Baseline Proposal for 1.6TbE PCS Lane Formation and AM Insertion

Eugene Opsasnick (Broadcom), Kapil Shrikhande (Marvell), Jeff Slavick (Broadcom)

IEEE P802.3dj Task Force Plenary Meeting, March 2023

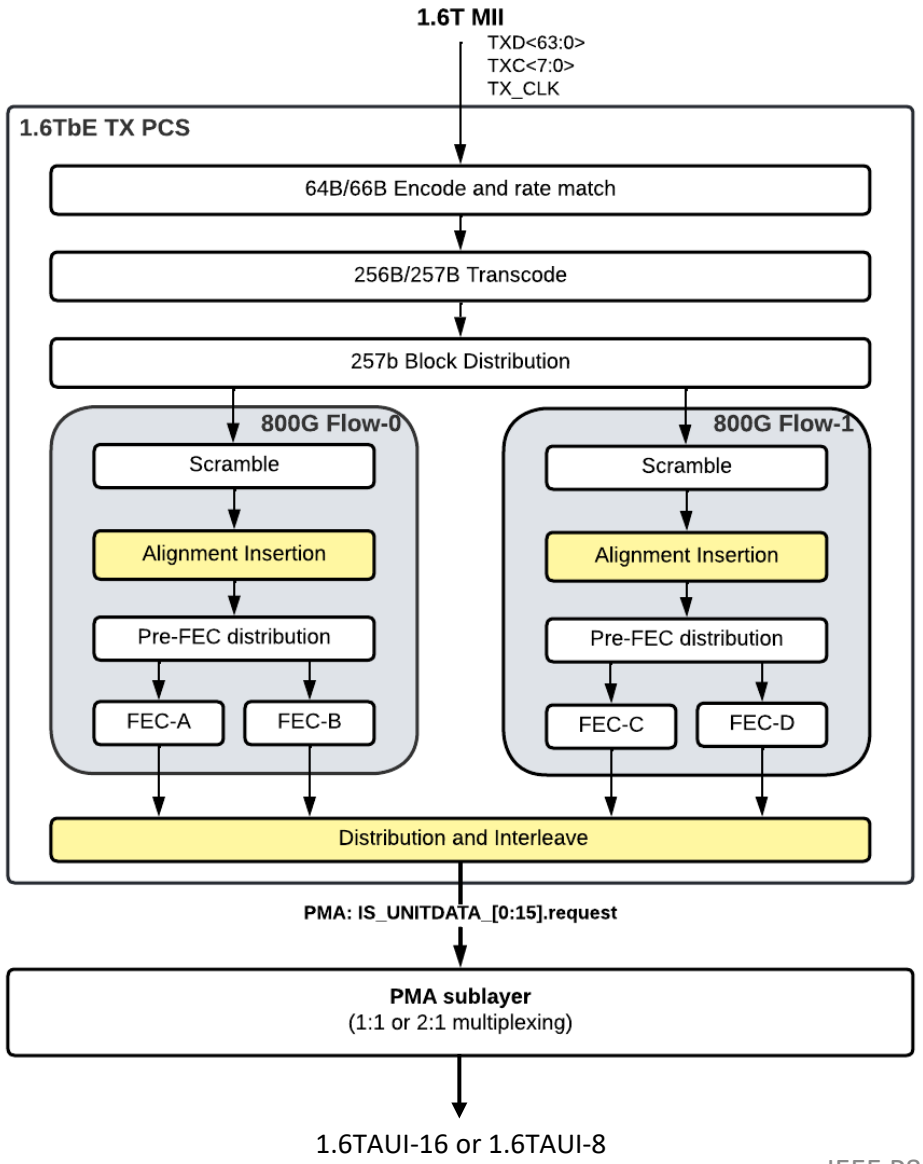
Contributors and Supporters

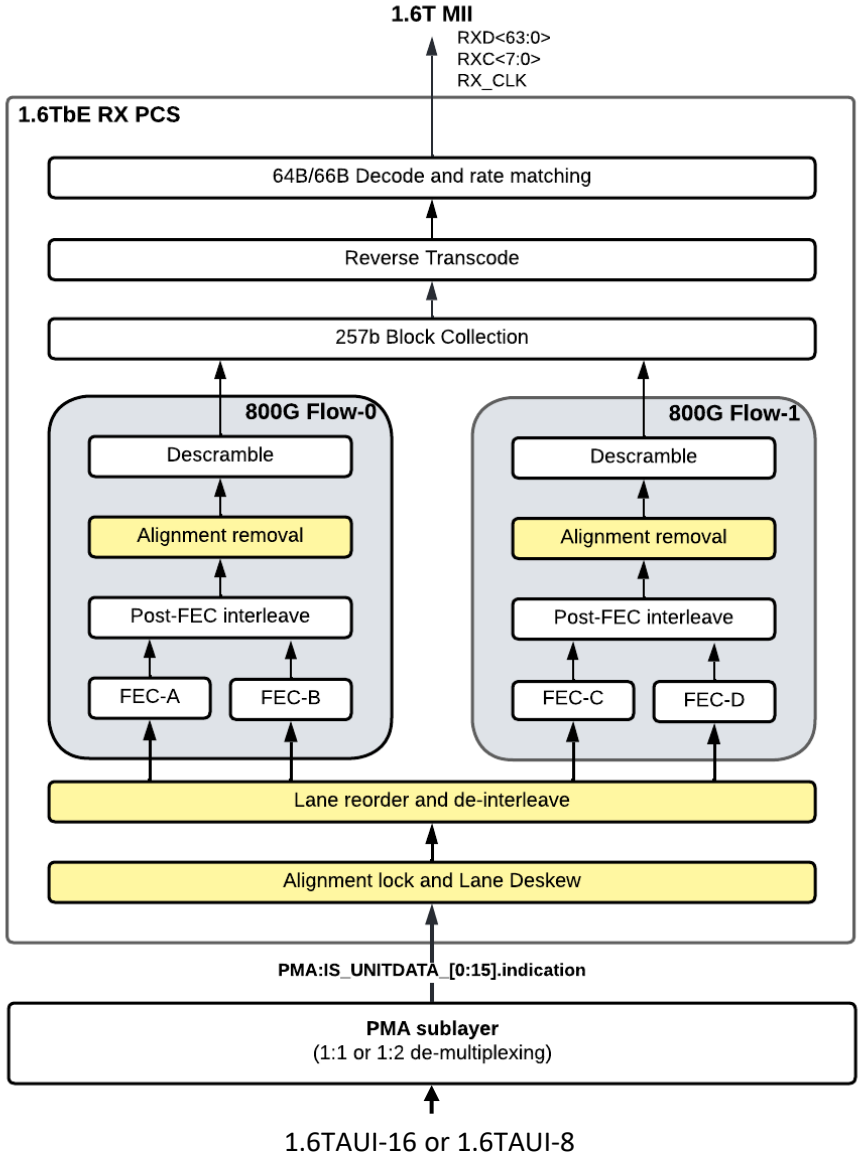
- Mark Gustlin, Cisco
- Adee Ran, Cisco
- Daniel Koehler, Synopsys
- Gary Nicholl, Cisco
- Arthur Marris, Cadence
- Eric Maniloff, Ciena
- Jerry Pepper, Keysight
- Kent Lusted, Intel
- Matt Brown, Huawei
- David Ofelt, Juniper
- Shawn Nicholl, AMD
- Tom Huber, Nokia
- Ted Sprague, Infinera
- Xinyuan Wang, Huawei
- Xiang He, Huawei
- Zvi Rechtman, Nvidia
- Ben Jones, AMD
- Rick Rabinovich, Keysight
- Viet Tran, Keysight
- Howard Heck, Intel
- Jeffery Maki, Juniper Networks
- Shimon Muller, Enfabrica
- Ed Nakamoto, Spirent
- Dave Estes, Spirent
- Chris DiMinico, PHY-SI/SenTekse/MC Communications
- Cedric Lam, Google
- Kenneth Jackson, Sumitomo Electric
- Frank Effenberger, Futurewei Technologies
- Roberto Rodes, Coherent
- Chris Cole, Quintessent
- Leon Bruckman, Huawei
- David Malicoat, Malicoat Networking Solutions

Scope

- 1.6TbE PCS Baseline adopted 2/6/23
  - gustlin_3dj_01b_230206.pdf (slides 6-12) - Motion #10.
  - Includes all PCS blocks except PCS lane formation and AM Insertion/Removal
- Scope of this presentation:
  - Specify the PCS Lane Formation and associated AM Insertion details
    - See highlight sub-blocks on next slide
  - FEC degrade signaling and HI_SER monitor across the PCS flows
- Together, the two presentations complete the 1.6TbE PCS baseline for 802.3dj
  - Any 800GbE PCS function not explicitly called out is to be included

