400GbE PCS and PMA Baseline Proposals

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

July 2015 Hawaii

Mark Gustlin – Xilinx
Arthur Marris - Cadence
Gary Nicholl – Cisco
Dave Ofelt – Juniper
Jerry Pepper – Ixia
Jeff Slavick – Avago
Andre Szczepanek - Inphi
Steve Trowbridge - ALU

Supporters

Ghani Abbas - Ericsson
Pete Anslow – Ciena
Thananya Baldwin - Ixia
Brad Booth – Microsoft
Paul Brooks – JDSU
Matt Brown - APM
Adrian Butter – Globalfoundries
Francesco Caggioni - APM
Juan-Carlos Calderon - Inphi
Wheling Cheng – Ericsson
Don Cober - Comira
Faisal Dada – Xilinx
Piers Dawe – Mellanox
Ian Dedic - Socionext
Dan Dove – DNS
Mike Dudek - Qlogic
Dave Estes – Spirent
John Ewing - Globalfoundries
Eric Fortin - ALU
Adam Healey – Avago
Scott Irwin – MoSys
Jonathan Ingham - Avago
Tom Issenhuth - Microsoft
Jonathan King - Finisar
Martin Langhammer - Altera
Scott Kipp – Brocade
Daniel Koehler – MorethanIP
Kenneth Jackson - Sumitomo
Ryan Latchman - Macom
David Lewis - JDSU
Mike Peng Li - Altera
Jeffery Maki – Juniper Networks
Andy Moorwood – Ericsson
Ed Nakamoto - Spirent
Mark Nowell – Cisco
John Petrilla - Avago
Rick Rabinovich - ALE
Adee Ran – Intel
Sam Sambasivan – ATT
Omer Sella - Mellanox
Ted Sprague – Infinera
Steve Swanson - Corning
Jeff Twombly – Credo Semiconductor
Winston Way - Neophotonics
Brian Welch - Luxtera
Oded Wertheim - Mellanox

References

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

http://www.ieee802.org/3/bs/public/15_03/wertheim_3bs_01a_0315.pdf
http://www.ieee802.org/3/bs/public/15_03/wang_t_3bs_01a_0315.pdf
http://www.ieee802.org/3/bs/public/15_03/trowbridge_3bs_01_0315.pdf
http://www.ieee802.org/3/bs/public/15_03/gustlin_3bs_02a_0315.pdf
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

