400GbE PCS Options

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

July 2014 San Diego

Mark Gustlin – Xilinx
Jerry Pepper - Ixia
Andre Szczepanek – Inphi
Steve Trowbridge - ALU
Tongtong Wang – Huawei

Introduction

- The following slides explore 400GbE PCS options for use with various PMDs

References

400GbE Architecture:

http://www.ieee802.org/3/bs/public/14_05/gustlin_3bs_02_0514.pdf

400G PCS options:

http://www.ieee802.org/3/bs/public/14_05/wang_x_3bs_01_0514.pdf

http://www.ieee802.org/3/bs/public/14_05/wang_t_3bs_01_0514.pdf

http://www.ieee802.org/3/bs/public/14_05/trowbridge_3bs_02_0514.pdf

http://www.ieee802.org/3/400GSG/public/14_01/wang_400_01a_0114.pdf

http://www.ieee802.org/3/400GSG/public/13_11/gustlin_400_02a_1113.pdf

http://www.ieee802.org/3/400GSG/public/13_11/song_400_01_1113.pdf

http://www.ieee802.org/3/400GSG/public/13_11/wang_400_01_1113.pdf

http://www.ieee802.org/3/400GSG/public/13_09/wang_400_01_0913.pdf

http://www.ieee802.org/3/400GSG/public/13_09/begin_400_01_0913.pdf

http://www.ieee802.org/3/400GSG/public/13_09/ghiasi_400_01_0913.pdf

http://www.ieee802.org/3/400GSG/public/13_09/song_400_01_0913.pdf

http://www.ieee802.org/3/400GSG/public/13_09/wang_z_400_01_0913.pdf

http://www.ieee802.org/3/400GSG/public/13_07/gustlin_400_02_0713.pdf

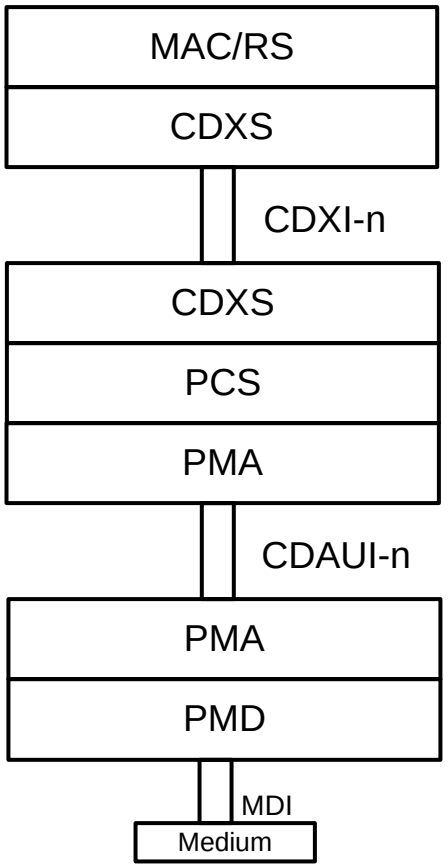
http://www.ieee802.org/3/400GSG/public/13_07/wang_400_01_0713.pdf

http://www.ieee802.org/3/400GSG/public/13_07/ghiasi_400_01_0713.pdf

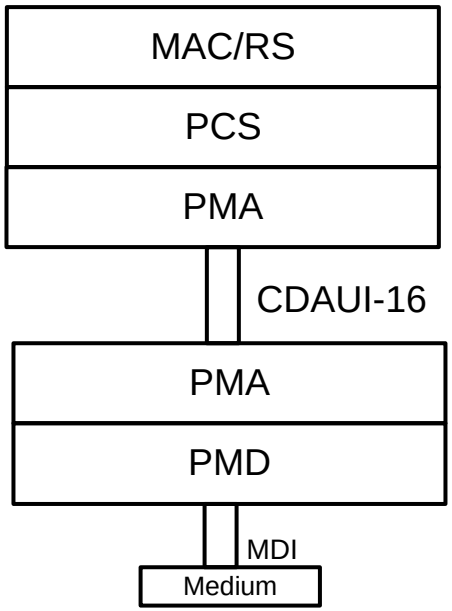
http://www.ieee802.org/3/400GSG/public/13_05/ghiasi_400_01a_0513.pdf

