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

Kapil Shrikhande (Marvell), Eugene Opsasnick (Broadcom), Gary Nicholl (Cisco), Matt Brown (Huawei), Xinyuan Wang (Huawei), Mark Gustlin (Cisco), David Ofelt (Juniper), Eric Maniloff (Ciena), Shawn Nicholl (AMD), Jeff Slavick (Broadcom), Adee Ran (Cisco), Kent Lusted (Intel), Jerry Pepper (Keysight)

October 2022
IEEE 802.3df Task Force
Supporters

- Rob Stone, Meta
- Lenin Patra, Marvell
- Arthur Marris, Cadence
- Howard Heck, Intel
- Venu Balasubramonian, Marvell
- Leon Bruckman, Huawei
- Yan Zhuangyan, Huawei
- Chris Cole, Quintessent
- Jeffery Maki, Juniper Networks
- Nathan Tracy, TE Connectivity
- Daniel Koehler, Synopsys
- David Malicoat, Malicoat Networking Solutions
- Frank Effenberger, Huawei
- Joshua Kim, Hirose Electric
- Kenneth Jackson, Sumitomo Electric
- Rick Rabinovich, Keysight
- Vipul Bhatt, II-VI
- Tom Palkert, MACOM and Samtec
- Mau-Lin Wu, Mediatek
- Ted Sprague, Infinera
- Dave Estes, Spirent
- Ed Nakamoto, Spirent
- Chris DiMinico, PHY-SI/SenTekse
- Ben Jones, AMD
- Ali Ghiasi, Ghiasi Quantum LLC
- Paul Brooks, Viavi Solutions
- Megha Shanbhag, TE Connectivity
- Pavel Zivny, Tektronix
- Mike Dudek, Marvell
- Shimon Muller, Enfabrica Corp.
- Xiang He, Huawei
- Viet Tran, Keysight
- Brian Welch, Cisco
- Phil Sun, Credo
- Lokesh Kabra, Synopsys
- Roberto Rodes, II-VI
- Brad Booth, Microsoft
Outline

• Introduction
• PCS/FEC/PMA Baseline proposal
• Conclusions
Goals

• Fast time to an 800GbE PCS/FEC/PMA specification for PMDs and AUIs using 100G/lane
  • Reuse 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 PMD and AUI specifications

• Maximize the reuse 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

**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 2km</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 8 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>TBO</td>
<td>Over 4 lanes</td>
<td>Over 4 pairs</td>
<td>1) Over 4 pairs</td>
<td>2) Over 4 pairs</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>Over 8 pairs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200 Gb/s</td>
<td></td>
<td>Over 8 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>
</tbody>
</table>

**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 AUls and electrical PMDs
- Development 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

---

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

* 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)
AUI and PMD assumptions

• 802.3df Task Force has adopted 800GbE 8-lane AUI baseline proposals leveraging existing 100G/lane AUI specs, drafts
  • [https://www.ieee802.org/3/df/public/22_03/lusted_3df_01a_220315.pdf](https://www.ieee802.org/3/df/public/22_03/lusted_3df_01a_220315.pdf)

• 802.3df Task Force has adopted 800GbE 8-lane PMD baseline proposals leveraging existing 100G/lane PMD specs, drafts
  • [https://www.ieee802.org/3/df/public/22_03/lusted_3df_01a_220315.pdf](https://www.ieee802.org/3/df/public/22_03/lusted_3df_01a_220315.pdf)
  • [https://www.ieee802.org/3/df/public/22_02/welch_3df_01a_220222.pdf](https://www.ieee802.org/3/df/public/22_02/welch_3df_01a_220222.pdf)
  • [https://www.ieee802.org/3/df/public/22_03/murty_3df_01a_220315.pdf](https://www.ieee802.org/3/df/public/22_03/murty_3df_01a_220315.pdf)

• 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
• Conclusions
Architecture

