Proposal for BER budget allocation for AUIs in Type 1 and Type 2 PHYs

Adee Ran, Cisco

Acknowledgements

Supporters

- Brian Welch, Cisco
- Zvi Rechtman, NVIDIA
- Itamar Levin, Intel
- Piers Dawe, NVIDIA

Contributors

- Kent Lusted, Intel
- Mark Nowell, Cisco
- Adam Healey, Broadcom
- Piers Dawe, NVIDIA

Background

- Achievable BER/DER for C2M channels at 106.25 GBd has been a discussion topic for a long time
- Recent work:
  - li_3dj_02a_2303 presented COM results with various channels and COM parameter sets, including DER_0 of 1e-5, 5e-5, 1e-4
  - Straw polls 8 & 9 in the March 2023 meeting (motions_3dfdj_2303) showed support for either 1e-5 or 5e-5 for “medium BER” AUIs and 5e-5 for “high BER” AUIs
  - brown_3dj_elec_01_230420 analyzed latency for combination of PHY types, AUI BER allocations, inner FEC interleaving
  - ran_3dj_elec_01_230420 presented the effect of various AUI BER allocations on the output BER required from the inner FEC
  - Straw polls 1 & 2 in the April 23 ad hoc (straw_polls_3df_elec_adhoc_230420) showed split support for 1e-5/1e-5, 5e-5/5e-5, and 8e-5/2e-5 allocations for C2C and C2M
- Additional presentations in this meeting address related issues

Problem statement

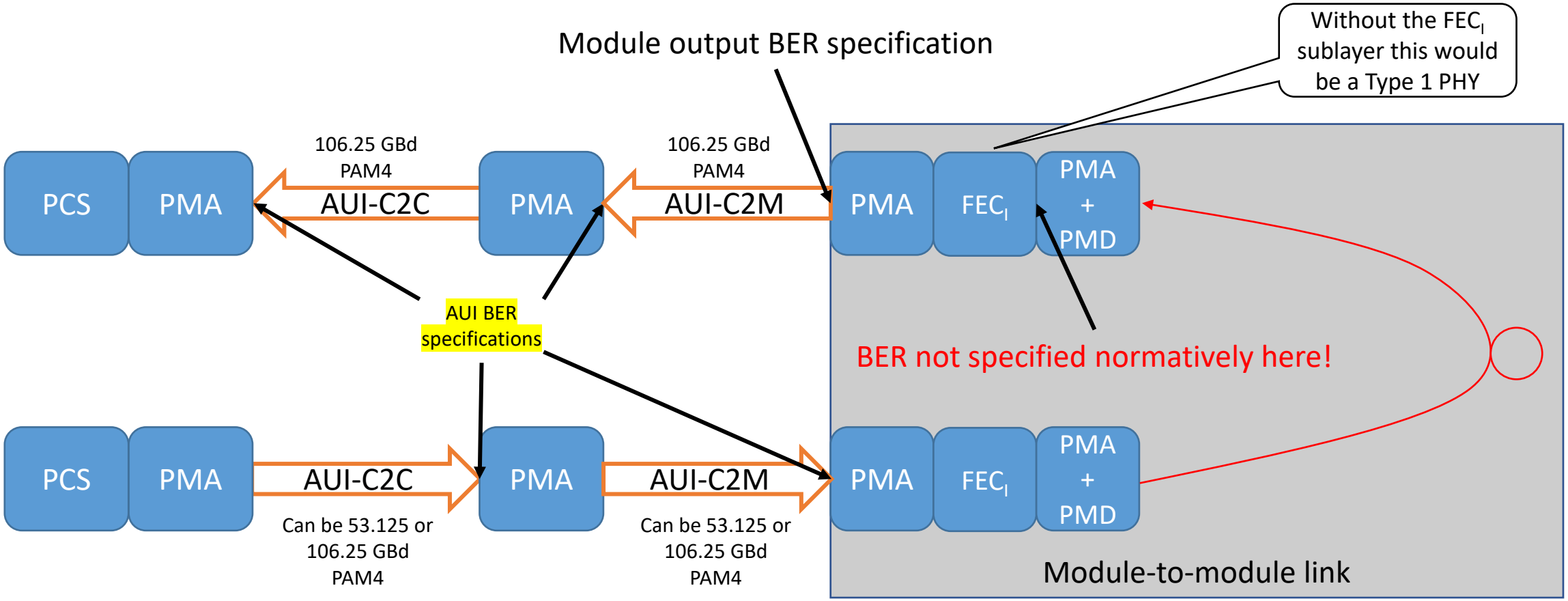
- The performance specification for 106.25 GBd AUIs determines the required performance for all modules (PMDs and associated FEC_l) in Type 1 and Type 2 PHYs that can include these AUIs
  - The single interaction between electrical specs and optical specs…
  - Arguing over it prevents both tracks from doing other work!
- Locking down the BER should be the top priority of our AUI efforts
  - COM strongly depends on DER_0 (which depends on the BER allocation for the AUI)
  - We have limited data on other electrical parameters (including COM parameters)
  - Numerous channel samples are available, and selecting which ones should be supported (to cover real systems) is still open
  - Too many moving parts…
- **We should make decisions on error budgeting and move forward**

