# Clause 173 PMA bit-muxing constraints

(In support of comment #27 against D1.1)

Adee Ran, Cisco

# Support

- Paul Brooks, Viavi
- Zvi Rechtman, NVIDIA
- Xiang He, Huawei
- Hao Ren, Huawei
- Ali Ghiasi, Ghiasi Quantum
- Daniel Koehler, Synopsys
- Kent Lusted, Intel
- Piers Dawe, NVIDIA
- Zhuangyan Yan , Huawei
- Leon Bruckman, Huawei
- Upen Reddy Kareti, Cisco

# Recap

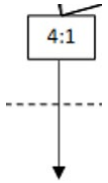
- The adopted PMA baseline places some restrictions on the PCSL muxing (173.4.2.1, 173.4.2.2, and 173.4.2.3)
  - These restrictions ensure all four codewords are present on each physical lane, which is good for FEC performance with correlated errors
- As presented in ran_3df_01a_2212, the existing restriction still allows muxing where one of two flows always gets the LSB of the PAM4 symbols
  - Labeled as “Option B”
  - Muxing this way **increases the FLR by x34 compared to 200G/400G with the same BER**
  - Muxing that shares the LSB equally across flows (“Option A”) avoids this degradation
  - Full analysis is available in the previous presentation
- Comment #6 against D1.0 suggested restricting the muxing further, but was rejected
  - Straw polls #1 and #2 were dominated by “need more information” (see comment report)
  - No change in D1.1…

# Illustration – 32:8 PMA

PCSL content of each PAM4 symbol in one physical lane
In this example, the lane muxes PCSLs 0, 1, 16, and 17 (per the restrictions in D1.1)

## Option A

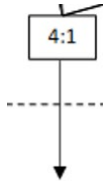
From PCS
Lanes 0, 1, 16, 17




| UI# | MSB | LSB |
|---|---|---|
| 21 | PCSL 16 | PCSL 17 |
| 20 | PCSL 0 | PCSL 1 |
| 19 | PCSL 16 | PCSL 17 |
| 18 | PCSL 0 | PCSL 1 |
| 17 | PCSL 16 | PCSL 17 |
| 16 | PCSL 0 | PCSL 1 |

## Option A
## 1 bit shifted

From PCS
Lanes 1, 16, 17, 0

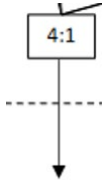



| UI# | MSB | LSB |
|---|---|---|
| 21 | PCSL 17 | PCSL 0 |
| 20 | PCSL 1 | PCSL 16 |
| 19 | PCSL 17 | PCSL 0 |
| 18 | PCSL 1 | PCSL 16 |
| 17 | PCSL 17 | PCSL 0 |
| 16 | PCSL 1 | PCSL 16 |

The LSB is shared by both flows
and all 4 codewords (A/B/C/D)
Average BER is the same for all codewords

## Option B

From PCS
Lanes 0, 16, 1, 17

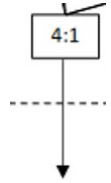



| UI# | MSB | LSB |
|---|---|---|
| 21 | PCSL 1 | PCSL 17 |
| 20 | PCSL 0 | PCSL 16 |
| 19 | PCSL 1 | PCSL 17 |
| 18 | PCSL 0 | PCSL 16 |
| 17 | PCSL 1 | PCSL 17 |
| 16 | PCSL 0 | PCSL 16 |

The LSB is only assigned to one flow
Two codewords (C/D) always get the LSB
Average BER is higher for these codewords

## (400G 16:4 PMA)

From PCS
Lanes 0, 1, 2, 3




| UI# | MSB | LSB |
|---|---|---|
| 21 | PCSL 2 | PCSL 3 |
| 20 | PCSL 0 | PCSL 1 |
| 19 | PCSL 2 | PCSL 3 |
| 18 | PCSL 0 | PCSL 1 |
| 17 | PCSL 2 | PCSL 3 |
| 16 | PCSL 0 | PCSL 1 |