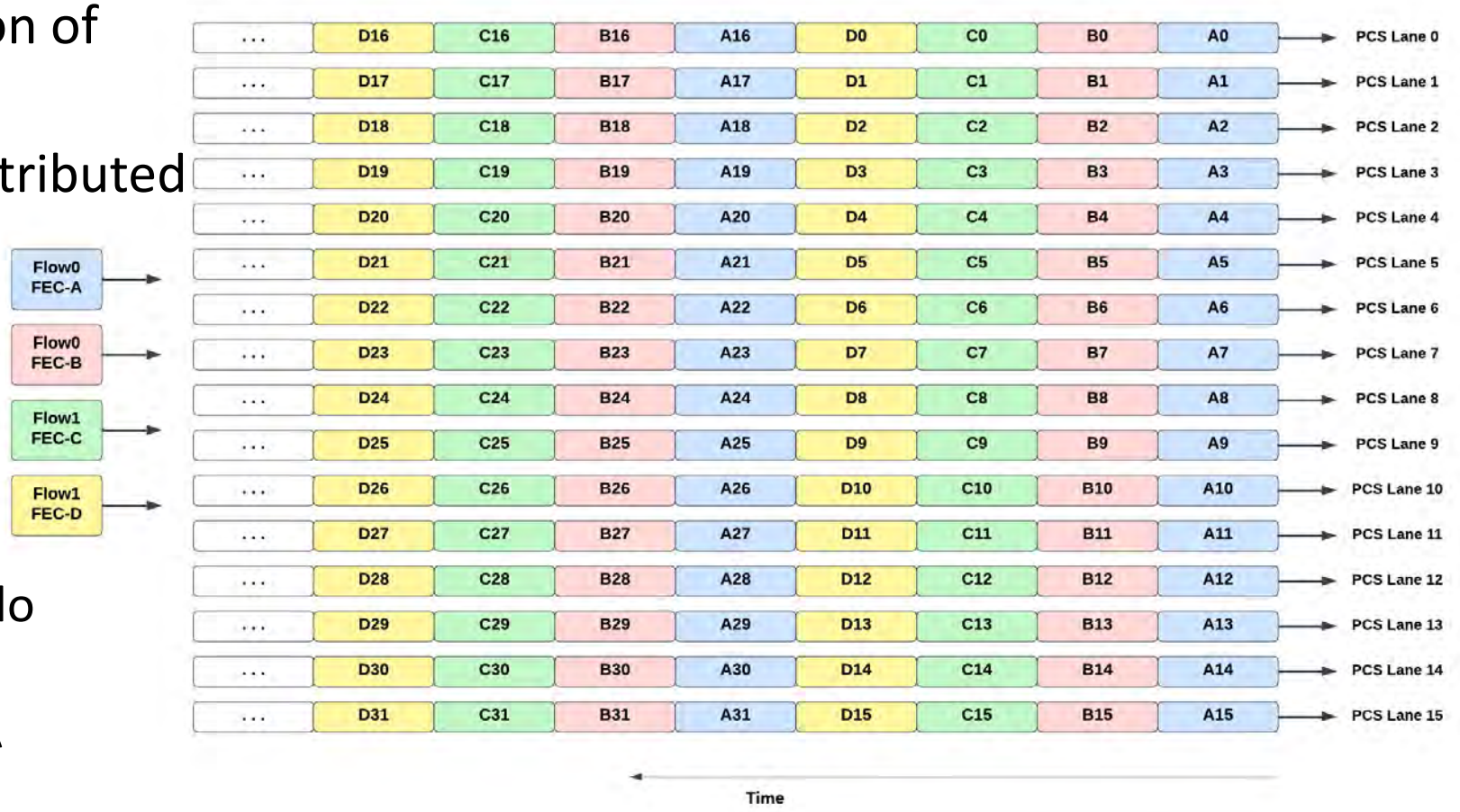
1.6TbE PCS Functional Blocks – TX and RX





1.6TbE PCS Lane Formation

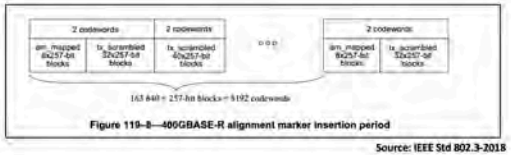
- Round-robin distribution of 10-bit RS-FEC symbols
- All 4 RS-FEC CW are distributed into 16 PCS Lanes
- 100G per PCS lane
- No “checkerboard” pattern on PCS lanes, assuming:
  - PMA for 1.6TbE will do symbol muxing
  - No bitmuxing in PMA



