Muxing and Interleaving

IEEE P802.3bs 400 Gb/s Ethernet Task Force

March 2015, Berlin

Oded Wertheim – Mellanox Technologies

FEC & PMA Muxing

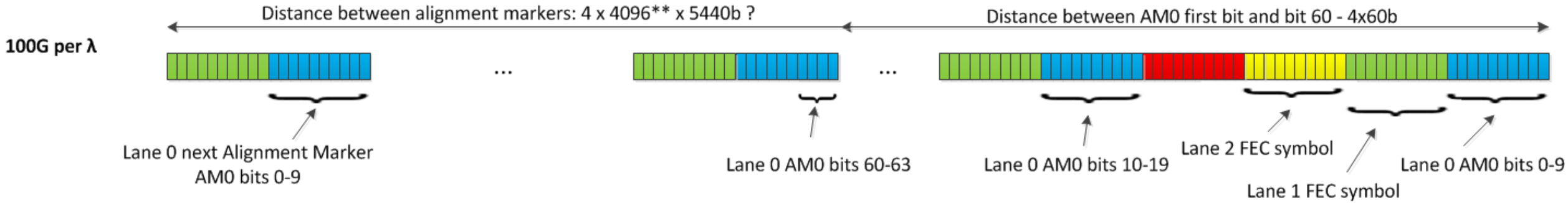
- The challenge
  - 400G links may include segments of different width: 16 lanes ⇔ 8 lanes ⇔ 4 lanes

- Options considered for PMA muxing*:
  - Bit muxing
    - Simple
    - Not aware to protocol
    - Results in interleaving between FEC symbols which may reduce the correction gain under correlated errors
      - Addressed by FOM bit muxing / FOM Pre-interleaving bit muxing
  - FEC symbol muxing
    - Protocol specific
      - Need to identify symbol boundaries
    - Maintains the FEC correction gain under correlated errors
  - A different scheme may be used when baud rate changes / remains the same

* slavick_3bs_01a_0115

RS-FEC 10b Symbol Multiplexing Complexity

- Option – Identify FEC codeword boundaries using alignment markers
  - Has to be performed in high speed (up to 100G per lambda)
  - Requires alignment lock / lose logic – under high BER
    - Clause 91 FEC alignment lose is based on FEC uncorrectable codewords so the alignment markers are protected.
  - Requires 4 x 10b symbol store and forward
  - **Conclusion: complex → high cost / high power modules**



- Option - Change the PMA encoded data to identify FEC symbol boundaries
  - 2 different alignment mechanisms in the protocol stack
  - A PMA alignment symbols will require bandwidth – higher signaling rate
  - Requires new alignment lock and alignment lose state machine
  - Requires 4 x 10b symbol store and forward
  - **Conclusion: complex → high cost / high power modules**

