Google

FEC Requirements for 800GbE/1.6TbE Optics - An Operator's Perspective

Shuang Yin, Xiang Zhou, and Cedric Lam

IEEE 802.3df July Plenary Meeting

Outline

- FEC requirements
  - Performance
  - Latency
  - Power
  - Backward compatibility
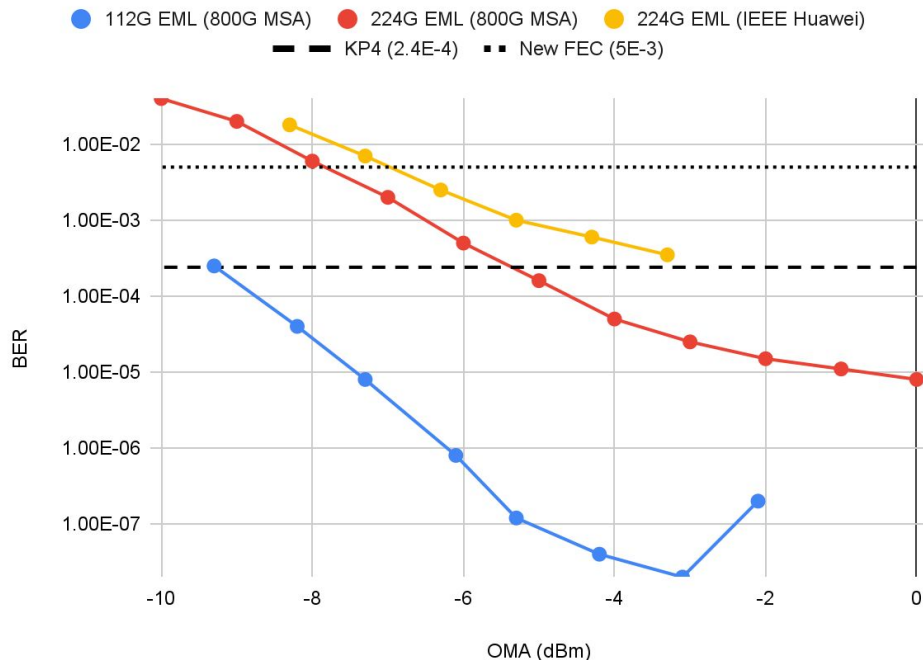- Potential FEC candidates
- Other thoughts

FEC Requirements - Performance

PAM4 Receiver Performance

- 112G EML (800G MSA)
- 224G EML (800G MSA)
- 224G EML (IEEE Huawei)
- KP4 (2.4E-4)
- New FEC (5E-3)

