800GbE PCS/FEC/PMA Baseline Proposal for PHYs using 8 x 100G PMD lanes - Update

Kapil Shrikhande (Marvell), Eugene Opsasnick (Broadcom), Gary Nicholl (Cisco), David Ofelt (Juniper), Eric Maniloff (Ciena), Shawn Nicholl (AMD), Jeff Slavick (Broadcom)

July 12, 2022
IEEE 802.3df Plenary meeting, July 2022
Supporters

• Rob Stone, Meta
• Brad Booth, Microsoft
• Kent Lusted, Intel
• Brian Welch, Cisco
• Lenin Patra, Marvell
• Venu Balasubramonian, Marvell
• Piers Dawe, Nvidia
• Bill Simms, Nvidia
• Arthur Marris, Cadence
• Liav Ben Artsi, Marvell
• Jerry Pepper, Keysight

• Chris Cole, Quintessent
• Ted Sprague, Infinera
• Dave Estes, Spirent
• Adee Ran, Cisco
• Chris DiMinico, PHY-SI/SenTekse
• Ben Jones, AMD
• Jeffery Maki, Juniper Networks
• Ali Ghiasi, Ghiasi Quantum LLC
• Paul Brooks, Viavi Solutions
• Nathan Tracy, TE Connectivity
This Talk

• Review updates to the Baseline

• Summary of work since May’22 interim
Outline

• Introduction
• PCS/FEC/PMA Baseline proposal
• Implementation considerations
• Architecture considerations
• Summary of work since May’22 interim
• Conclusions
Goals

• Fast time to an 800GbE PCS/FEC/PMA specification for PMDs using 100G/lane
  • Re-use 400GbE PCS/FEC (CL119) as much as possible
  • Support 800GbE with simple modification to the 400GbE PCS/FEC
  • Leverage 802.3bs Cl120 PMA; leverage 802.3ck 100G/lane PMA and AUI specifications

• Maximize the re-use of existing logic sub-blocks used in 400GbE PCS/FEC
  • Leverage industry investment in 400GbE technology

• Enable systems using current 8-lane 800G connectors (OSFP / QSFP-DD) to also support 800GbE
  • E.g. 8-lane C2M AUIs used as: 8 x 100GAUI-1 / 4 x 200GAUI-2 / 2 x 400GAUI-4 and 1 x 800GAUI-8
Scope

Scope of this Baseline: 800GbE PCS/FEC/PMA for all PHY objectives that use 8 x 100G PMDs and AUIs

### 802.3df Adopted PHY Objectives*

<table>
<thead>
<tr>
<th>Ethernet Rate</th>
<th>Assumed Signaling Rate</th>
<th>AUI</th>
<th>BP</th>
<th>Cu Cable</th>
<th>MMF 50m</th>
<th>MMF 100m</th>
<th>SMF 500m</th>
<th>SMF 1km</th>
<th>SMF 10km</th>
<th>SMF 40km</th>
</tr>
</thead>
<tbody>
<tr>
<td>200 Gb/s</td>
<td>200 Gb/s</td>
<td>Over 1 lane</td>
<td>Over 1 pair</td>
<td>Over 1 Pair</td>
<td>Over 1 Pair</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>400 Gb/s</td>
<td>200 Gb/s</td>
<td>Over 2 lanes</td>
<td>Over 2 pairs</td>
<td>Over 2 Pair</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>800 Gb/s</td>
<td>100 Gb/s</td>
<td>Over 6 lanes</td>
<td>Over 8 pairs</td>
<td>Over 8 pairs</td>
<td>Over 8 pairs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200 Gb/s</td>
<td>Over 4 lanes</td>
<td>Over 4 pairs</td>
<td>Over 4 pairs</td>
<td>1) Over 4 pairs</td>
<td>2) Over 4 1/2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TBD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.6 Tb/s</td>
<td>100 Gb/s</td>
<td>Over 16 lanes</td>
<td>Over 8 pairs</td>
<td>Over 8 pairs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200 Gb/s</td>
<td>Over 8 lanes</td>
<td>Over 8 pairs</td>
<td>Over 8 pairs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Table from [https://www.ieee802.org/3/B400G/public/21_1028/B400G_overview_c_211028.pdf](https://www.ieee802.org/3/B400G/public/21_1028/B400G_overview_c_211028.pdf)