** based on 802.3bj – TBD for 400G

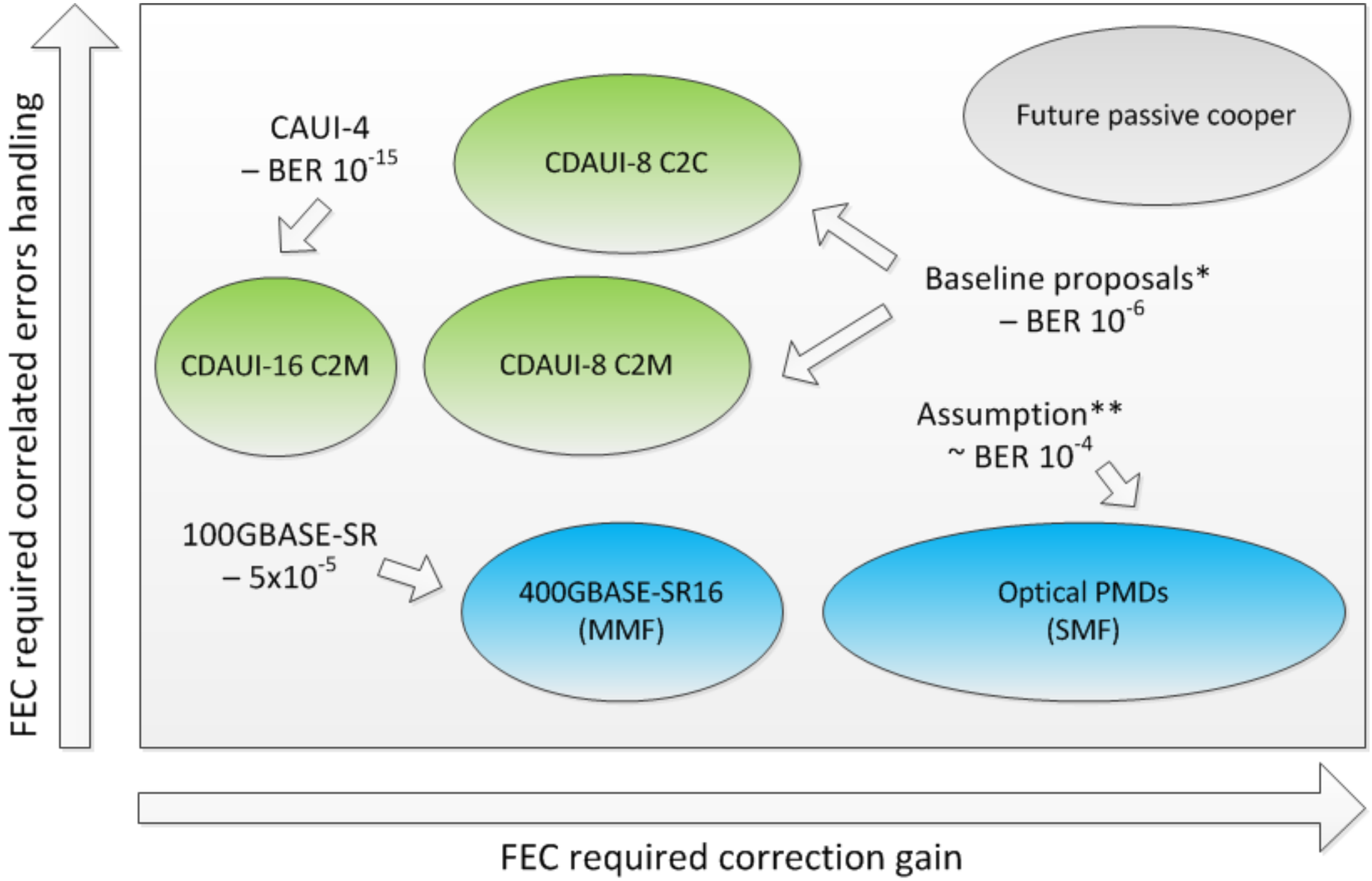
Protocol Unaware Modules and Protocol Specific Modules

- Protocol Unaware Modules
  - Enables to use the same optical modules for other protocols:
    - Forward compatibility to future IEEE protocols
    - Breakout – enables to use CDAUI-16 modules for 4x100G / 16x25G
    - Other standards, for example: ITU-T, Fibre Channel, Infiniband

  → **Lower cost → Economic feasibility and broad market potential**

- Lessons from the 100Gb/s generation – Did we do something right ? Yes
  - 100Gb/s (and 40Gb/s) products benefit from having modules that are not protocol aware
  - 25G over single lane (802.3by)
    - Can reuse 100G modules
    - Uses different alignment markers / codeword marker format and size
  - MLG3.0
    - Can reuse 100G eco-system
    - Uses different alignment markers content

FEC Considerations for 400G links



* li_3bs_01a_0115
* brown_3bs_01a_0115
* palkert_3bs_02a_0115
** gustlin_3bs_04_0115

