400GbE PCS and PMA Baseline Proposals

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

March 2015 Berlin

Mark Gustlin – Xilinx
Arthur Marris - Cadence
Gary Nicholl - Cisco
Dave Ofelt - Juniper
Jerry Pepper - Ixia
Andre Szczepanek – Inphi
Steve Trowbridge – ALU
Tongtong Wang - Huawei

References

- This work is based on much of these preceding slide decks/work

http://www.ieee802.org/3/bs/public/15_01/marris_3bs_01_0115.pdf
http://www.ieee802.org/3/bs/public/15_01/wang_x_3bs_01a_0115.pdf
http://www.ieee802.org/3/bs/public/15_01/slavick_3bs_01a_0115.pdf
http://www.ieee802.org/3/bs/public/15_01/gustlin_3bs_02_0115.pdf
http://www.ieee802.org/3/bs/public/14_11/gustlin_3bs_03a_1114.pdf
http://www.ieee802.org/3/bs/public/14_11/dambrosia_3bs_01_1114.pdf
http://www.ieee802.org/3/bs/public/14_09/anslow_3bs_02_0914.pdf
http://www.ieee802.org/3/bs/public/14_09/wang_z_3bs_01_0914.pdf
http://www.ieee802.org/3/bs/public/14_09/wang_t_3bs_01a_0914.pdf
http://www.ieee802.org/3/bs/public/14_07/wang_x_3bs_01_0714.pdf
http://www.ieee802.org/3/bs/public/14_07/trowbridge_3bs_01_0714.pdf
http://www.ieee802.org/3/bs/public/14_07/wang_t_3bs_01_0714.pdf
http://www.ieee802.org/3/bs/public/14_07/gustlin_3bs_04_0714.pdf
http://www.ieee802.org/3/bs/public/14_07/gustlin_3bs_02_0714.pdf
http://www.ieee802.org/3/bs/public/14_05/wang_x_3bs_01_0514.pdf
http://www.ieee802.org/3/bs/public/14_05/trowbridge_3bs_01_0514.pdf
http://www.ieee802.org/3/bs/public/14_05/barrass_3bs_01_0514.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

Table Of Contents

Introduction

- This looks at a baseline PCS and PMA proposal, there are still some open issues that are being investigated