This presentation

- Recaps the effect of BER allocations
- Suggests possible scenarios
- Proposes a “Random BER” allocation for 106.25 GBd AUIs

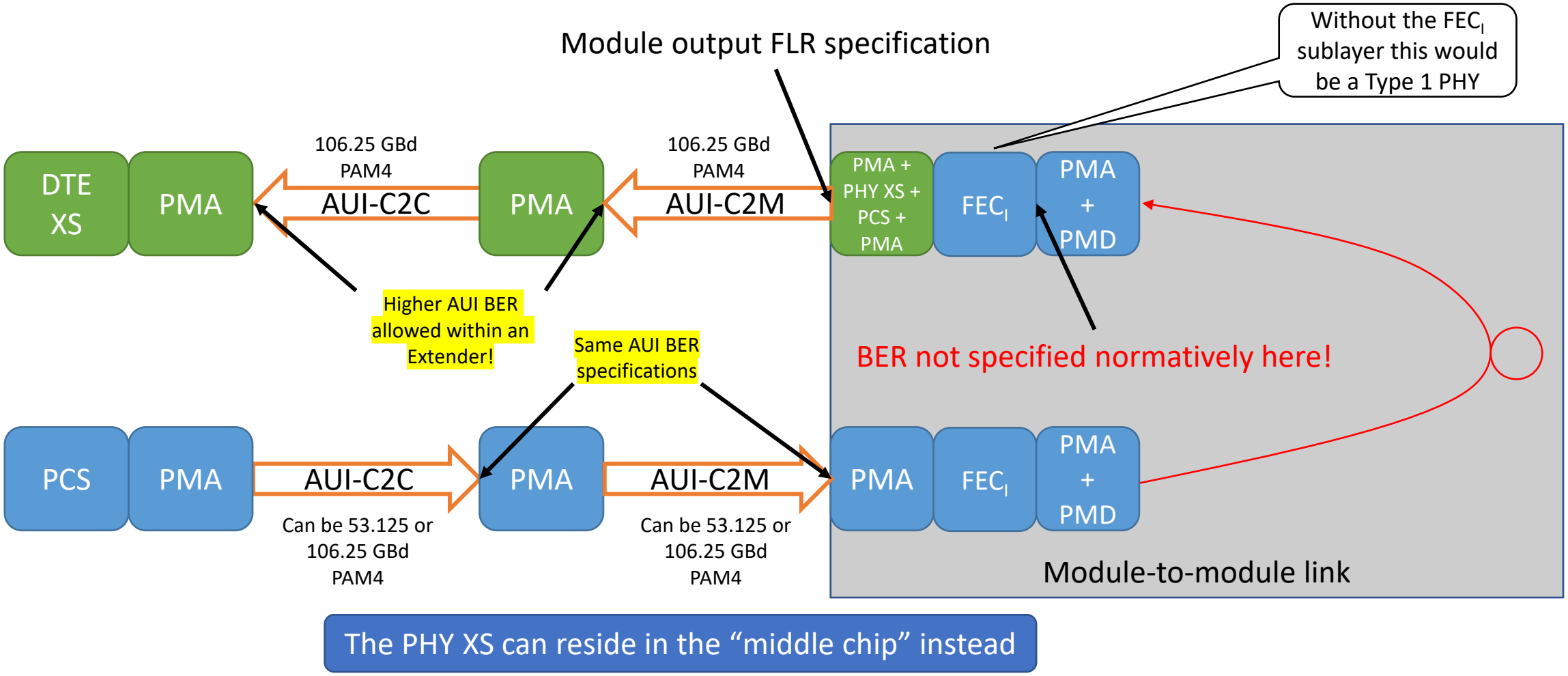
Link diagram – Type 2 PHY without Extender