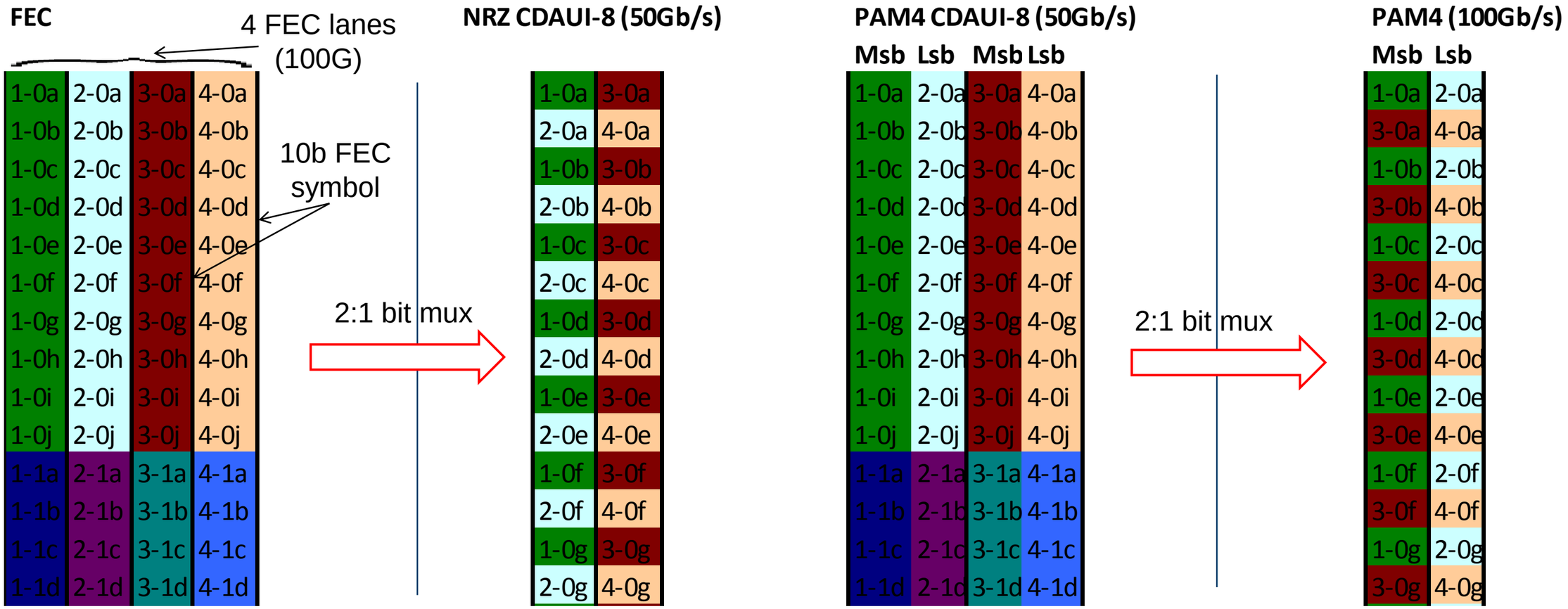
FEC Considerations for 400G links

- Electrical interfaces
  - 50Gb/s CDAUI-8 C2M/C2C , 25Gb/s CDAUI-16
  - Error model – Expect DFE in the CDAUI-8 C2C receivers – error bursts
  - Proposed required gain: $10^{-6} \rightarrow 10^{-13}$
    - We may strengthen the requirement to a lower BER - $10^{-7}/10^{-8}$ if there is a value.
  - **Conclusions:**
    - **400G FEC should address correlated errors on the CDAUI-8 interface**
    - **No need for a high gain as in the optical interfaces**

- Optical interfaces
  - 25G per λ MMF, 50G / 100G per λ SMF
  - Error Model – error bursts were not identified as an issue on the optical PMDs.
  - Required gain: ~ $10^{-4} \rightarrow 10^{-13}$
  - **Conclusions:**
    - **400G FEC does not need to be optimized for correlated errors on the optical link**