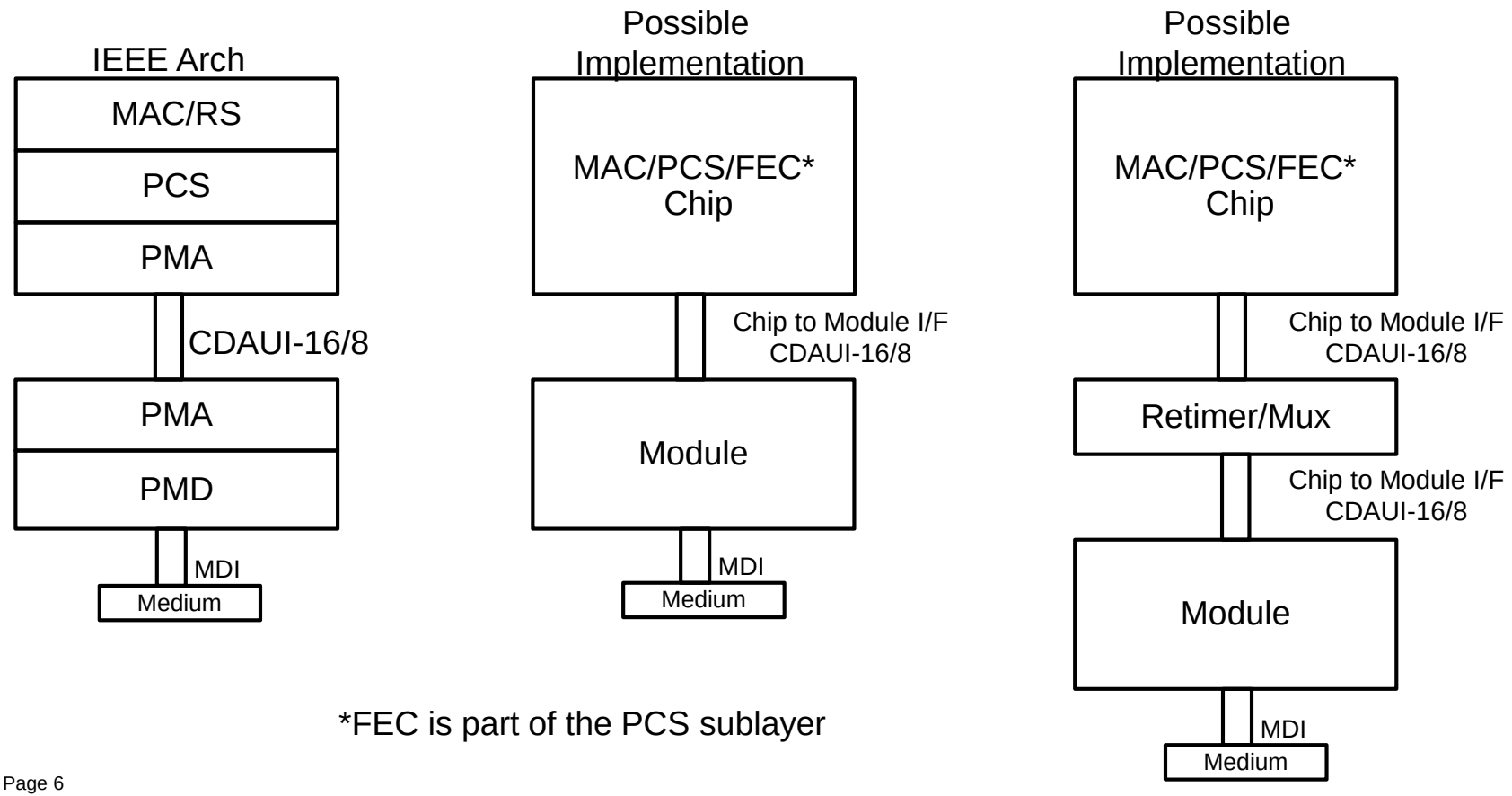
Review of the Sublayer Functions

Sublayer	10GbE	100GbE	400GbE (proposed)
MAC	Framing, addressing, error detection	Framing, addressing, error detection	Framing, addressing, error detection
Extender	XGS (PCS + PMA)	N/A	CDXS (PCS)
PCS	Coding (8B/10B, 64B/66B), lane distribution, EEE	Coding (64B/66B), lane distribution, EEE	Coding, lane distribution, EEE, FEC
FEC	FEC, transcoding	FEC, transcoding, align and deskew	N/A
PMA	Serialization, clock and data recovery	Muxing, clock and data recovery, HOM	Muxing, clock and data recovery, HOM??
PMD	Physical interface driver	Physical interface driver	Physical interface driver

Note that there are variations with a single speed, not all are captured in this table

PCS Architecture

- Based on the adopted system architecture
- You can mix the two, just PMA Muxing to go back and forth
- In this instance a single FEC is used, across up to 5 interfaces (in the PCS sublayer)
- Assuming a single FEC covers up to 5 interfaces