400GbE Architecture for SR16 PMDs

Generic Architecture



Possible PMD Specific Arch



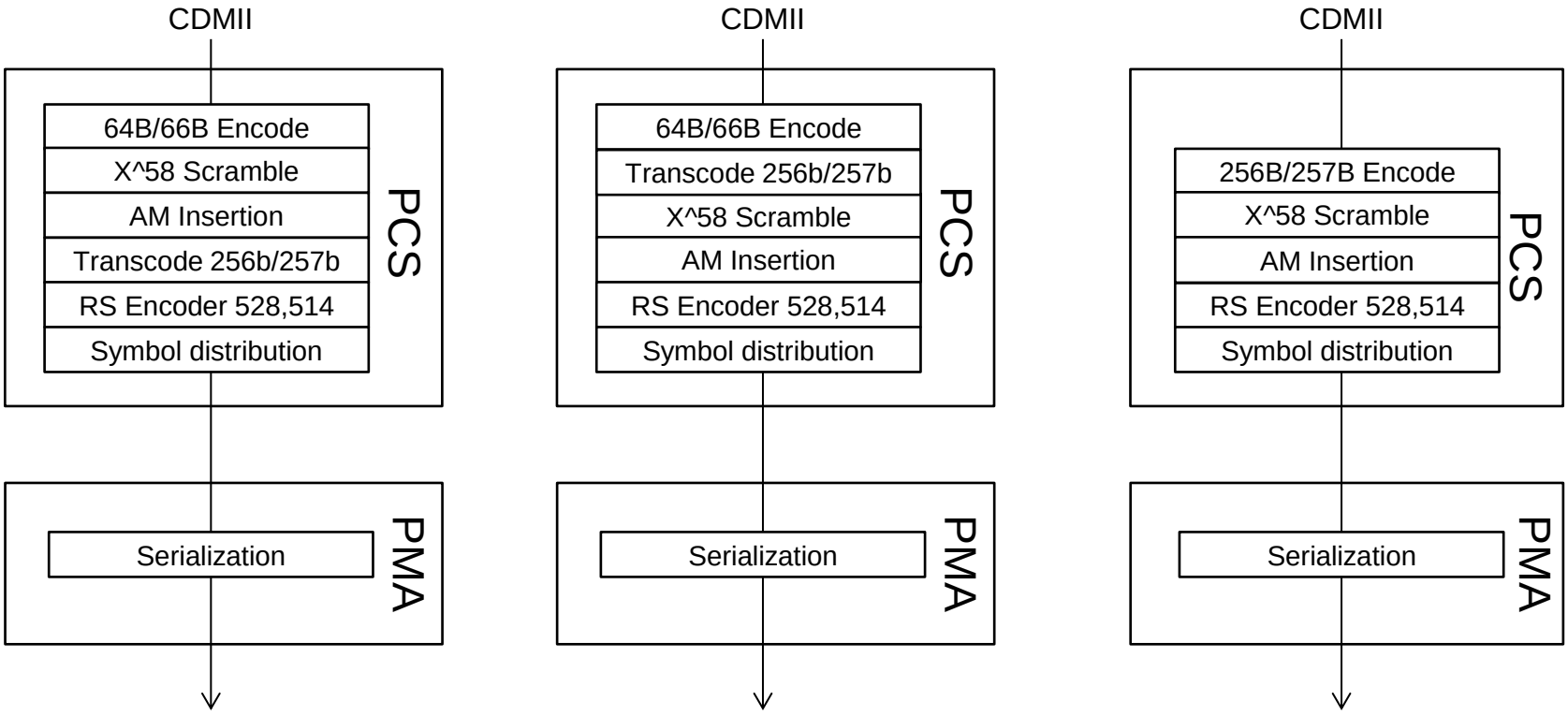
No need for an extender sublayer if the electrical and optical interfaces require the same coding