1.6TbE Alignment Markers

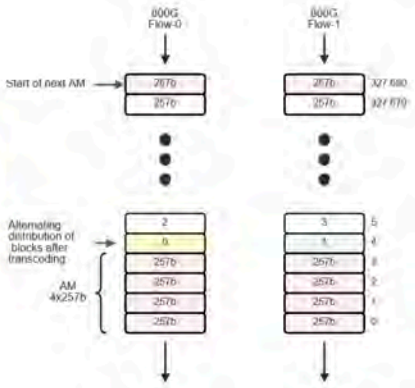
- From adopted baseline gustlin_3dj_01b_230206.pdf :

Alignment Marker Insertion



- Markers inserted at consecutive 257b blocks across both 800G flows
- Uses 16 PCSL
  - Total size of markers is same as 400GbE
  - Increased spacing between markers to maintain frequency of arrival.

GbE	#PCSL	AM group size (x 257b)	Spacing (in CWs)	#PCS flows	AM Spacing (in 257b per flow)
200	8	4	4k	1	81,920
400	16	8	8k	1	163,840
800	32	16	16k	2	163,840
1600	16	8	32k	2	327,680



Alignment Marker Encoding

- With 16 PCSL
- CM0-CM5 and UP0-UP2 are unchanged from 400GbE CL119
- UM0-UM5 are inverted from 400GbE
- Resulting UMs differ from 400GbE and 800GbE
- Clock Content and Baseline Wander Analysis – TBD
- UP and UM values can be adjusted if necessary
- Open issue: How to form the AMs in a coherent way so they appear correctly on physical lanes

PCS Lane #	Encoding														
	CM0	CM1	CM2	UP0	CM3	CM4	CM5	UP1	UM0	UM1	UM2	UP2	UM3	UM4	UM5
0	0x9A	0x4A	0x26	0xB6	0x65	0xB5	0xD9	0xD9	**0xFE**	**0x8E**	**0x0C**	0x26	**0x01**	**0x71**	**0xF3**
1	0x9A	0x4A	0x26	0x04	0x65	0xB5	0xD9	0x67	**0xA5**	**0x21**	**0x81**	0x98	**0x5A**	**0xDE**	**0x7E**
2	0x9A	0x4A	0x26	0x46	0x65	0xB5	0xD9	0xFE	**0xC1**	**0x0C**	**0xA9**	0x01	**0x3E**	**0xF3**	**0x56**
3	0x9A	0x4A	0x26	0x5A	0x65	0xB5	0xD9	0x84	**0x79**	**0x7F**	**0x2F**	0x7B	**0x86**	**0x80**	**0xD0**
4	0x9A	0x4A	0x26	0xE1	0x65	0xB5	0xD9	0x19	**0xD5**	**0xAE**	**0x0D**	0xE6	**0x2A**	**0x51**	**0xF2**
5	0x9A	0x4A	0x26	0xF2	0x65	0xB5	0xD9	0x4E	**0xED**	**0xB0**	**0x2E**	0xB1	**0x12**	**0x4F**	**0xD1**
6	0x9A	0x4A	0x26	0x3D	0x65	0xB5	0xD9	0xEE	**0xBD**	**0x63**	**0x5E**	0x11	**0x42**	**0x9C**	**0xA1**
7	0x9A	0x4A	0x26	0x22	0x65	0xB5	0xD9	0x32	**0x29**	**0x89**	**0xA4**	0xCD	**0xD6**	**0x76**	**0x5B**
8	0x9A	0x4A	0x26	0x60	0x65	0xB5	0xD9	0x9F	**0x1E**	**0x8C**	**0x8A**	0x60	**0xE1**	**0x73**	**0x75**
9	0x9A	0x4A	0x26	0x6B	0x65	0xB5	0xD9	0xA2	**0x8E**	**0x3B**	**0xC3**	0x5D	**0x71**	**0xC4**	**0x3C**
10	0x9A	0x4A	0x26	0xFA	0x65	0xB5	0xD9	0x04	**0x6A**	**0x14**	**0x27**	0xFB	**0x95**	**0xEB**	**0xD8**
11	0x9A	0x4A	0x26	0x6C	0x65	0xB5	0xD9	0x71	**0xDD**	**0x99**	**0xC7**	0x8E	**0x22**	**0x66**	**0x38**
12	0x9A	0x4A	0x26	0x18	0x65	0xB5	0xD9	0x5B	**0x5D**	**0x09**	**0x6A**	0xA4	**0xA2**	**0xF6**	**0x95**
13	0x9A	0x4A	0x26	0x14	0x65	0xB5	0xD9	0xCC	**0xCE**	**0x68**	**0x3C**	0x33	**0x31**	**0x97**	**0xC3**
14	0x9A	0x4A	0x26	0xD0	0x65	0xB5	0xD9	0xB1	**0x35**	**0x04**	**0x59**	0x4E	**0xCA**	**0xFB**	**0xA6**
15	0x9A	0x4A	0x26	0xB4	0x65	0xB5	0xD9	0x56	**0x59**	**0x45**	**0x86**	0xA9	**0xA6**	**0xBA**	**0x79**