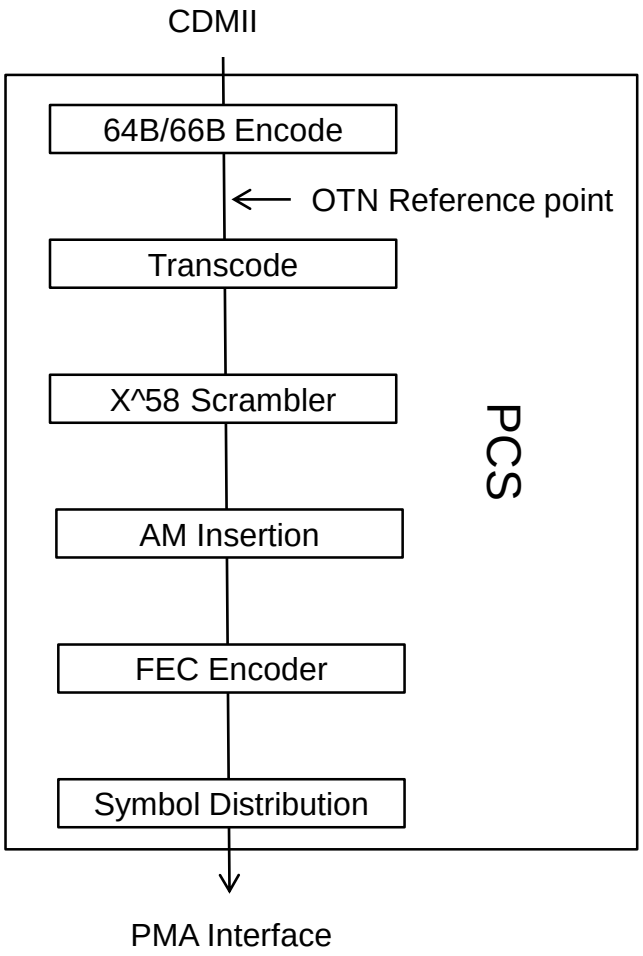


*FEC is part of the PCS sublayer

Table Of Contents

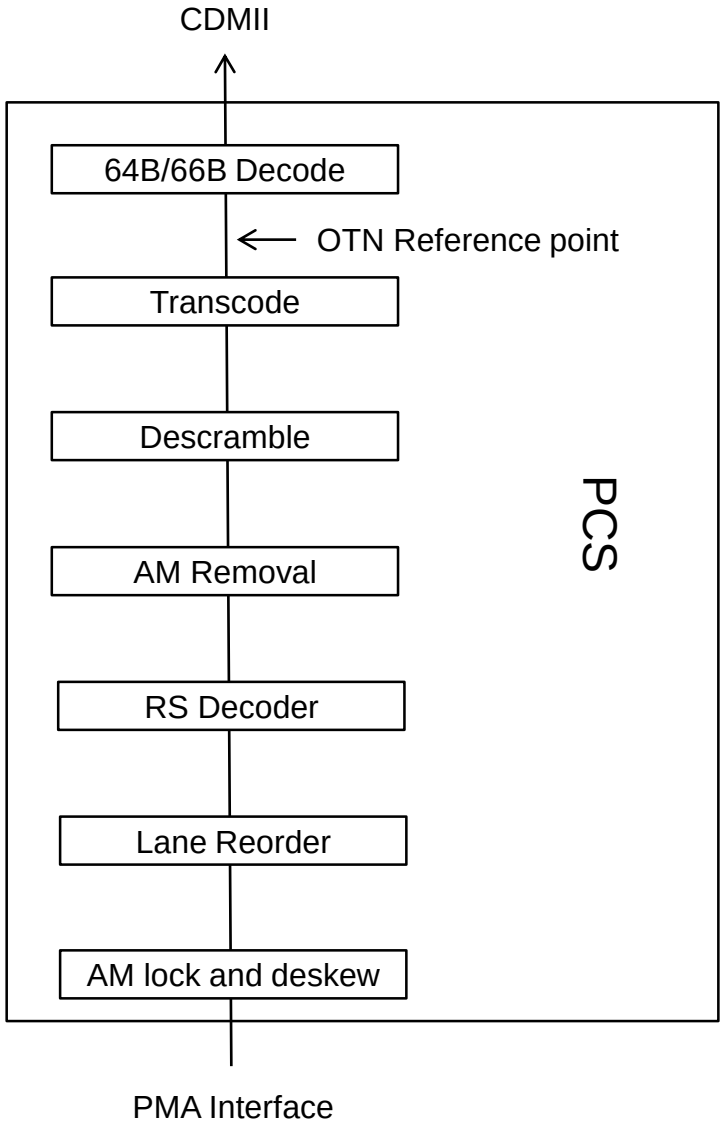
Proposed TX PCS Data Flow

- 64B/66B encode based on clause 82
- Transcode to 256B/257B based on clause 91
- Scrambler is moved to after the Transcoding to simplify the flow
- FEC Encoder is described later
- 16 PMA lanes (similar to PCS/FEC lanes)
- Location of the OTN reference point is as shown and adopted in the January meeting



Proposed RX PCs Data Flow

- Reverse of TX
- Allows for arbitrary lane arrival



Scrambling

- Re-use the X^58 self synchronous scrambler, but after the transcoding
- Run it across all payload information, but not the AMs
- Scrambling includes all 257 bits
  - Note that this is slightly different and simpler than 802.3bj

Table Of Contents

Which FEC to use?

- This baseline proposal uses RS FEC (544,514,10)
  - It is possible to use a different FEC if a PMD requires it in the adopted architecture, but this RS FEC is the best starting point now given the various PMDs and their stated requirements
  - Assume all PMDs and all electrical interfaces are covered by this FEC
    - This means that SR16 is covered by this RS FEC for example, even though it only requires a KR4 FEC (is the overspeed acceptable for SR16 interfaces?)
    - One exception would be a DMT PMD, it is proposed to have a stronger FEC on the optical module, the overall adopted architecture supports this
- We need to adopt PMD choices and understand their FEC requirements to finalize this choice

1x400G vs. 4x100G FEC

- Decision points:
  - Do we need FOM for muxing and to preserve gain? -> Choose 4x100G architecture
  - Otherwise go with 1x400G architecture to allow lowest latency and cleanest solution for the long run
  - Other things under consideration
  - Processing latency is implementation dependent, y and z can be similar

Category	1x400G	4x100G
Block Latency	~12ns	~50ns
Processing Latency	y	z
Synergy with 100GbE	Some	Higher
Muxing		Allows for FOM
Implementation Size	1x	1.3-0.9x*