PCS Re-use

- When defining the a PCS, we have to decide between different goals:
  - High re-use from 100GbE, allows for more compact 4x100G/1x400G designs, but might limit our flexibility
  - Lower re-use from 100GbE, could allow us to leverage industry learning for a more flexible interface, but leads to less compact multi-speed implementations
- Possible areas of re-use
  - 64B/66B and/or 256B/257B encoding
  - 256B/257B transcoding
  - X^58 scrambling
  - RS-FEC (KR4 and KP4 can be considered)
  - Alignment Markers
  - BIP
  - Bit Muxing

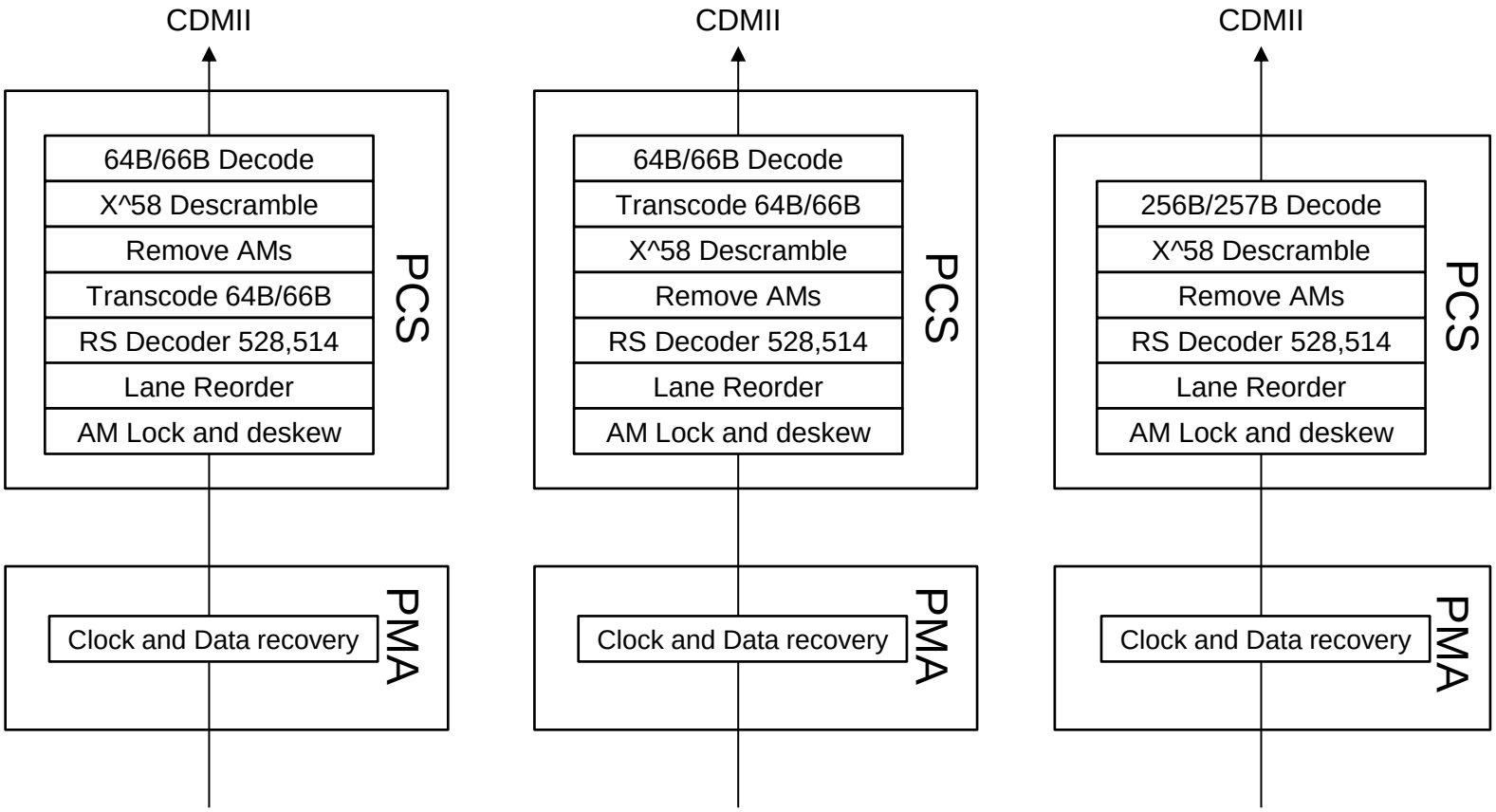
