TAM causes larger buffers and higher latency variation than necessary
- For 400G, if PCS and FEC are not co-located, simple 5:1 bit-level mux/demux operations map the 80 PCS lanes onto 16 CDAUI lanes and back.

Option 4: use 25G per PCS lane

- Re-use clause 82 PCS, but with a variable number of 25 Gb/s lanes
  - One lane for 25G, 2 lanes for 50G, 16 lanes for 400G
  - Can re-use existing AM encodings from clause 82 even for 400G.
  - Change AM period to a multiple of 20 66-bit blocks (say, 20*512=10240) to enable alignment with RS-FEC codewords.
- Several RS-FEC changes compared to clause 91:
  - Input from PCS: Lane sync, alignment, and reordering have a different number of lanes.
  - Transcoding:
    - For 25G and 50G, a TAM is a part of a 257-bit block. Can be treated as a data block for transcoding purposes (so it occupies 64 bits or 128 bits respectively).
    - For 400G, a TAMs spans 4 full 257-bit blocks.
  - AMs on output lanes:
    - 25G: arbitrary bit pattern
    - 50G and 400G: a bit pattern defined such that it creates unique patterns per lane with a common prefix (not a simple re-insertion).
- For 400G with CDAUI-16, even if PCS and FEC are not co-located, same number of lanes removes need for bit-muxing, this way is more tolerant to error bursts.
- 25G and 50G can also have an AUI above or below the RS-FEC.

Option 5: no AMs

- Based on clause 49 PCS behavior when clause 74 FEC is in used during LPI wake (EEE).
- Use scrambler bypass to allow fast codeword synchronization after AN and training are completed (even if EEE is not used).
  - Alignment can be found quickly based on known incoming data, instead of testing 5140 possible codeword alignments.
- Natural PCS choice would be clause 49 (no AMs)
  - Would be useful for OTN (same PCS encoding regardless of FEC usage)
  - Required additions: new states and timers for scrambler_bypass control in transmit and receive state diagrams
- Re-use most of clause 91 RS-FEC
  - Input: Lane sync, alignment, and reordering are not required; no AM removal
  - Output: no AM re-insertion
- Does not work for MLD PHYs (400G and possibly 2-lane 50G)

Summary

- In 25G terms, the presented 5 options are:
  1. Use 4 AMs, 257-bit TAMs (baden_25GE_01_1114)
  2. Use 20 AMs, 1285-bit TAMs (lo_25GE_01_1114)
  3. Use 5 AMs, TAMs occupy non-integer number of transcoded blocks
  4. Use 1 AM, TAMs occupy a part of a transcoded block
  5. Use no AMs
- Each option has different merits
  - For 25G only, options 1, 2 and 5 are simplest and seem most suitable
  - Options 1, 2, 3 and 4 can be used for both 25G and 50G
  - Option 4 seems most suitable for 400G
  - Option 5 can help operation over OTN
- We don’t have to choose now – but we have some available solutions

What’s next?

- Once we become a task force, discuss options and hopefully build consensus around one
- Create a baseline proposal for 25G RS-FEC
- Possibly re-use for 400G