#### Technology Reuse

- Leverage existing or work-in-progress 100 Gb/s per lane (e.g., 3cu, 3ck, 3db) to higher lane counts
- Develop 200 Gb/s per lane electrical signaling for 1/2/4/8 lane variants of AUIs and electrical PMDs
- Develop 200 Gb/s per optical fiber for 1/2/4/8 fiber based optical PMDs and 4 lambda WDM optical PMD
- Potential for either direct detect and / or coherent signaling technology

Making it all work together
AUI and PMD assumptions

• 802.3df Task Force has adopted 800GbE 8-lane AUI baseline proposals leveraging existing 100G/lane AUI specs, drafts

• 802.3df Task Force has adopted 800GbE 8-lane PMD baseline proposals leveraging existing 100G/lane PMD specs, drafts

• 802.3bs CL119 PCS works for all 100G/lane AUIs and PMDs for 400GbE

• Similarly, this PCS/FEC Baseline (leveraging CL119) works for all adopted 800GbE 8-lane AUIs and PMDs
Outline

• Introduction
• PCS/FEC/PMA Baseline proposal
• Implementation considerations
• Architecture considerations
• Summary of work since May’22 interim
• Conclusions
Architecture

* PCS and FEC are in the PCS sub-layer (same as CL119)

Note: Not showing layering diagram for Cu PMD (will be same as other Cu PMD layering diagrams in 802.3)
End-End PCS/FEC scheme for 800GbE (8 x 100G) PMDs

Note: This End-End PCS/FEC works with optional Chip to Chip AUJIs and a combination of Chip to chip and Chip to module (same as 400GAIU-4 in 802.3ck)
Tx PCS/FEC Data Flow

• Based on two 802.3bs, CL119 sub-layers in parallel
  • Two 400G FEC flows (flow-0 and flow-1)
• 66b round robin distribution into two 400G flows after 64B/66B encode
• Sub-blocks shown within each flow are identical to CL119, except:
  • AM values are made unique across the two flows
  • AM insertion is aligned across the two flows
• 32 PCS lanes per 800GbE PCS
  • 16 PCS lanes per 400G flow
• Any 4 PCS lanes to any PMA output lane
  • 4:1 bitmuxing
Tx 66b Block Distribution

• Round Robin among two ‘400G Flows’
Alignment Marker Insertion

- **802.3bs 400G AM structure**
  - AM size = 8 x 257b
  - Spacing = 160k x 257b = 8192 CWs
- **AM total sizing for 800G = 2x400G**
  - AM size = 16 x 257b
  - Spacing = 320k x 257b = 16384 CWs
- **Markers inserted at consecutive 257b blocks across both 400G flows**
  - Flow-0 is first in time carrying the even encoded 4x66b blocks
  - Flow-1 carries odd encoded 4x66b blocks

Source: IEEE Std 802.3-2018
AM Marker Encoding

- CM0-CM5 and UP0-UP2 are unchanged from 400GbE CL119
- UM0/UM3 for PCS lanes 0-15 are inverted from 400GbE
- UM1/UM2/UM4/UM5 for PCS lanes 16-31 are inverted from 400GbE
- Prevents lock with 400GbE ports
- Maintains DC balance

Note: in table above, bolded text indicates changes from CL 119 AM values
Rx PCS/FEC Data Flow

- **Alignment Lock and Deskew**
  - AM lock: per lane, same as CL119
  - De-skew: across 32 PCS lanes

- **Lane reorder (and split)**
  - Reorder and split 32 PCS lanes into 2 groups of 16
    - Lanes 0-15: Flow-0
    - Lanes 16-31: Flow-1

- **FEC decode, de-scramble, transcode decode – same as CL119**

- **Round robin block collection must be aligned across Flow-0/1 based on Alignment Marker location**
Rx 66b Block Collection

• Round Robin 66b Block Collection is opposite of Tx Block Distribution
Re-use CL119 State Diagrams