Data Flow - TX

- Possible protocol flows
- Should we direct encode to 256B/257B encoding?
- OTN reference point could be intra sublayer interface, just after the 64B/66B encoder (or 256B/257B)?



Data Flow - RX

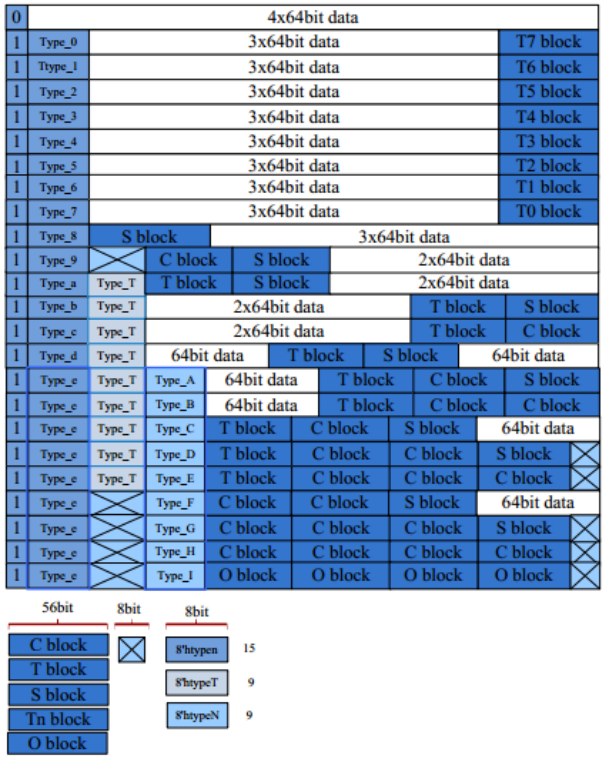
- Possible protocol flows

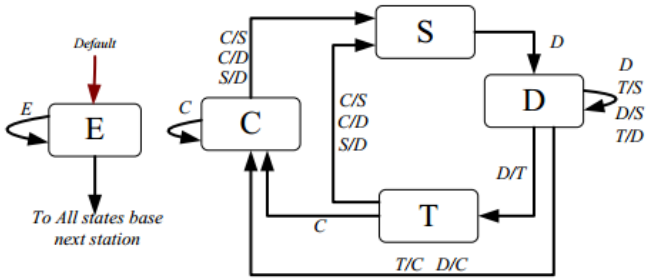


Direct Encoding

- Proposal is to direct encode to 256B/257B which is compatible with 802.3bj

An Example of 256b/257b Direct Coding





- **256b/257b direct coding is possible and straightforward**
- **256b/258b direct coding can simply extend the block header bit to two bits**