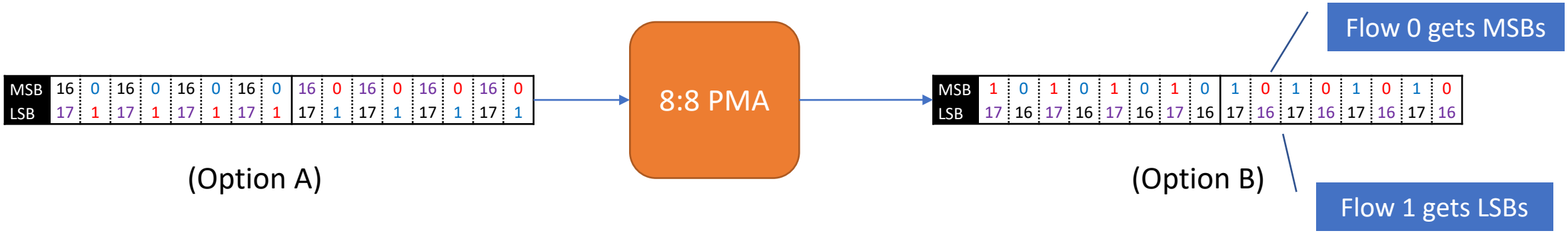
The LSB is shared by both
codewords (A/B)
Average BER is the same

# Illustration – 8:8 PMA

Currently:
"The order of PCSLs from an input lane does not have to be maintained on the output lane"

So even if the 32:8 PMA does the muxing wisely, a retimer or module is allowed to **reorder the bit stream from (0, 1, 16, 17) to (0, 16, 1, 17)…**

This would cause the following effect:



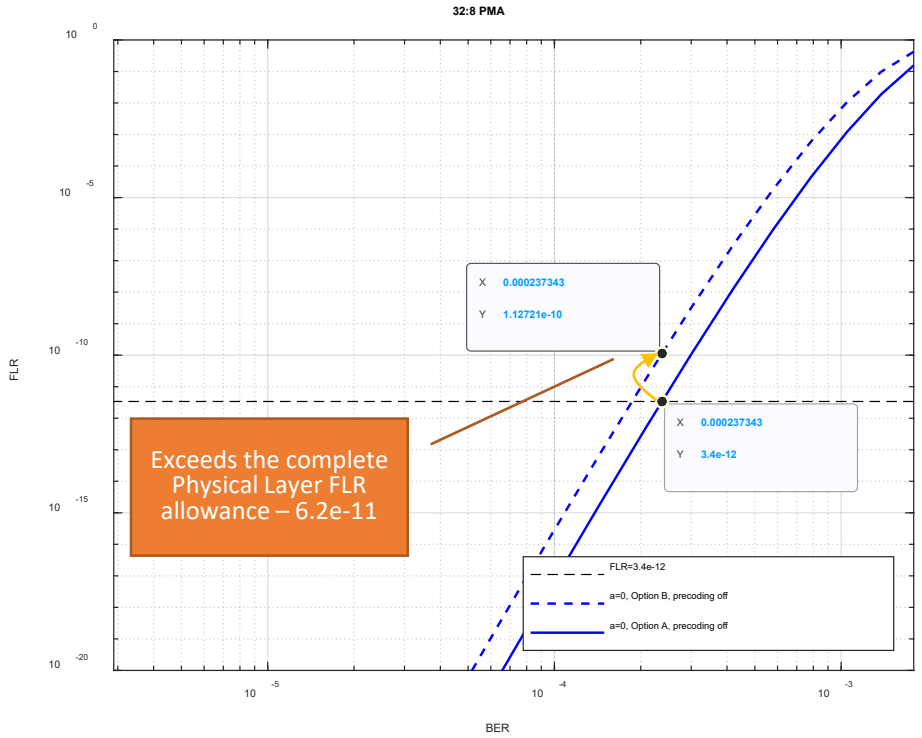
# Implications

- An uncorrectable codeword invalidates the 3 other codewords interleaved with it
  - Causing loss of up to 33 MAC frames
  - x2 FLR compared to 400G/200G PHYs (2-CW interleaving) – already discussed; insignificant
  - Data loss event frequency is the same
- Having only LSBs in a codeword creates a 33% higher BER for that codeword compared to balanced MSB/LSB (as in 400G/200G PHYs)
  - 0.15 dB (electrical) SNR penalty
  - Frequency of uncorrectable codewords (data loss event) increase x34 (due to FEC coding gain)
  - Combined FLR is x68 compared to 400G/200G PHYs!
  - "Once per month" in 2x400G becomes "more than daily" in 1x800G.
  - "Once per year" in 2x400G becomes "more than weekly" in 1x800G.
- A device that uses "option B" on its transmit path creates a higher FLR and more data loss events **for its link partner.**
  - This can hinder troubleshooting…
- This happens:
  - Regardless of where the errors happen (any optical/AUI receiver whose transmitter uses the "option B" order)
  - Regardless of receiver structure.