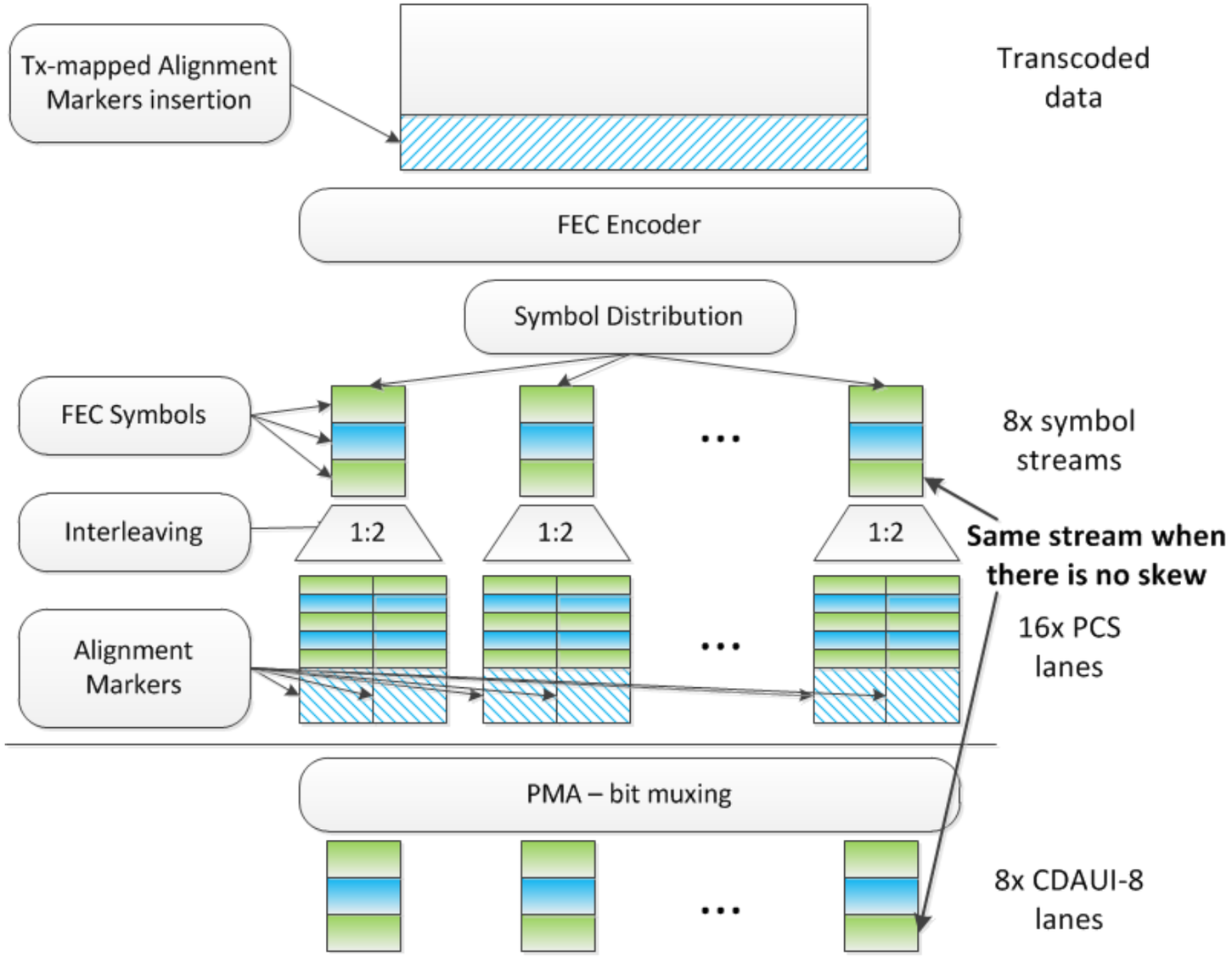
Blind Bit Multiplexing Example

- 100G (4 PCS lanes) bit multiplexing example (1/4 of a 400G link)