• Re-use all of the following
  • Figure 119–12—Alignment marker lock state diagram
  • Figure 119–13—PCS synchronization state diagram
  • Figure 119–14—Transmit state diagram
  • Figure 119–15—Receive state diagram

• Minor modification to the following
  • Add restart_lock<y> variable per 400G flow
    • restart_lock = restart_lock<0> OR restart_lock<1>
  • Add hi_ser<y> variable per 400G flow
    • hi_ser = hi_ser<0> OR hi_ser<1>
PMA

• PMA functions as defined in CL120, with latest 802.3ck updates for 100G/lane
  • Bit-multiplexing (4:1)
  • Modulation (PAM4)
  • AUI Physical lane instantiation (8 lane)
  • Signaling lane rate (106.25Gb/s)
  • Coding (Gray, precoding)
  • Clock and data recovery
  • Loopbacks
  • Test patterns

• Per lane AUI specifications from 802.3ck
PMA Muxing

• Any 32 PCS Lanes to Any 8 PMA Lanes
  • 4:1 Bit-multiplexing of data from any 4 PCS lanes to any 1 PMA lane
  • The receiver can receive PMA lanes in any order and has a full 32 lane reorder block
  • Clock content is same as a 400GE CL119 stream
    • Analysis completed and presented in wong_3df_logic_220630
Outline

• Introduction
• PCS/FEC/PMA Baseline proposal
• **Implementation considerations**
• Architecture considerations
• Summary of work since May’22 interim
• Conclusions
Latency considerations

• Two 400GbE FEC encode/decode engines in parallel
• FEC latency for this baseline proposal same as 400GbE FEC latency
Many 800G implementations will support 100/200/400/800GbE Ethernet ports

- 32 PCS lanes already exist to support 2 x 400GbE / 4 x 200GbE / 8 x 100GbE!
- Reuse of per lane PCS alignment logic

<table>
<thead>
<tr>
<th>800Gb/s Block config</th>
<th>PCS/FEC lanes per Ethernet port</th>
<th>Total PCS/FEC lanes per 800Gb/s block</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 x 800GbE port</td>
<td>32 lanes @ 25G</td>
<td>32</td>
</tr>
<tr>
<td>2 x 400GbE ports</td>
<td>16 lanes @ 25G</td>
<td>32</td>
</tr>
<tr>
<td>4 x 200GbE ports</td>
<td>8 lanes @ 25G</td>
<td>32</td>
</tr>
<tr>
<td>8 x 100GbE ports</td>
<td>4 lanes @ 25G</td>
<td>32</td>
</tr>
</tbody>
</table>

Choice of 32 PCS lanes can enable implementations over 16 x 50G AUI lanes
- If needed (e.g. test equipment)
Other Implementation Considerations

• This baseline benefits from the use of two 400GbE PCSs in parallel
  • Reuse of logic blocks from 400GbE PCS possible
  • FEC engines, transcoder, scramblers running at same bandwidth as 400GbE
  • Per lane alignment lock running at same speed as 400GbE
  • Minimizes new development and verification

• This baseline follows the approach taken by the adopted 800GbE 8-lane AUIs and PMD baselines
  • 800GbE 8-lane AUIs and PMDs are doubling number of lanes from 400GbE
    • Example 1: 800GAUI-8 is 2 x 400GAUI-4 in parallel
    • Example 2: 800GBASE-DR8 is 2 x 400GBASE-DR4 in parallel
  • Allows re-use of specifications, maximize use of technology and investment from 400GbE
Outline

• Introduction
• PCS/FEC/PMA Baseline proposal
• Implementation considerations
• Architecture considerations
• Summary of work since May’22 interim
• Conclusions
IEEE P802.3df Architecture: FEC schemes

End-to-End FEC scheme
(FEC1 used for AUIs and PMD)

Concatenated FEC scheme
(FEC2 is added on top of FEC1. FEC1 for AUIs, FEC1+FEC2 for PMD)

Segmented FEC scheme
(FEC2 replaces FEC1. FEC1 used for local AUI only. FEC2 for PMD only)
800GbE Architecture : FEC schemes over AUI-8