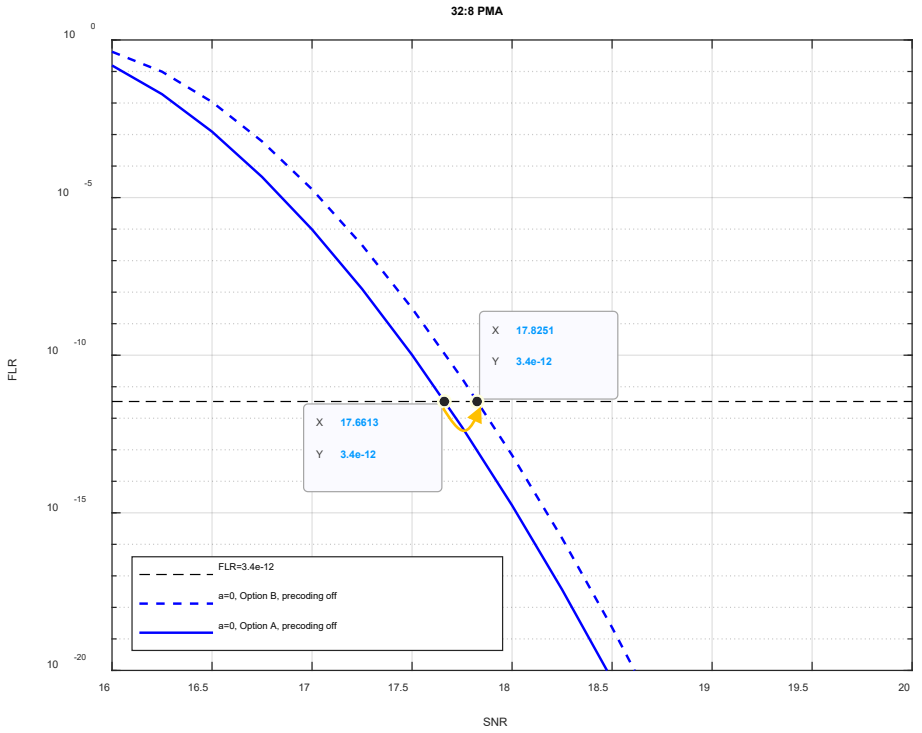
  Does not happen if precoding is used (but currently precoding is defined only for electrical links).

# FLR effect of option B with low error correlation (e.g. optical PMD)

**The specified PMD BER (2.4e-4) creates a x34 higher FLR than with option A**




**FEC coding gain is reduced by 0.15 dB**



# Addressing concerns that have been raised Part 1 – 32:8 and 8:32 PMAs

Adding more restrictions than we had in the adopted baseline (which was in large consensus) may make products in development non-compliant

- A product that uses option B muxing will cause poor link performance in its partner; this can be identified easily (now that the problem is known)
  - Even if option B is not forbidden by the standard, it will be a known flaw.
  - Fixing the problem should be easy (a very minor and localized design change).
- A device that is made “non-compliant” by the suggested change would still be compatible with compliant devices (albeit with higher FLR)
  - Having pre-standard products not fully compliant but compatible is considered acceptable.
- Same goes for 8:32 PMAs (which exist only in modules/retimers with PHY 800GXS).

Test equipment that already exists may not be able to meet the restrictions

- Test equipment typically uses FPGAs, possibly with time-interleaving of two 50G/lane SerDes
- Even if traffic is generated with 50G SerDes and time-interleaved with uncontrolled skew, the muxing selection is controllable.
- The proposed change does not impose any skew restrictions on two “half rate” lanes.
- Test equipment for electrical or optical parameters uses specific test patterns and can’t have unrestricted skew (see next slide).

# Addressing concerns that have been raised Part 2 – 8:8 PMAs

8:8 PMAs (in modules and retimers) were explicitly given freedom to re-order the PCSLs on each physical lane – are we breaking existing products?