- **Target BER of ~5e-3** for 200G per lane PMD to lower Rx SNR requirement

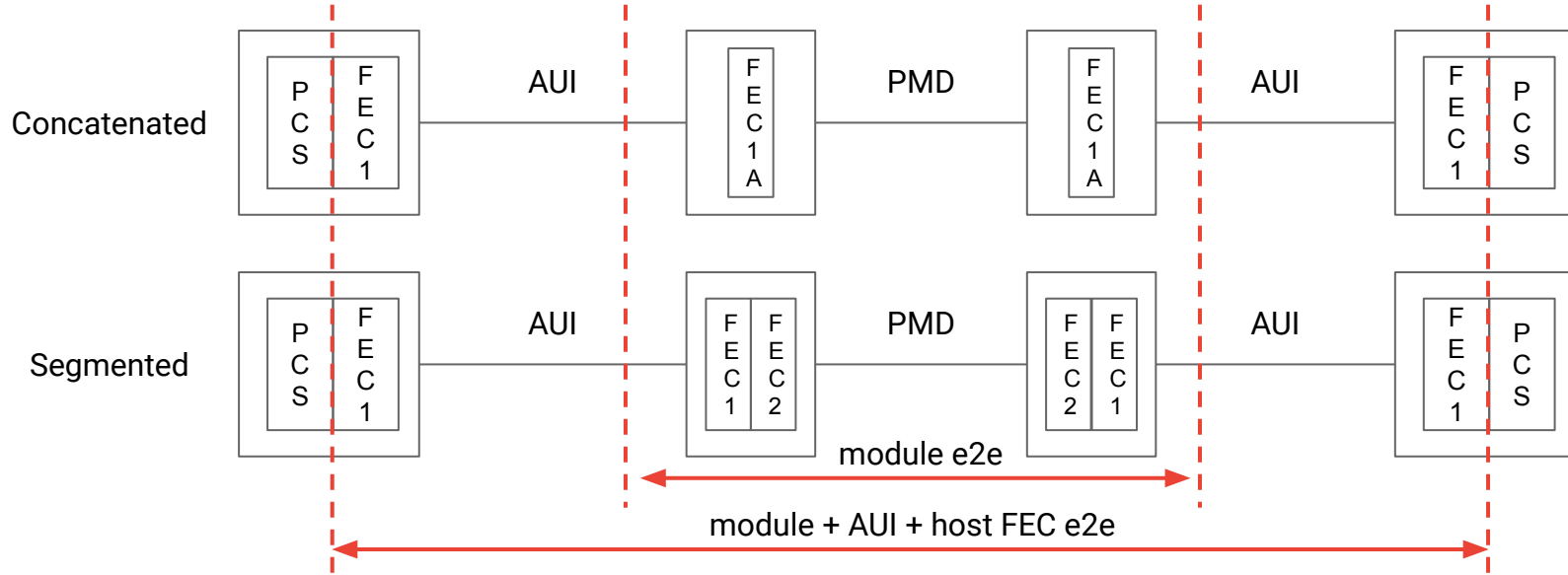# FEC Requirements - Latency

- **<u>Latency matters</u>**
  - AI/HPC, disaggregated architecture (RDMA/RoCE), etc.
    - High performance Ethernet NEA (<u>zhuang_nea_01_220622</u>)
    - Aquila* (low-latency fabric for data center)
    - Both cases advocated for sub-10 us cluster/network fabric, i.e., **~100 ns on FEC matters**
  - Traditional network operators on private line related applications (<u>wang_3df_01a_220609</u>)
- Current 100G per lane 400GbE module, e.g., 400G-FR4, has ~200 ns e2e latency (~100 ns module DSP** + ~100 ns host KP4 FEC**)
- For 200G per lane 800GbE/1.6TbE module, target **<u>< 400 ns</u>** e2e latency
  - Allocate additional latency for a higher coding gain FEC
  - Support majority of intra data center use cases
  - Align with 800G LR1 requirement in <u>oif2021.369.02</u> (300 ns module e2e + ~100 ns host KP4 FEC)

* <u>D. Gibson et al., Aquila: A unified, low-latency fabric for datacenter networks, NSDI'22</u>.
** <u>R. Nagarajan et al., Low Power DSP-Based Transceivers for Data Center Optical Fiber Communications (Invited Tutorial), JLT 39 (16), 2021.</u>

Google

FEC Latency Example

FEC1: KP4
FEC1A: Hamming (128, 120)
FEC2: KP4 + Hamming (128, 120)