End-to-End FEC scheme
Targeted by this Baseline

- MAC/RS
- PCS
- PMA (32:8)
- PMA (8:8)
- PMD
- MDI

800GAUI-8

800GBASE-CR8/KR8, 800GBASE-VR8/SR8, 800GBASE-DR8/DR8+

Concatenated FEC scheme

- MAC/RS
- PCS
- PMA (32:8)
- PMA (8:32)
- FEC
- MDI

800GAUI-8

800GBASE-CR8/KR8, 800GBASE-VR8/SR8, 800GBASE-DR8/DR8+

Segmented FEC scheme

- MAC/RS
- XS
- PMA (32:8)
- PMA (8:32)
- FEC
- MDI

800GAUI-8

800GBASE-CR8/KR8, 800GBASE-VR8/SR8, 800GBASE-DR8/DR8+

Other PMDs (TBD)

Included in this baseline

Other FEC schemes / evolution remains open
**800GbE Architecture: FEC schemes over AUI-4**

**End-to-End FEC scheme**

- MAC/RS
- PCS
- PMA

**Concatenated FEC scheme**

- MAC/RS
- PCS
- FEC
- PMA

**Segmented FEC scheme**

- MAC/RS
- XS

*FEC1 could be the FEC proposed in this Baseline, or it could be a different FEC. Evolution options remain open.*
Outline

• Introduction
• PCS/FEC/PMA Baseline proposal
• Implementation considerations
• Architecture considerations
• **Summary of work since May’22 interim**
• Conclusions
Summary of work since May’22

• FLR analysis completed and presented in Logic Ad hoc (06/30/22)
  • See opsasnick_3df_logic_220630a
  • Baseline meets the 6.2E-11 FLR requirement corresponding to the 1E-13 BER objective
  • Addressed questions raised by X. Wang

• FLR analysis using burst error model completed and presented in Logic Ad hoc (06/30/22)
  • See opsasnick_3df_logic_220630a
  • Burst error performance looks good, no FLR floor observed
  • Some FEC gain is possible using a cross-flow bit-muxing to interleave bits from 4 codewords

• Clock content analysis completed and presented in Logic Ad hoc (06/30/22)
  • See wong_3df_logic_220630
  • Clock content is same as a 400GE CL119 stream
  • Analysis was pending from May’22 baseline presentation
Outline

• Introduction
• PCS/FEC/PMA Baseline proposal
• Implementation considerations
• Architecture considerations
• Summary of work since May’22 interim
• Conclusions
Conclusions

• This Baseline: 800GbE PCS, FEC and PMA for 8 x 100G PMDs and 8 x 100G AUls
• Supports all adopted 802.3df copper and optical PMDs baselines using 100G/lane
• Highly leverages existing 400GbE specifications
  • 2 x 400GbE (Clause 119) with minor modifications to the specifications
• Highly leverages existing 400GbE implementations
  • Enable re-use of per-lane AM lock, FEC interleaving, FEC encode/decode, scrambler, transcoder
• Meets the FLR requirement corresponding to 1E-13 BER objective
• Clock Content is same as a 400GbE CL119 stream
• Enables faster time-market for 800GbE (8 x 100G/lane) implementations
  • Maximizing technology reuse and existing industry investments
• Fits into an overall 800GbE Logic Architecture, and does not constrain future FEC schemes using 200G/lane AUls and PMDs and/or Coherent PMDs
• 1.6TbE PCS/FEC can be chosen independently of 800GbE
  • Decisions made in this baseline will not restrict options / choices for 1.6TbE
Thanks !
Backup – FLR Analysis Data for Random and Burst Errors

- FLR data from opsasnick_3df_logic_220630a
- Additional data added for 400GbE for comparison
BER_{in} and SNR Requirements with Random Errors