*PCS and FEC are in the PCS sublayer (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 AUIs and a combination of Chip to chip and Chip to module (same as 400GAUI-4 in 802.3ck)
Tx PCS/FEC Data Flow

- Based on two 802.3bs, CL119 sublayers 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 Flow lanes per 800GbE PCS
  - 16 per 400G flow
- Specific Flow lanes mapped to a given PMA output lane
  - 4:1 bit-muxing
  - Lanes chosen so all 4 FEC codewords are equally represented on each PMA output lane
  - Bitmux can be specified to occur in either the PCS or PMA sublayer (TBD).
Tx 66b Block Distribution

• Round Robin among two ‘400G Flows’

From 64B/66B encoding

Transcodes into 257b

400G flow-0

400G flow-1
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
AM Marker Encoding

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

<table>
<thead>
<tr>
<th>Flow Lane #</th>
<th>Encoding</th>
</tr>
</thead>
<tbody>
<tr>
<td>CM0</td>
<td>0x9A</td>
</tr>
<tr>
<td>CM1</td>
<td>0x4A</td>
</tr>
<tr>
<td>CM2</td>
<td>0x26</td>
</tr>
<tr>
<td>UP0</td>
<td>0x86</td>
</tr>
<tr>
<td>CM3</td>
<td>0x65</td>
</tr>
<tr>
<td>CM4</td>
<td>0x05</td>
</tr>
<tr>
<td>CM5</td>
<td>0x09</td>
</tr>
<tr>
<td>CM6</td>
<td>0x09</td>
</tr>
<tr>
<td>UP1</td>
<td>0x0F</td>
</tr>
<tr>
<td>UM0</td>
<td>0x3F</td>
</tr>
<tr>
<td>UM1</td>
<td>0x26</td>
</tr>
<tr>
<td>UM2</td>
<td>0x26</td>
</tr>
<tr>
<td>UM3</td>
<td>0x26</td>
</tr>
<tr>
<td>UM4</td>
<td>0x26</td>
</tr>
<tr>
<td>UM5</td>
<td>0x26</td>
</tr>
<tr>
<td>UP2</td>
<td>0x26</td>
</tr>
<tr>
<td>UP3</td>
<td>0x26</td>
</tr>
<tr>
<td>UP4</td>
<td>0x26</td>
</tr>
<tr>
<td>UP5</td>
<td>0x26</td>
</tr>
</tbody>
</table>

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

Diagram:
- Round Robin 66b-block collection
- 4x66b blocks from reverse transcode
- 400G flow-0
- 400G flow-1
- To 64B/66B Decode
Reuse CL119 State Diagrams

• Reuse 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\(<1>\)
Flow lane Muxing

• 32 Flow Lanes to 8 PMA Lanes such that
  • Each PMA lane is a result of bitmux of 2 flow lanes from Flow 0 and 2 flow lanes from Flow 1
    • This applies to all PMAs in the PHY
  • The PCS receiver includes full 32 lane reorder and deskew block so that
    • Any PMA output lane can connect to any PMA input lane
    • There can be non-zero skew between the 32 lanes (same skew limits as CL120)
Flow lane Muxing (Tx): Example illustrating proposed change

Baseline proposal from July 2022

Example bitmuxing that meets new proposal
PMA

• PMA functions as defined in CL120, with latest 802.3ck updates for 100G/lane
  • Bit-multiplexing (4:1) [if PMA Service interface below the PCS is 32 lanes]
  • 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
Outline

• Introduction
• PCS/FEC/PMA Baseline proposal
• Conclusions
Conclusions

• This Baseline: 800GbE PCS, FEC and PMA for 8 x 100G PMDs and 8 x 100G AUIs

• Supports all adopted 802.3df copper and optical PMDs baselines using 100G/lane

• Fits into an overall 800GbE Logic Architecture
  • Does not constrain future PCS/FEC/PMA schemes using 200G/lane AUIs, 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