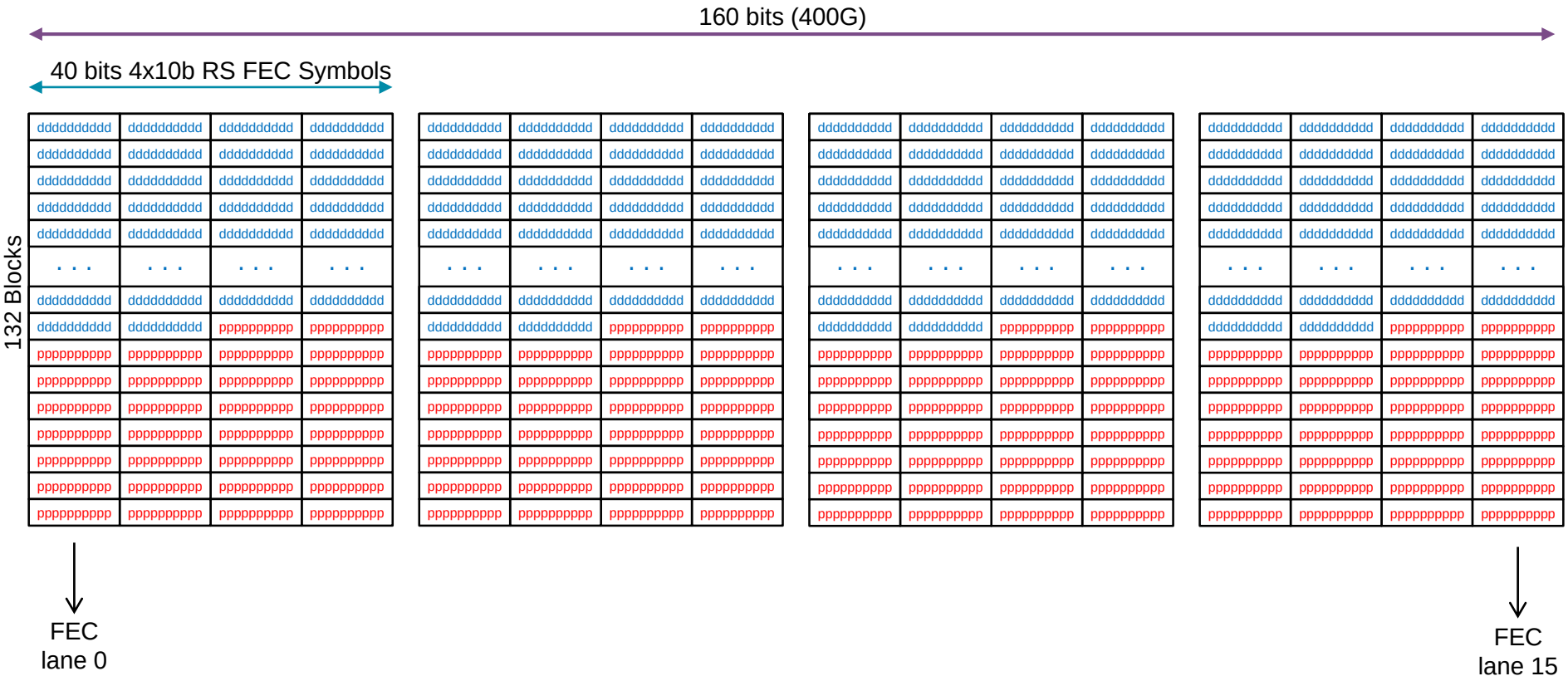
* Depends on assumptions, is 4x100G already part of the chip etc.