<table>
<thead>
<tr>
<th>RS(544,514) FEC</th>
<th>FLR Target</th>
<th>FSF</th>
<th>CER Required</th>
<th>BER_{in} Required</th>
<th>PAM4 DER Required</th>
<th>SNR (dB) Required</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Interleave</td>
<td>6.2E-11</td>
<td>1.125</td>
<td>5.49E-11</td>
<td>3.20E-4</td>
<td>6.40e-4</td>
<td>17.45</td>
</tr>
<tr>
<td>2 CW Interleave</td>
<td>6.2E-11</td>
<td>2.125</td>
<td>2.92E-11</td>
<td>3.06E-4</td>
<td>6.13E-4</td>
<td>17.48</td>
</tr>
<tr>
<td>4 CW Interleave</td>
<td>6.2E-11</td>
<td>4.125</td>
<td>1.50E-11</td>
<td>2.93E-4</td>
<td>5.85E-4</td>
<td>17.52</td>
</tr>
</tbody>
</table>

- 100G/lane PMDs assume BER_{in} = 2.4E-4 or better.
  - See “Bit Error Ratio” in Clauses 124, 140, 151, etc.
  - Expand requirement to include two AUI on each end of the link, adds 4 * 1E-5 = 2.8E-4

- Even if BER_{in} is worse than 2.8E-4, all Interleaves meet the 6.2E-11 FLR Target
- SNR increase to meet the same FLR from 2-way to 4-way FEC interleave is ≈ 0.04dB (negligible)
### Burst Error Results for 8x100 PCS Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Required FLR</th>
<th>1+0.1D no precoding (a=0.01)</th>
<th>1+0.5D no precoding (a=0.375)</th>
<th>1+D with precoding (a=0.75)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Required SNR</td>
<td>Required DER</td>
<td>Required SNR</td>
<td>Required DER</td>
</tr>
<tr>
<td>1.a</td>
<td>6.20E-11</td>
<td>17.49</td>
<td>6.09E-04</td>
<td>5.40E-04</td>
</tr>
<tr>
<td></td>
<td>17.57</td>
<td>18.09</td>
<td>2.49E-04</td>
<td></td>
</tr>
<tr>
<td>1.b</td>
<td>6.20E-11</td>
<td>17.49</td>
<td>6.07E-04</td>
<td>17.79</td>
</tr>
<tr>
<td></td>
<td>17.96</td>
<td>18.22</td>
<td>2.03E-04</td>
<td></td>
</tr>
<tr>
<td>2.a</td>
<td>6.20E-11</td>
<td>17.52</td>
<td>5.79E-04</td>
<td>18.41</td>
</tr>
<tr>
<td></td>
<td>18.44</td>
<td>18.44</td>
<td>1.39E-04</td>
<td></td>
</tr>
<tr>
<td>2.b</td>
<td>6.20E-11</td>
<td>17.52</td>
<td>5.80E-04</td>
<td>17.83</td>
</tr>
<tr>
<td></td>
<td>17.83</td>
<td>18.06</td>
<td>2.61E-04</td>
<td></td>
</tr>
<tr>
<td>400Gbe</td>
<td>6.20E-11</td>
<td>17.49</td>
<td>6.07E-4</td>
<td>18.36</td>
</tr>
<tr>
<td></td>
<td>18.40</td>
<td>18.40</td>
<td>1.50E-04</td>
<td></td>
</tr>
</tbody>
</table>

- 1+0.1D: Nearly random (a=0.01)
  - Option 1.a, 2 CW interleave, is 0.03dB better than 4 CW Interleave

- 1+D: High burst correlation (a=0.75)
  - Option 2.b, 4:1 bitmux across 4 CW, is best option by 0.03dB

Option 1.a (SM + CI2, FSF=2.125) is for proposal from [wang_3df_logic_220623a.pdf](#)
Option 2.a (BM4 + CI2, FSF=4.125) and Option 2.b (BM4 + CI4, FSF=4.125) is for this baseline.
400GbE uses (BM4 + CI2, FSF=2.125)
Option 1.a (random, FSF=2.125) & (SM + CI2, FSF=2.125) (light blue) is for proposal from wang_3df_logic_220623a.pdf
Option 2 (random, FSF=4.125) & 2.a:(BM4 + CI2, FSF=4.125) (solid red) and 2.b:(BM4 + CI4, FSF=4.125) (dark blue) is for this baseline.
400Gbe uses (BM4 + CI2, FSF 2.125) (dashed red)