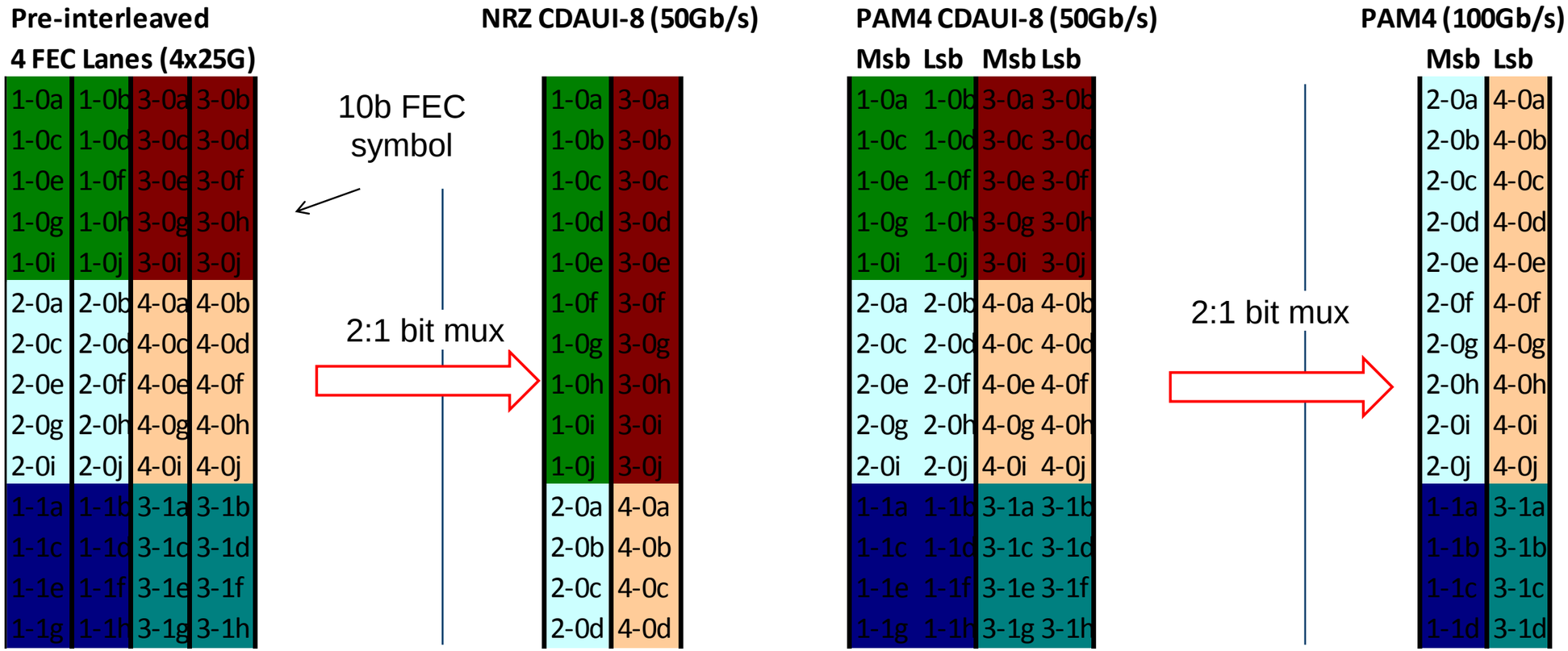
- Each column indicates a lane – (PAM4 lanes are divided to Msb/Lsb)
- Each color indicates a FEC 10b symbol
- Each “x-yz” indicates a bit: x – FEC lane number / y – FEC symbol index / z – bit with a symbol

50G (2 lanes) FEC Pre-interleaving



- Simple interleaving in the FEC sublayer between 2 x 25Gb/s lanes
  - Each FEC symbol is bit-multiplexed to 2 PCS lanes
  - Tx-mapped Alignment markers are inserted such that alignment markers will appear on each PCS lane after the symbol distribution and interleaving. (similarly to 802.3bj CL91 AM mapping)
- When there's no CDAUI-16 instance (no skew between the FEC sublayer and the CDAUI-8 PMA) the PMA bit-mux reconstructs the 10b FEC symbols
- When there is a CDAUI-16 instance (results in skew between the interleaved lanes). The KP4 RS-FEC correction gain should be sufficient for CDAUI-8 C2M/C2C I/Fs. (may require a stronger BER requirement on the C2M/C2C I/Fs ~ 10^{-7} / 10^{-8})

50G (2 lanes) FEC Pre-interleaving & Blind Bit Multiplexing



- The example shows 4 PCS lanes (100G) multiplexing to CDAUI-8 and to 100G per λ
- 100G per λ => CDAUI-8 de-muxing maintains each CDAUI-8 lane bit ordering
  - May reorder the CDAUI-8 lanes

Summary

- Protocol unaware modules are highly desired
  - Enables re-use for multiple protocols (future protocols, breakout, other standards)
  - Removes the need for complex state machine in the modules
  - Lower cost, lower power