See wangx_01_1214_logic for comparative FEC sizing details

Table Of Contents

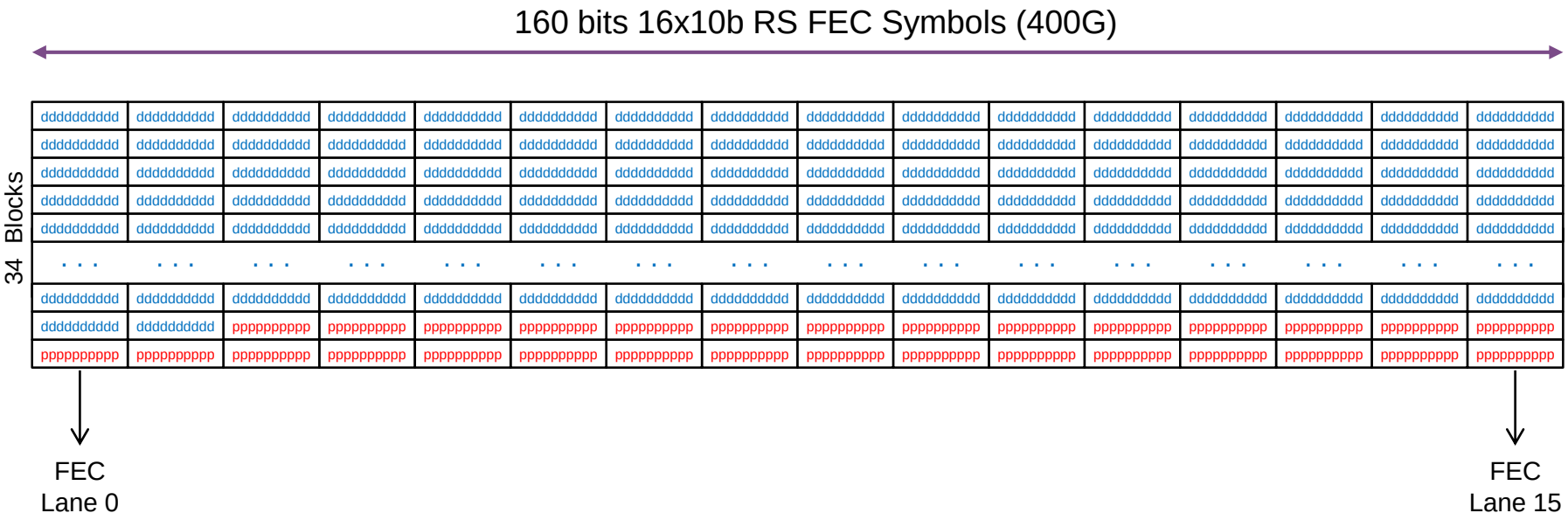
400GbE Data Distribution – 4x100G

- Below the RS-FEC sublayer, with using 4x802.3bj KP4 FEC, you would naturally have 16 FEC lanes
  - ddddddddddd = protected data
  - ppppppppppp = FEC Parity addition



400GbE Data Distribution – 1x400G

- Below the RS-FEC sublayer, with using 1x802.3bj KP4 FEC (400G single FEC instance), you would naturally have 16 FEC lanes
  - ddddddddddd = protected data
  - ppppppppppp = FEC Parity addition