From: http://www.ieee802.org/3/400GSG/public/13_09/song_400_01_0913.pdf

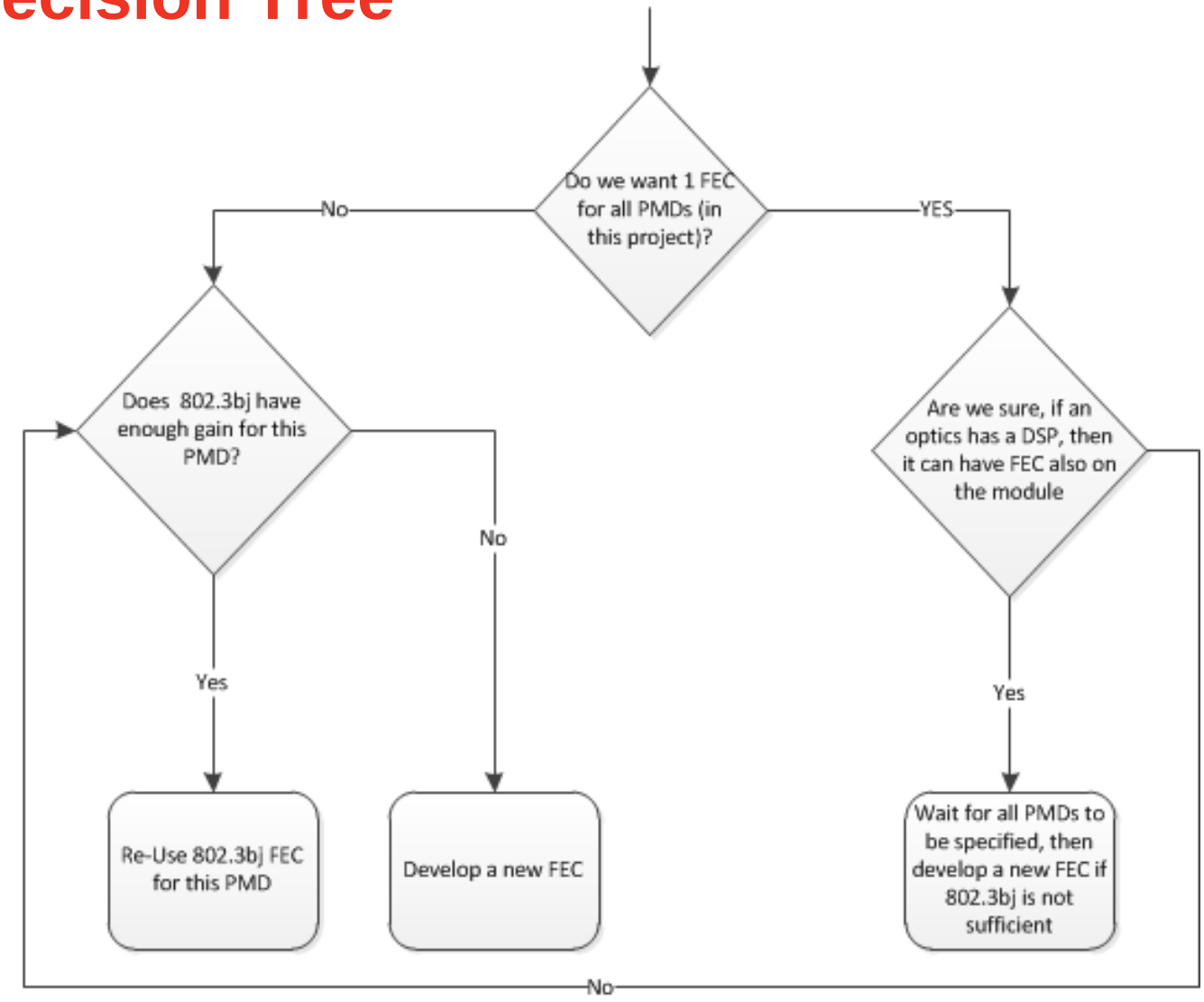
Scrambling

- Seems to be lots of consensus on re-using the current X^58 scrambler, self synchronous
- There are open questions on if the scrambler should be across the aggregate payload (like 100GbE)
  - Options:
    - Aggregate scrambler (like 100G is)
    - Scrambler per lane (PCS or FEC)
    - Scrambler at some other level, such as on 100G chunks
  - Add sizing comparisons? Size might be small compared to FEC, but everything matters?
- Why did 100GbE go with a single aggregate scrambler?
  - There was concern over per PCS lane scramblers, it is difficult to analyze if spill-out errors impact the CRC32 detection capability
    - If we always have FEC this does not seem to be a concern now
- When to do scrambling?
  - For 802.3bj, data remains scrambled even when doing transcoding, you could wait to do scrambling until after transcoding (or direct encode)

FEC

- There have been many proposals to re-use the RS-FEC from 802.3bj, and use 4 instances of it in parallel, one instance per four physical lanes
  - Could look at KR4 and KP4 re-use if needed
- Does this cover all PMDs?
  - Might not even cover the 100m MMF PMD?
- What Considerations do we have for the FEC choices?
  - Gain
  - Latency
  - Complexity
  - Re-use

FEC Decision Tree



This assumes that if the 802.3bj FEC has sufficient gain we would use it, unless another issue comes up.