(only one direction shown)



Link diagram – Type 2 PHY with Extender

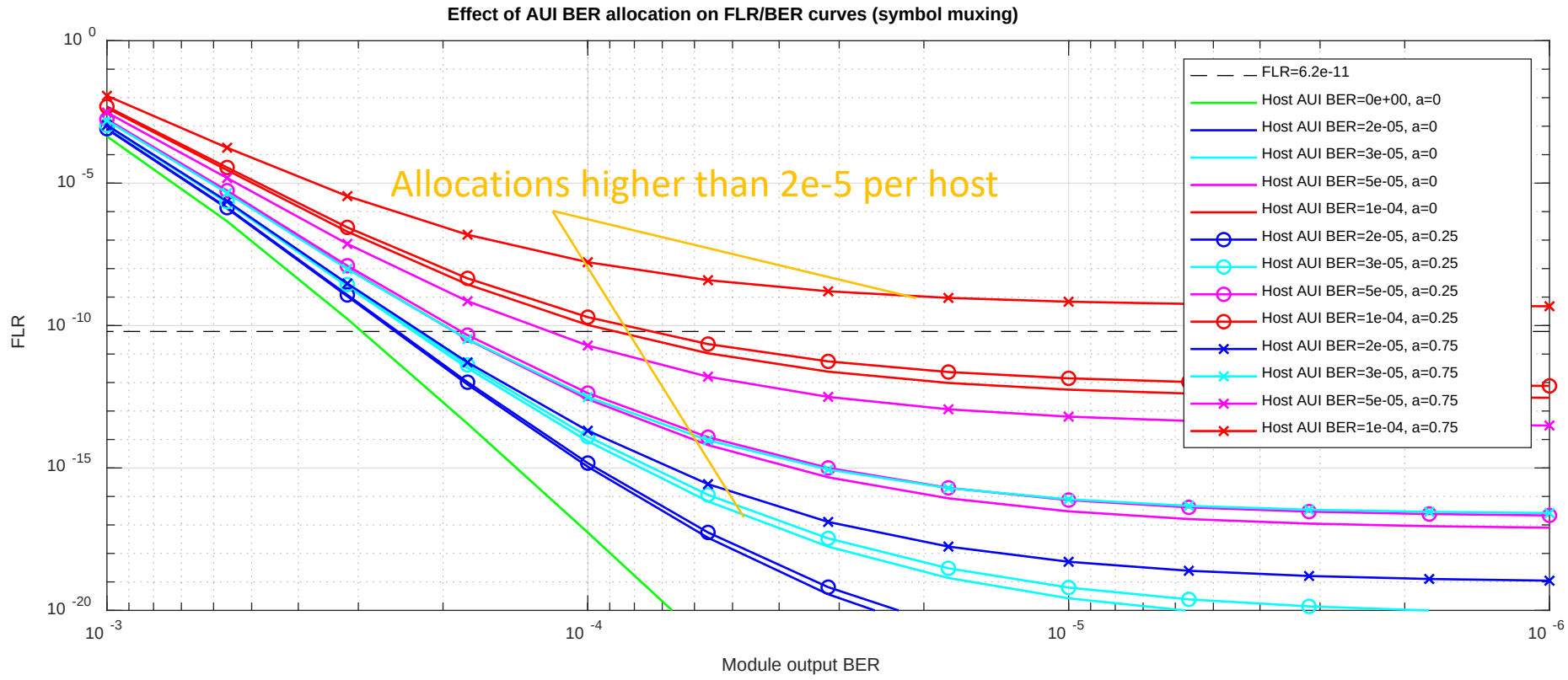
(only one direction shown)



The PHY XS can reside in the “middle chip” instead

Effect on optical module requirement

This is similar to data presented in ran_3dj_elec_01_230420 (but with clarified labels). Module errors assumed to be uncorrelated.