- 100G/lane retimers are typically “two SerDes back-to-back” on each channel
  - Changing the muxing order or changing the skew within one physical lane is unnatural in this architecture.
  - SerDes bus width is typically even; **there is no physical serial bit stream that could be shifted to change the PAM4 bit pairing**.
- Any existing device practically must meet the proposed restriction already:
  - Keeping the same PCSLs in the input and output physical lanes is already required.
  - On a given physical lane, doing anything other than relaying of the input PAM4 symbol stream to the output would turn a test pattern (PRBS31Q, PRBS13Q, SSPRQ, or square wave) into something else (not just a delay).
  - This could make measurements of electrical/optical specifications (TDECQ, EH/VEC, stressed input tests, etc.) impracticable.
  - A device that behaves differently is also not testable with PRBS31Q (per-lane test) using external pattern generators and checkers. If tested in external loopback with a host, it will not return the same pattern, so BER can’t be checked at the PMA.

# Suggested remedy – part 1

### 173.4.2.1 32:8 PMA bit-level multiplexing

*Change the second list item as shown:*

– The multiplexing function has an additional constraint that each of the 8 output lanes contain two ~~unique~~ PCSLs from PMA client lanes $i$ = 0 to 15 ~~and~~ followed by two ~~unique~~ PCSLs from PMA client lanes $i$ = 16 to 31

### 173.4.2.2 8:32 PMA bit-level multiplexing

*Change the second list item as shown:*

– The multiplexing function has an additional constraint that each of the 8 output lanes contain two ~~unique~~ PCSLs from service interface lanes i = 0 to 15 ~~and~~ followed by two ~~unique~~ PCSLs from service interface lanes i = 16 to 31.

This wording allows option A (with possible bit shift), and forbids option B

# Suggested remedy – part 2

### 173.4.2.3 8:8 PMA bit-level multiplexing

*Change the second list item as shown:*

– The 4 PCSLs received on ~~any~~ an input lane shall be mapped to ~~the same~~ an output lane such that the Gray-coded PAM4 symbol sequence on the output lane is identical to the Gray-coded PAM4 symbol sequence on the input lane (see 173.4.7.1). ~~The order of PCSLs from an input lane does not have to be maintained on the output lane.~~

This prevents a module or retimer from converting option A into option B

Note: "Gray coded PAM4 symbol" definition in 120.5.7.1 (referenced by 173.4.7.1) creates a 1:1 correspondence between symbols and pairs of bits. The wording above requires bit pairs to have the same order, which may be more restrictive than necessary. If flexibility in bit pair ordering is desired, the wording can be changed to:

such that the Gray-coded PAM4 symbol sequence on the output lane is identical to the Gray-coded PAM4 symbol sequence on the input lane, except for possible swapping of each bit pair (see 173.4.7.1)

# Summary

- Bad muxing choice will create significant and noticeable performance degradation in the link partner.
- The additional constraints suggested in the comment solve the issue.
- The concerns that were raised have been addressed:
  - 8:8 constraint is unlikely to affect any existing retimer implementation.
  - Test equipment is unlikely to be impacted by the suggested change.
  - The PCS receive function handles any PCSL order, so the constraints preserve compatibility of devices that do not meet them.
- Let’s do the right thing for the market.

# Backup

# 2/3 of random errors occur in the LSB

**Burst error model 2**

The second aspect of this table is that of the six possibilities giving bits in error, two have errors in the first bit while four have errors in the second bit.

| Correct level | Received level | | Error pattern | |
|---|---|---|---|---|
| | One up | One down | One up | One down |
| 3 | 3 | 2 | ✓, ✓ | ✓, × |
| 2 | 3 | 1 | ✓, × | ×, ✓ |
| 1 | 2 | 0 | ×, ✓ | ✓, × |
| 0 | 1 | 0 | ✓, × | ✓, ✓ |

The analysis in the remainder of this contribution therefore assumes that if a given symbol is in error, the probability of a bit error in the first bit is 1/3 and in the second bit is 2/3.

11

Source: anslow_3ck_adhoc_01_072518

This means that, with option B:

- The two codewords that get the MSBs (A/B) have 2/3 of the average BER
- The two codewords that get the LSBs (C/D) have 4/3 of the average BER
- Uncorrectable errors occur more often in C and D
- Any uncorrectable error corrupts all four codewords

Note: if precoding is used, the decoding operation spreads errors equally across MSB and LSB, so this only applies to the non-precoded case