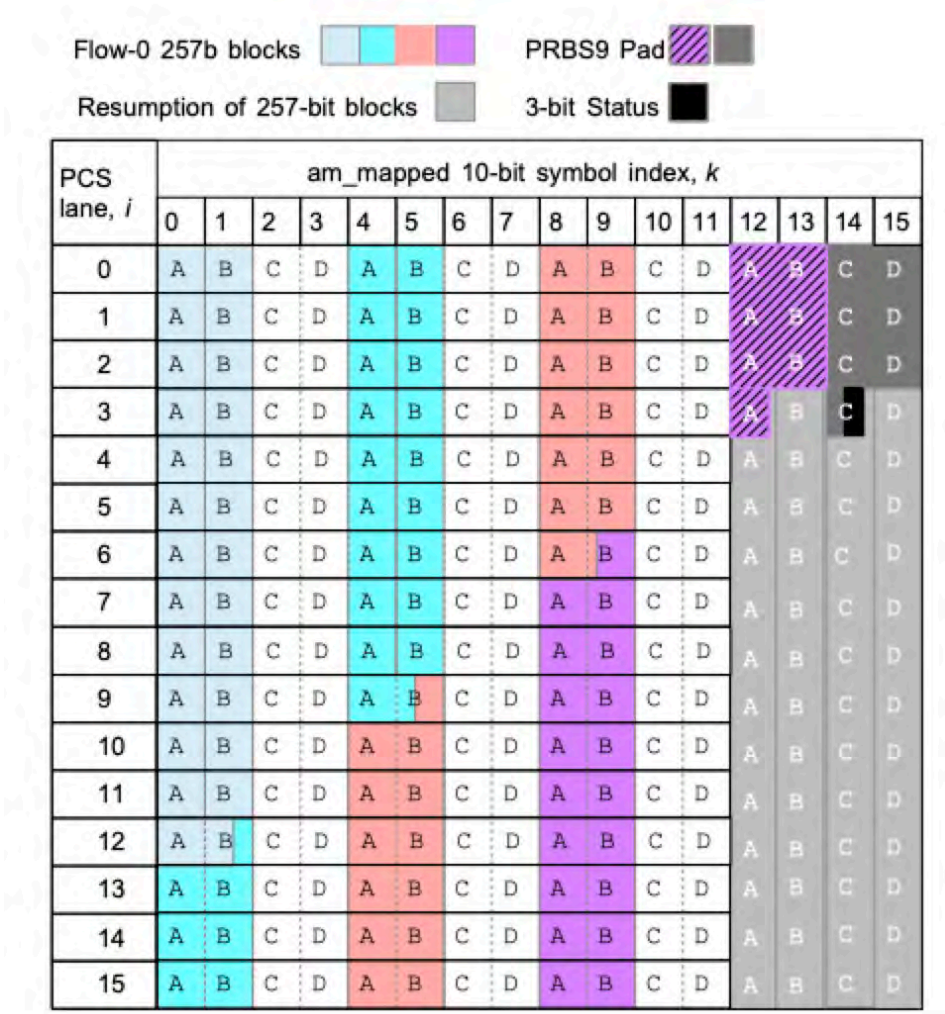
Note: in table above, bolded text indicates inverted values from CL 119 AM values

11

1.6TbE AM pattern across PCS lanes

- AM data is eight 257-bit blocks
  - 4x257b inserted by each flow
- 120-bit AM marker per PCS lane

  In figure to right:
  - Symbols 0-11 of each lane are defined AM values
  - 257-bit AM blocks inserted by Flow-0 (CW-A and CW-B) highlighted in color
  - Padding in PCSLs 0-3, Symbols 12-15
    - 68-bits padding in Flow-0
    - 65-bits padding + 3-bit status in Flow-1
  - PRBS9 padding data in each flow is independently generated per flow.
    - Each flow should use different seeds for the PRBS9 pattern
  - TX AM status Field (tx_am_sf<2:0>)
    - Only at end of padding in Flow-1
    - Status is based on all 4 CWs