Alignment

- Do we keep with something similar to 802.3bj for alignment?
  - Repeated pattern across all lanes, then other identifying markers for lane identification?
  - How important is re-use in this area from 100G?
- What other information should be communicated?
  - BIP? How useful is this with FEC?
  - If there is BIP, is it end to end, or segment by segment?
  - We do need a way to tell if errors are happening, and possible trigger on an error threshold

802.3bj AMs

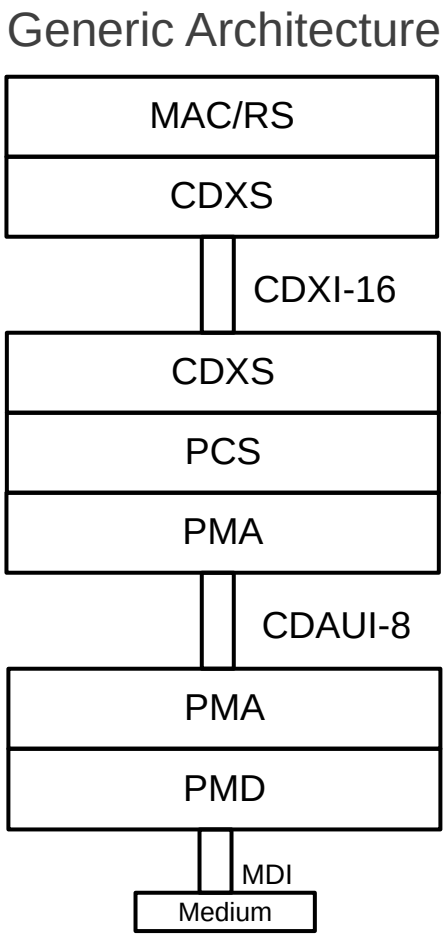
- Clause 91 defines how Alignment Markers are mapped when sent across the 4 FEC lanes
  - They are re-mapped to the FEC lanes so they appear consecutively on a given FEC lanes
  - A 5b pad is added to the end to round make them fit within a even number of 257b blocks (20*64+5 = 257*5)
  - AM0 and AM16 are repeated on all 4 FEC lanes to make it less logic intensive to find block alignment
  - The remaining AMs uniquely identify the 4 FEC lanes

FEC Lane	Reed-Solomon symbol index (10 bit symbols																																
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
0	AM0 (0 … 63)										AM4					AM8					AM12					AM16								
1	AM0										AM5					AM9					AM13					AM16								
2	AM0										AM6					AM10					AM14					AM16								
3	AM0										AM7					AM11					AM15					AM16								

5b pad

Do We Need to Define a CDXS?

- Will we have a 16 and an 8 lane electrical interface in this project?
- If yes then we likely need a CDXS
- 16 lane interface out of a large ASIC, and 8 lane interface to at least some PMDs would require the CDXS



Striping

- We will need to decide how to stripe data to multiple lanes
  - In the past we striped at the encoded block level; 802.3ba used 66b block striping, 802.3bj uses 10b FEC block striping
  - We might want to stripe at a FEC block size, 256B/257B block size or at some other granularity
  - We need to decide the encoding/FEC strategy then we can decide on the striping

Muxing

- When we need to multiplex logical lanes together, we have several options that have been used or identified:
  - Bit Muxing, used in 802.3ba
  - FEC Orthogonal Multiplexing
  - Block muxing
  - There might be a complex line encoding that has a unique multiplexing method

Summary

- These slides explored some of the technology options we have around a 400GbE PCS
- We can look at practical re-use of logic between 100GbE and 400GbE, but we need progress on the PMD choices to make a lot of progress

Thanks!