400GbE 257b Block Mapping

- This shows how the 257b blocks fit within the FEC block

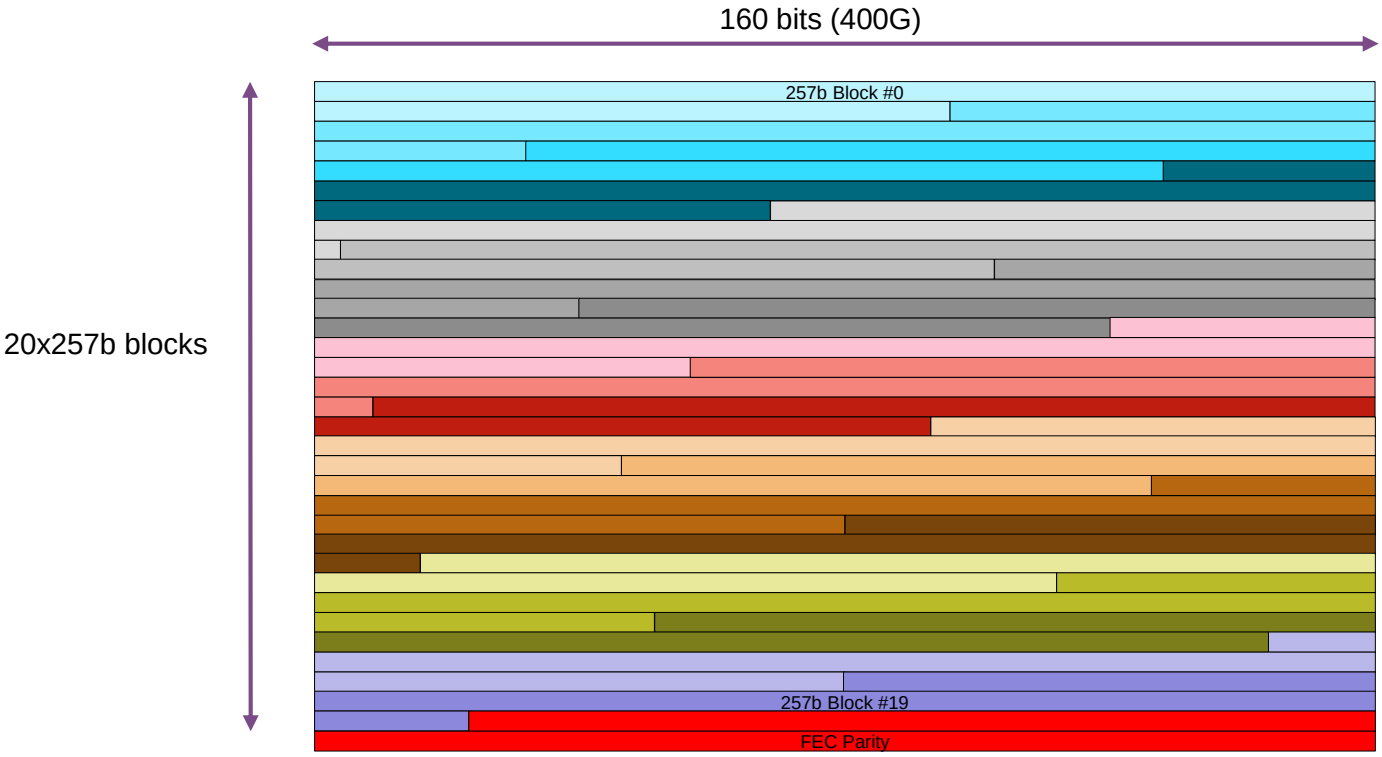


Table Of Contents

802.3bj AMs

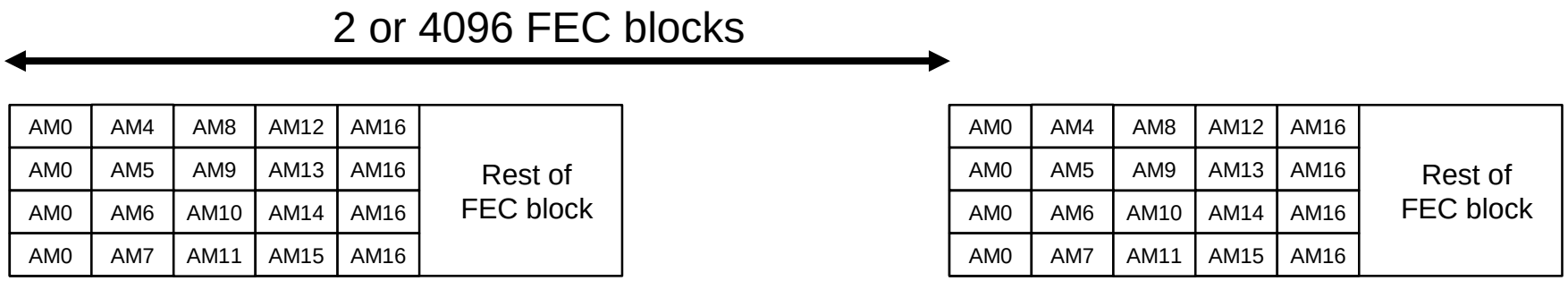
- Clause 91 defines how Alignment Markers are mapped when sent across the 4 FEC lanes
  - They are re-mapped to the FEC lanes so they appear consecutively on a given FEC lanes
  - A 5b pad is added to the end to round make them fit within a even number of 257b blocks (20*64+5 = 257*5)
  - AM0 and AM16 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