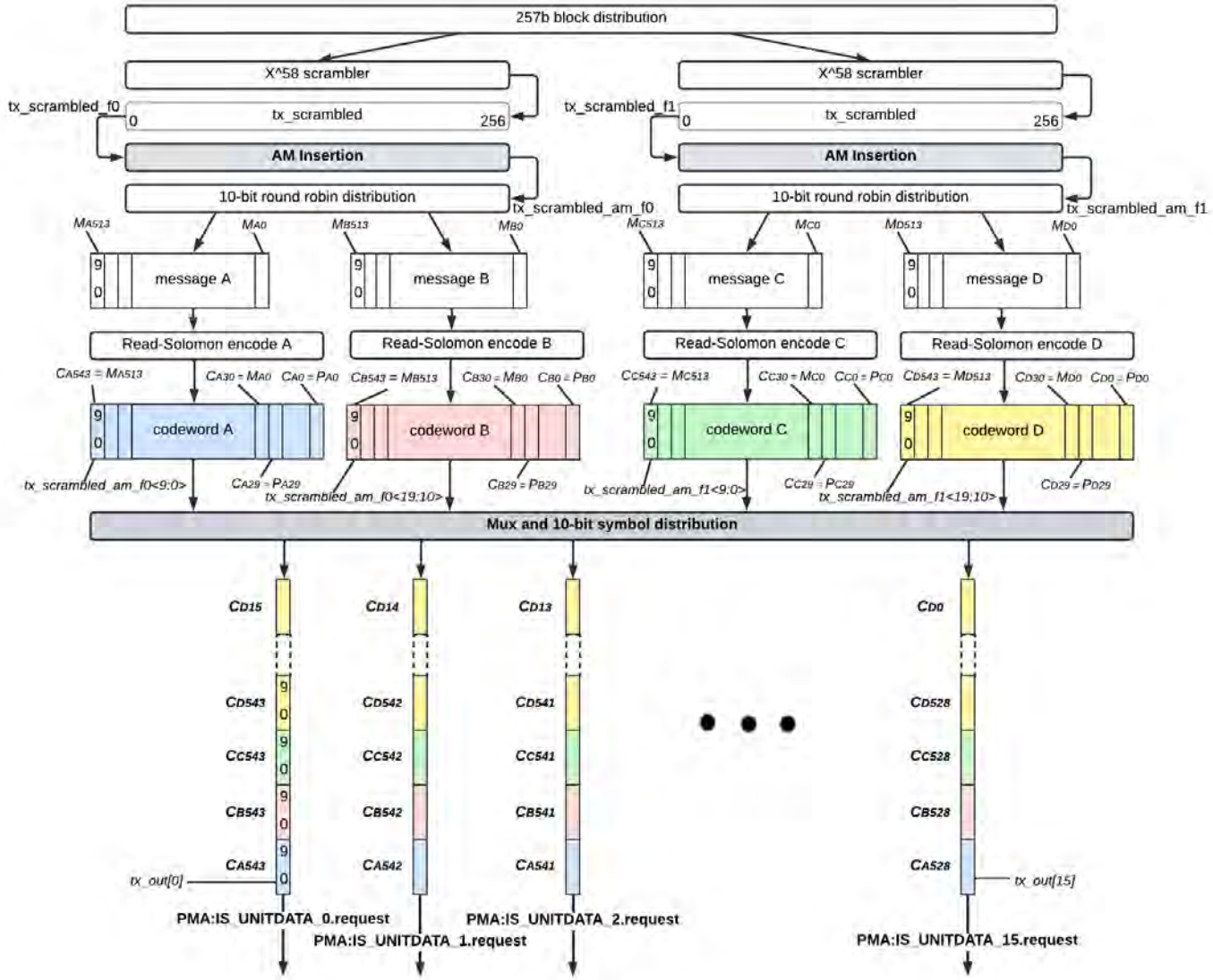


1.6TBASE-R Alignment marking mapping to PCS lanes

1.6TbE AM Insertion

- Each flow requires a unique insertion pattern definition to make the AMs appear on the PCS lanes correctly since all 4 CWs are inserted into each PCS lane.
- For each of the sixteen 120-bit AMs:
  - AM bits 0-19, 40-59 and 80-99 come from Flow-0
  - AM bits 20-39, 60-79 and 100-119 come from Flow-1
  - Plus additional padding is added to align to a 257b boundary

1.6TbE PCS Transmit Symbol Distribution and Bit Ordering



AM Mapping Pseudo-Code (in CL 119 style)