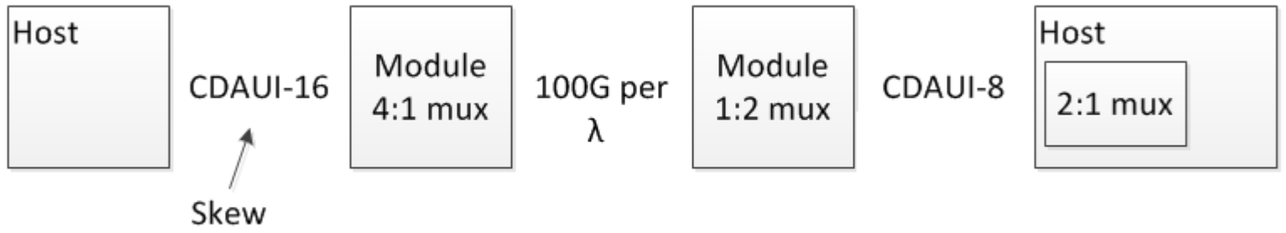
- Different FEC consideration for different link segments
  - Electrical interfaces
    - Correlated errors
    - No need for a high gain as in the optical interfaces
  - Optical interfaces
    - Need not to be optimized for correlated errors
    - Need for a higher correction gain

- A FEC pre-interleaving alternative with PMA bit muxing is proposed:
  - Enables to maintain the correction gain under error bursts for future 8x50G high loss passive interfaces
  - In addition to existing valid bit muxing proposals such as FOM bit muxing.

- **PMA Bit multiplexing is the right decision**

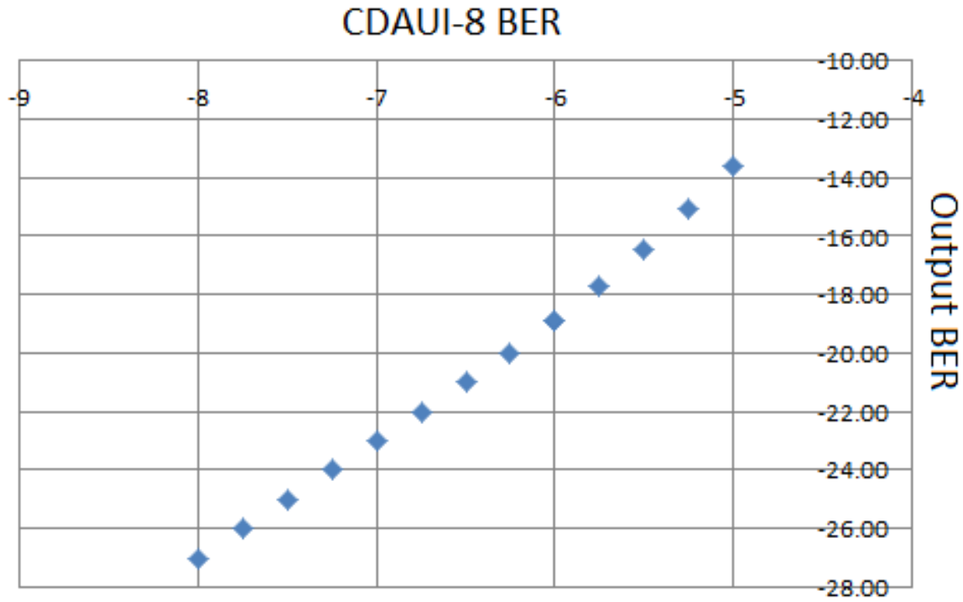
Thank You

50G (2 lanes) FEC Pre-interleaving BER Evaluation Under Skew



- The difficult use case is when the link has a 16 wide segment (CDAUI-16) and a CDAUI-8 I/F, resulting in skew between the interleaved lanes on the CDAUI-8
- From a burst errors analysis, the worst case is PAM4 CDAUI-8 13 bit skew between the lanes
  - Assuming 2 level errors

13 bit skew - 2 level PAM4					
	Symbol errors				
Burst length	**1**	**2**	**3**	**4**	**5**
2b	100%	60%	0%	0%	0%
3b	100%	85%	20%	0%	0%
4b	100%	95%	48%	3%	0%
5b	100%	99%	71%	14%	0%
6b	100%	100%	88%	34%	0%
7b	100%	100%	96%	59%	6%
8b	100%	100%	99%	79%	22%



*An approximation of the output BER on a single CDAUI-8 link segment. alpha = 0.5