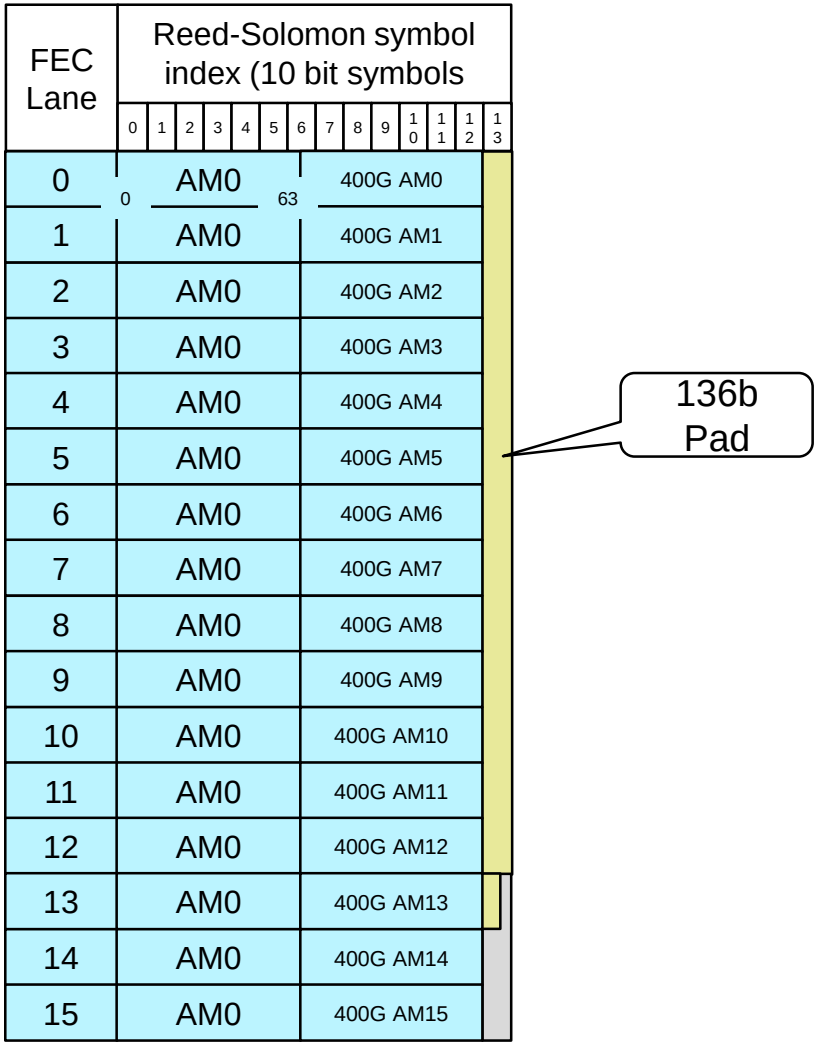
802.3bj AM Distance

- AMs are always aligned to the beginning of an RS-FEC block
- The repetition distance between AMs for normal operation in 802.3bj is once every 4096 FEC blocks
- When sending rapid alignment markers, they are sent every 2 FEC blocks for EEE support



Proposed 400Gb/s AMs

- Re-use AM0 from 802.3ba to allow common block lock between lanes of 100G and 400G, the rest is unique to 400GbE
- Have a 56b 400G unique AM per lane also
  - 56+64 = 120b, allows us to fit within 8 257b blocks evenly
  - Content is TBD



400 Gb/s AM Distance

- AMs are always aligned to the beginning of an RS-FEC block
- Repetition distance is 8192 FEC blocks (2x 802.3bj)

8192 FEC blocks

0	AM0	400G AM0	Rest of the FEC block
1	AM0	400G AM1	
2	AM0	400G AM2	
3	AM0	400G AM3	
4	AM0	400G AM4	
5	AM0	400G AM5	
6	AM0	400G AM6	
7	AM0	400G AM7	
8	AM0	400G AM8	
9	AM0	400G AM9	
10	AM0	400G AM10	
11	AM0	400G AM11	
12	AM0	400G AM12	
13	AM0	400G AM13	
14	AM0	400G AM14	
15	AM0	400G AM15	

0	AM0	400G AM0	Rest of the FEC block
1	AM0	400G AM1	
2	AM0	400G AM2	
3	AM0	400G AM3	
4	AM0	400G AM4	
5	AM0	400G AM5	
6	AM0	400G AM6	
7	AM0	400G AM7	
8	AM0	400G AM8	
9	AM0	400G AM9	
10	AM0	400G AM10	
11	AM0	400G AM11	
12	AM0	400G AM12	
13	AM0	400G AM13	
14	AM0	400G AM14	
15	AM0	400G AM15	

Table Of Contents

PMA Functions

- The following are the functions performed by the PMA sublayer
  - Provide appropriate multiplexing
  - Provide appropriate modulation (PAM4 for instance if required)
  - Provide appropriate coding as needed
    - Gray coding as appropriate
    - Pre-coding as appropriate
  - Provide per input-lane clock and data recovery
  - Provide clock generation
  - Provide signal drivers
  - Optionally provide local loopback to/from the PMA service interface
  - Optionally provide remote loopback to/from the PMD service interface
  - Optionally provide test-pattern generation and detection
  - Tolerate Skew Variation
- Not required
  - Extra overhead such as block termination bits or framing for that termination, even if PAM4 electrical interfaces are chosen

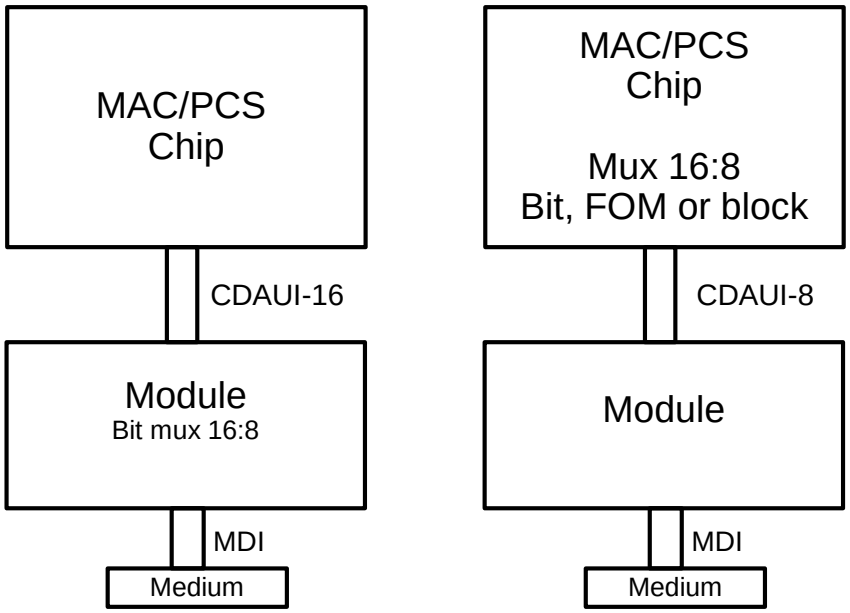
Muxing

Option	Pros	Cons
Bit muxing	Simple	Loses gain with correlated errors
FOM bit muxing	Retain majority of gain even with correlated errors	Lane restrictions, requires 4x100G architecture
FOM pre-interleaved bit muxing	Retain some gain even with correlated errors, no lane restrictions	Lose some gain with correlated errors, requires 4x100G architecture
Block muxing	Retains gain with correlated errors	More complicated, makes modules protocol specific

How does PAM4 impact the muxing decisions?