- AM mapping into TX data stream (AM insertion)

```
for all k=0 to 2
  for all j=0 to 15                                    ;; j = PCS lane number
    am_mapped_f0<320k+20j+9  : 320k+20j>    = am_{j}<40k+9  : 40k>     ;; CW A
    am_mapped_f0<320k+20j+19 : 320k+20j+10> = am_{j}<40k+19 : 40k+10>  ;; CW B
    am_mapped_f1<320k+20j+9  : 320k+20j>    = am_{j}<40k+29 : 40k+20>  ;; CW C
    am_mapped_f1<320k+20j+19 : 320k+20j+10> = am_{j}<40k+39 : 40k+30>  ;; CW D

am_mapped_f0<1027:960> = PRBS9_f0<67:0>   ;; Flow-0 padding
am_mapped_f1<1024:960> = PRBS9_f1<64:0>   ;; Flow-1 padding
am_mapped_f1<1027:1025> = tx_am_sf<2:0>   ;; tx am status field

tx_scrambled_am_f0<1027:0> = am_mapped_f0<1027:0>
tx_scrambled_am_f1<1027:0> = am_mapped_f1<1027:0>
```

Pre-FEC and Post-FEC Symbol Distribution Pseudo-Code

- Pre-FEC Symbol Distribution

```
for all i=0 to 513
  Ma<(513-i)> = tx_scrambled_am_f0<(20i+9):(20i)>
  Mb<(513-i)> = tx_scrambled_am_f0<(20i+19):(20i+10)>
  Mc<(513-i)> = tx_scrambled_am_f1<(20i+9):(20i)>
  Md<(513-i)> = tx_scrambled_am_f1<(20i+19):(20i+10)>
```

- Post-FEC Symbol Distribution

```
for all k=0 to 33
 for all j=0 to 15
    tx_out<64k+j   > = Ca<543-16k-j>
    tx_out<64k+j+16> = Cb<543-16k-j>
    tx_out<64k+j+32> = Cc<543-16k-j>
    tx_out<64k+j+48> = Cd<543-16k-j>
```

** tx_scrambled_am_f{0:1}, FEC messages (Ma, Mb, Mc, Md), Codewords (Ca, Cb, Cc, Cd), and tx_out correspond to labels in the figure on slide 9.

1.6TbE RX AM Lock, Reorder, De-interleave and AM Deletion

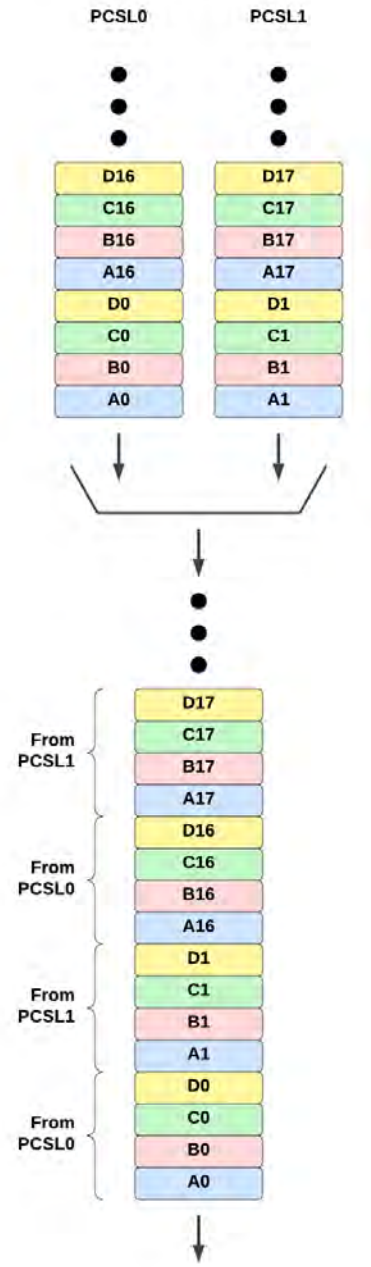
- Use the same method as 172.2.5.1 for AM lock and deskew for 16 PCS lanes
- After all PCS lanes are aligned, deskewed, and reordered, the PCS lanes are de-interleaved to reconstruct the original stream of four FEC codewords.
- Use the same method as 119.2.5.5 for AM removal and rx_am_sf assignment.