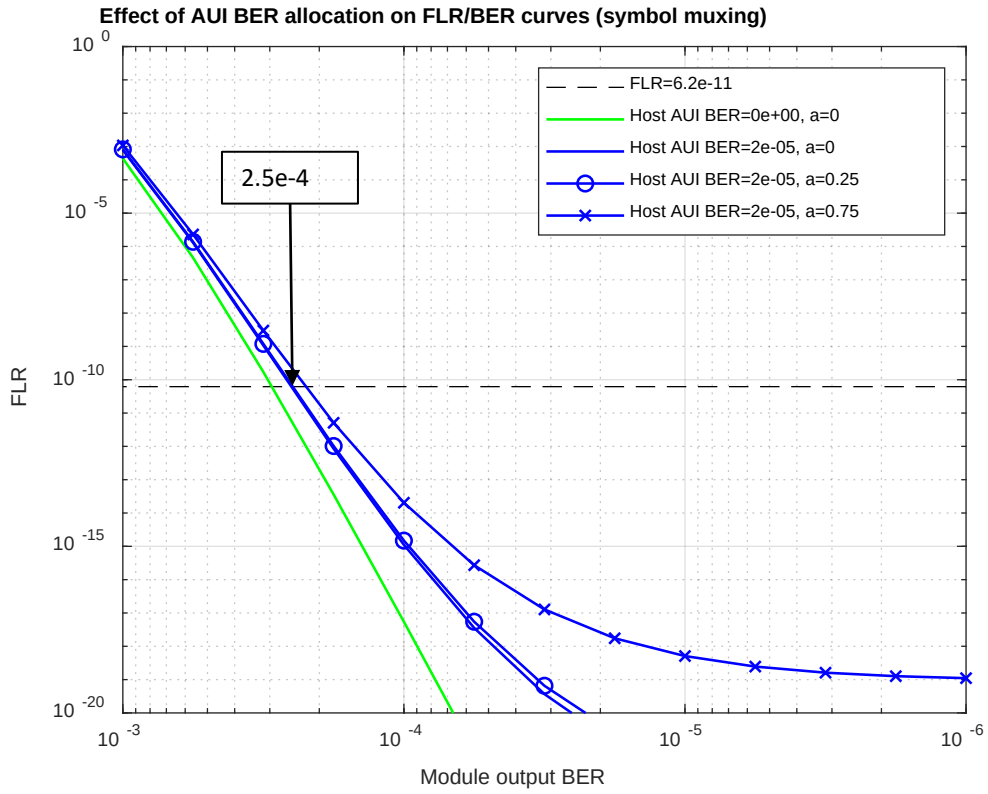
Per-host AUI **random BER** (see backup slide): 0 2e-5 3e-5 5e-5 1e-4

Implications

- Allocating more than a random BER of 2e-5 per host:
  - Creates a high FLR floor (~1e-17 for 5e-5, >1e-13 for 1e-4)
  - Creates noticeable penalty for the module-to-module link (lower BER for combined PMD + FEC_l)
  - Correlated errors on the AUIs raise the FLR floor and increase module-to-module link penalty
- If module output errors are correlated (e.g., FEC_l with incomplete interleaving), these effects will become even worse
- If the AUI BER is indeed so high – it would likely drive operation in "Extender mode" (segmenting the RS-FEC) in many applications that require low FLR
- Existing PMDs (100G/lane) assume no more than 2e-5 per host!
  - Higher BER will force "Extender mode" for interoperability

Implications (cont.)

- Allocating random BER of 2e-5 (total per host):
  - Creates low FLR floor (1e-19 with a=0.75, ~1e-25 for a=0.25)
  - Creates very small penalty for the module; enables output BER of **2.5e-4** (see backup slide)
  - Enables existing PMDs to work without forcing an MII Extender
- Due to symbol muxing, error propagation has a small effect
  - Even a=0.75 requires just a slight reduction in random BER on the AUIs



Can we meet 2e-5 per host?

- Currently it may seem challenging (with available channel data)
- But history shows performance exceeded initial estimates many times
  - Both channels and SerDes performance improve over time
  - Examples:
    - 802.3bj defined 100GBASE-KP4 (PAM4) based on initial channel data that showed low bandwidth – but lost to 100GBASE-KR4 (NRZ) – deployed channels were much better
    - 100GBASE-KR4 required FEC in all links but later 25GBASE-R added no-FEC mode (with reach beyond the expectation)
    - 802.3ck specified Copper cable reaches up to 2 m – but operation over 4 m is feasible today
- And new data rates always make older data rates look easy…
- Even if 2e-5 looks difficult today – we should not assume it's impossible!
  - As an easy fallback, modules can be configured to terminate the RS-FEC (become XS)