- Module e2e: ~300 ns (concatenated) vs ~500 ns (segmented)
- Module + AUI + host FEC e2e: ~400 ns (concatenated) vs ~600ns (segmented)
- Strongly prefer the concatenated FEC with ~200 ns lower latency per link
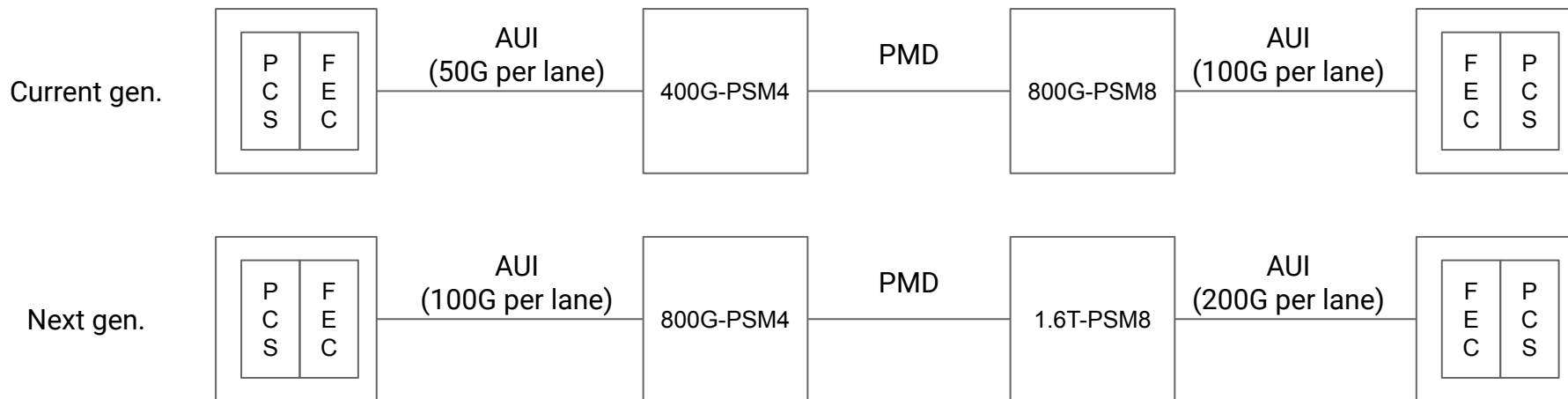
Google

FEC Requirements - Power

- Current 100G per lane 800GbE module at ~20 W
- Target at **<u>< 30 W</u>** for 200G-per-lane 1.6TbE module to allow 1RU design without RHS (close top OSFP)*
- Example study showed **<u>module FEC power of ~0.2 W/800G</u>** @5nm CMOS node (oif2022.274.00) is feasible in terms of performance and module power envelope

* OSFP-XD thermal feasibility and system level implications, OSFP-XD optics ad-hoc meeting 12/13/2021.

FEC Requirements - Backward Compatibility

- Host FEC backward compatibility
  - Enable smooth speed evolution without FEC adaptation in the module (simplifying module and DSP options)
  - Good to keep, don't break it unless absolutely have to!

Current gen.

| P C S | F E C | AUI (50G per lane) | 400G-PSM4 | PMD | 800G-PSM8 | AUI (100G per lane) | F E C | P C S |

Next gen.

| P C S | F E C | AUI (100G per lane) | 800G-PSM4 | PMD | 1.6T-PSM8 | AUI (200G per lane) | F E C | P C S |

Google

Potential FEC Candidates

FEC Scheme	FEC Proposal	Pre-FEC BER	E2E Latency*	Module FEC Power
Concatenated	RS(544, 514) + Hamming(128, 120) patra_3df_01_220518	4.85e-3	~400 ns	Low
End to end	RS(576, 514) he_b400g_01_210426	1.1e-3	~230 ns	0
Segmented	RS(544, 514) + [RS(544, 514) + BCH(144, 136)] he_b400g_01_210426	2.4e-3	~420 ns	High

* Module + AUI + host FEC e2e latency for fair comparison, assumed ~160 ns additional latency for non-FEC contributions, e.g., equalization, etc.

Google

Other Thoughts

- **For concatenated FEC scheme**
  - Proper interleaver design between outer and inner code is critical to overall FEC performance
  - Flexibility in interleaver is preferred to trade off between code gain/performance and latency (bliss_3df_01a_220517 and patra_3df_01_220518)
  - Appropriate inner code telemetry is needed for ease of operation
- **To maintain FEC backward compatibility in 200G per lane host (KP4)**
  - 200G C2M channel needs to operate better than 1e-4, preferably < 5e-5 (wang_3df_01_220517)
  - Early studies (tli_3df_01b_220316, oif2022.174.01, and ghiasi_3df_01a_220518) showed 200G C2M channel can work at 1e-5

Google

Summary

- FEC requirements for 800GbE/1.6TbE optics (200G per lane) are elaborated in terms of performance, latency and power.
- Concatenated RS(544, 514) + Hamming(128, 120) from patra_3df_01_220518 is a good baseline solution
- A fair comparison among different FEC approaches is needed
  - Optimizing for the lowest power and area (cost) while meeting the code gain (pre-FEC BER) and latency requirements

Google