1.6TbE FEC Degrade Signal Generation and HI_SER Monitor

- FEC degraded SER functionality is same as CL119 with following exception
  - Symbol errors counted on all PCS lanes, across all 4 FEC codewords
  - Single FEC degrade SER status bit for the 1.6TbE PCS (same as CL119)
    - Different from CL172 which has 2 FEC degrade SER bits that are logically combined into a FEC degrade SER for the 800GbE PCS
  - FEC degrade functionality is optional (same as CL119 and CL172)
- HI_SER Monitoring is same as CL119 with following exception
  - Symbol errors are counted across all 4 FEC codewords

Notes for PMA Symbol Muxing

- PMA for 16x100G 1.6TAUI-16:
  - No additional lane muxing (1:1 PCSL to PMA lanes)
- PMA for 8x200G 1.6TAUI-8
  - Requires 2:1 lane mux/de-mux using 40-bit blocks
  - No restrictions on which PCSLs to combine together
  - PMA must align to 4-RS-symbol boundaries
    - Deskew only to 40-bit boundaries
  - Example: Symbol muxing of PCSL #0 and PSCL #1 as shown
    - Resulting RS symbols order:
    - [(A0, B0, C0, D0), (A1, B1, C1, D1), (A16, B16, C16, D16), (A17, …), …]
    - With skew, other resulting patterns are possible (e.g. A17 could follow D0)
- A separate proposal is offered for the complete PMA definition



Summary

- Suggested course of action
  - Adopt the AM insertion and deletion, PCS lane formation, and FEC degrade signaling as outlined in this presentation for the 1.6TbE PCS.
  - Complete the analysis of Clock Content and Baseline Wander for AM encoding values as specified in gustlin_3dj_01b_23_0206.pdf
    - If necessary, adjust marker UP and UM values

Thanks

Proposed Straw Poll

- I would support adopting opsasnick_3dj_01a_2303, slides 3, 5-9, 12-13, as a supplement to the previously adopted 1.6TbE PCS baseline from gustlin_3dj_01b_230206.pdf. These two presentations together complete the baseline for the 1.6TbE PCS.

  Y:

  N:

  A:

  NMI:

400GbE AM pattern across PCS lanes (CL 119 – reference)

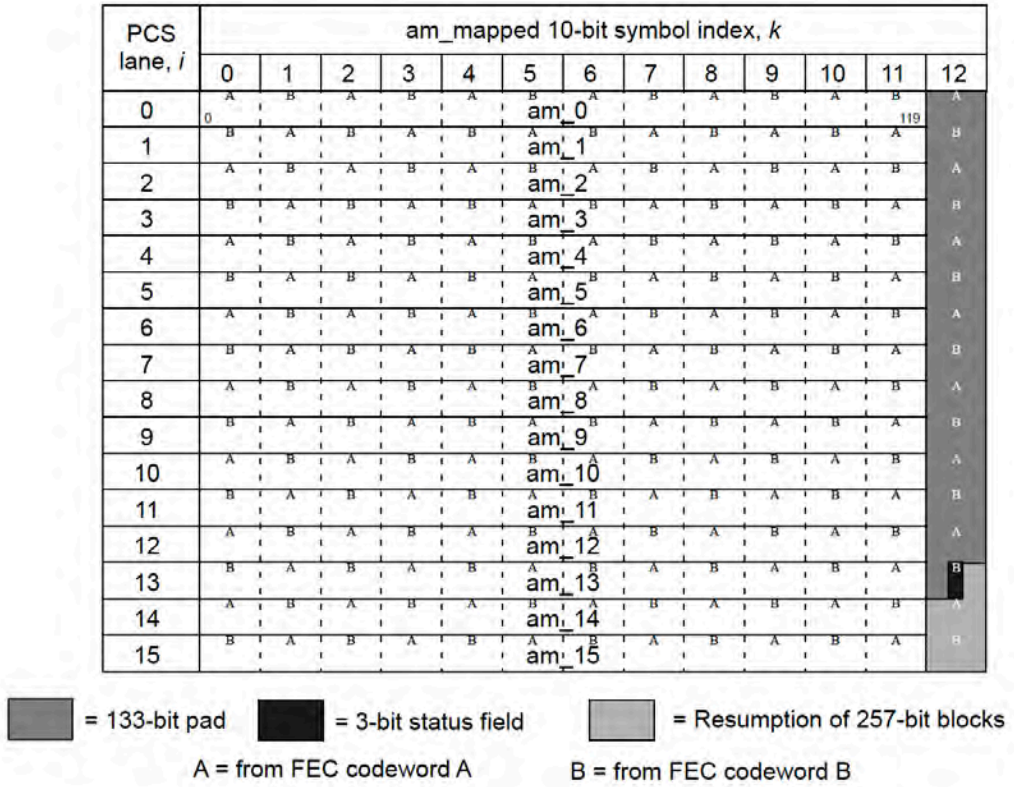


Figure 119–7—400GBASE-R alignment marker mapping to PCS lanes