PMA Multiplexing

- Multiplexing will be needed to go from 16 lanes down to fewer (only in factors of 2)
- When muxing, and if there are no correlated errors, you can bit mux without concern of the FEC block boundaries
- If there are correlated errors, then need to understand the error models to see if we can do bit muxing, or if we need to do FOM or block level muxing
- If we use a 400G FEC vs. 4x100G, that would rule out FOM for muxing



PMA Data Rate

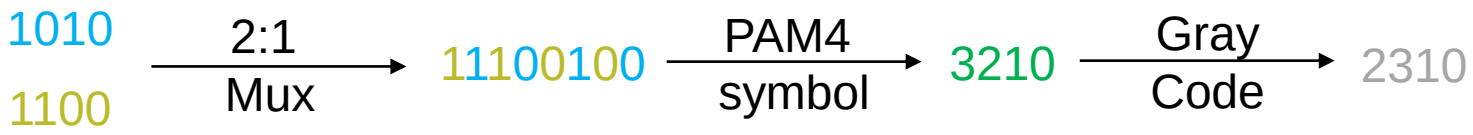
- With KP4 FEC the per lane data rate is:
  - 544/514*257/256*25G = 26.5625G
  - When running 16 lanes
  - When running 8 lanes it is 53.125G per lane

- PLL multiplier from 156.25MHz is 170 for a 26.5625G lane
- This means that SR16 lanes will run 3% faster than the current SR4 lanes

PMA Coding Example

- This is one example of Coding
- Gray mapping prevents more than one bit being in error most of the time

0,0 maps to 0
0,1 maps to 1
1,1 maps to 2
1,0 maps to 3



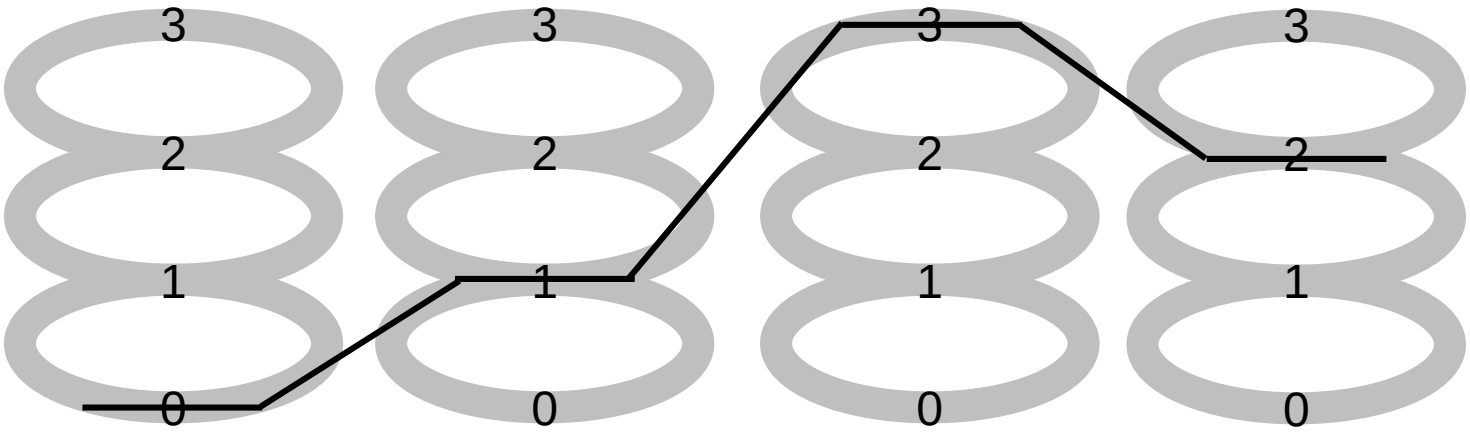


Table Of Contents

Conclusion

- This baseline proposes a single FEC for all PMD and electrical interfaces, focusing on an RS 544 code
- We need to make PMD and electrical interface choices in order to finalize some of the choices of this baseline

Logic Work Items

- Finalize the FEC choice
- 4x100G vs. 1x400G FEC architecture
- Details of the AMs patterns
- What muxing is used for each PMA instance

Thanks!