Possible options for a 2e-5 host allocation

A. Within a PHY, allocate random BER of **1e-5 for both C2C and C2M**
   - In COM, use $DER_0 = \frac{4}{3} \cdot 1e{-}5 = 1.33e{-}5$ (see backup slide)

B. Within a PHY, allocate random BER of **2e-5 for C2M** and **0 for C2C**
   - In COM, use $DER_0 = \frac{4}{3} \cdot 2e{-}5 = 2.67e{-}5$ for C2M (see backup slide)
   - C2C can only be used within an Extender (which may also span the C2M)

In either option, within an Extender, C2C and C2M can have much higher allocations
   - Specification for this case should be based on a small portion of the FLR allocated to the Extender

How systems can be built

- A network port that consists of a MAC/PCS ASIC + AUI-C2M + Module:
  - If the AUI-C2M meets the allocated BER (per option A or option B) – there is no need to terminate the RS-FEC; latency and power are minimized
  - If the AUI-C2M has higher BER (up to ~1e-4) – the module can be configured to terminate the RS-FEC, and the AUI-C2M becomes part of the MII extender; AUI reach is higher, but latency and power are higher too
  - Can be configured locally, transparent to the link partner
- A network port that has AUI-C2C in addition to the above:
  - In option A – can operate without terminating the RS-FEC, minimizing latency
  - In option B – the AUI-C2C can only be part of an MII extender (possibly including the AUI-C2M too)

Proposal

- For Type 1 and Type 2 PHYs, allocate:
  - BER of **2.4e-4*** to the module-to-module link, assuming uncorrelated errors
    - To be specified at the module's electrical output (e.g., after FEC_l decoding)
  - Total random BER of **2e-5** for the host AUI links, with an assumption of correlated errors with a=0.25
    - Choice of option A or B, to be decided later
- Specifications will be in terms of FLR or similar
  - Based on the random BER
  - Details to be defined later
- AUIs within a PHY extender will have different specifications
  - Not included in this proposal; to be defined later

* 2.4e-4 is required to support 100G/lane AUIs with bit muxing – see backup slide

Proposed straw poll

For AUIs within a Type 1 or Type 2 PHY (not within an MII extender), I would support specifications corresponding to a random BER of:

A. 2.4e-4 for the module-to-module link; 2e-5 for AUIs within each host
B. Lower value for module-to-module link; higher values for AUIs
C. Higher value for module-to-module link; lower values for AUIs
D. Need more information

Backup

“Random BER”?

- We often use the Gilbert error propagation model
  - Random BER is the ratio of “random” bit errors (not associated with bursts) to the number of bits; error propagation probability is a parameter, a
  - In this model, **errors occur in contiguous bursts of PAM4 symbols**, with mean burst length of $\frac{1}{1-a}$; mostly each symbol error causes one bit error
  - This model, the “Random BER” is $\frac{3}{4}\cdot$DER *
- A model for MLSE error propagation was recently described in shakiba_3dj_elec_02_230420
- These are partial models…
  - Actual error correlation depends on device implementation (and possibly link partner too)
  - **Error correlation may not be just due to contiguous bursts** – effect on RS-FEC can be different
- To predict the effect of correlated errors on FEC performance, we need the distribution of symbol errors per codeword…
  - This depends on error correlation in the implemented system, as well as the FEC interleaving and AUI muxing schemes, and whether precoding is used
  - **Average BER (product of random BER by mean burst length) is not useful for this purpose**
- Metrics other than average BER may be preferable

* Previously $\frac{1}{2}\cdot$DER was used, until Bill Kirkland noticed this is incorrect (following Jonathan King’s prior work). See kirkland_3dj_elec_01_230406

BER allocation for module-to-module link

- With symbol muxing and 4-way interleaving, allocation of 2e-5 for each host enables a module-to-module BER of 2.5e-4
  - In previous 100G/lane PMDs we used 2.4e-4 – why is it different?
  - Because symbol muxing is more tolerant to burst errors, compared to bit muxing; the same AUI BER has a lower impact on FLR (even with 4-way RS-FEC interleaving)
  - The AUI BER allocation is small part of the total error budget, so the effect on the remainder of the budget is small
- The good news – 200G/lane AUIs can be compatible with existing 100G/lane PMDs
- The bad news – 200G/lane modules still need to work with 100G/lane AUIs, which use bit muxing
  - In this case, the BER allocation for the module-to-module link can't be increased from 2.4e-4
- The proposal is therefore to use 2.4e-4 in all cases.