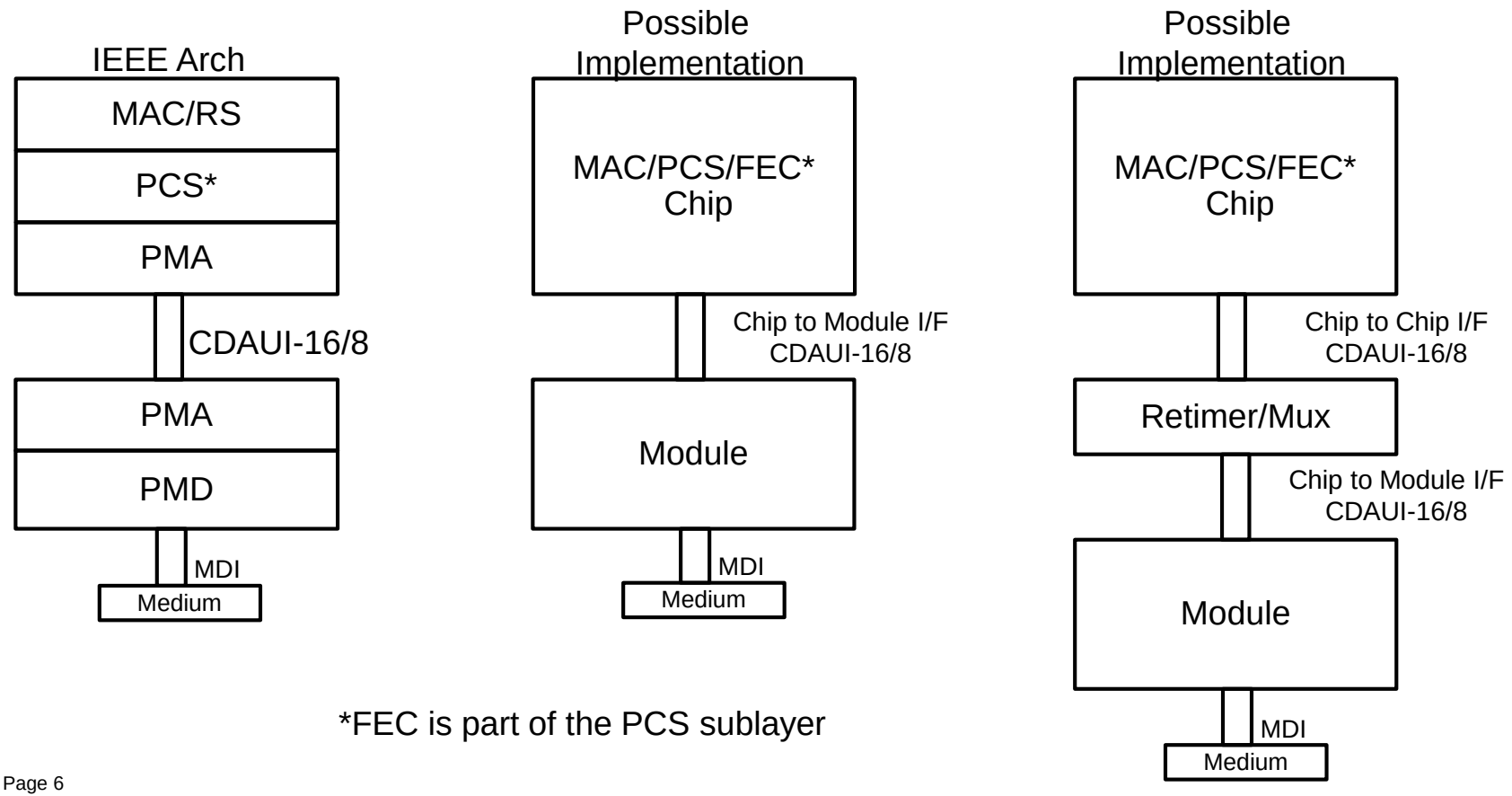
Table Of Contents

Introduction

- This looks at a baseline PCS and PMA proposal based on a 1x400G FEC architecture

PCS Architecture

- Based on the adopted system architecture
- A single FEC is used, across up to 5 interfaces (in the PCS sublayer)
- CDMII is an optional interface that is not shown in these figures, but is already adopted and may be present in a given implementation

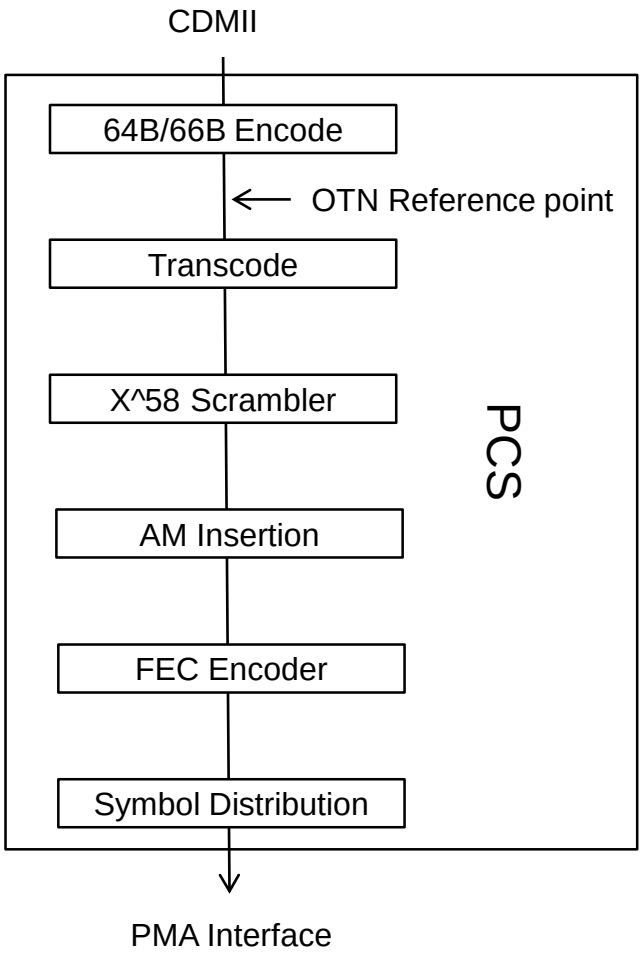


*FEC is part of the PCS sublayer

Table Of Contents

Proposed TX PCS Data Flow

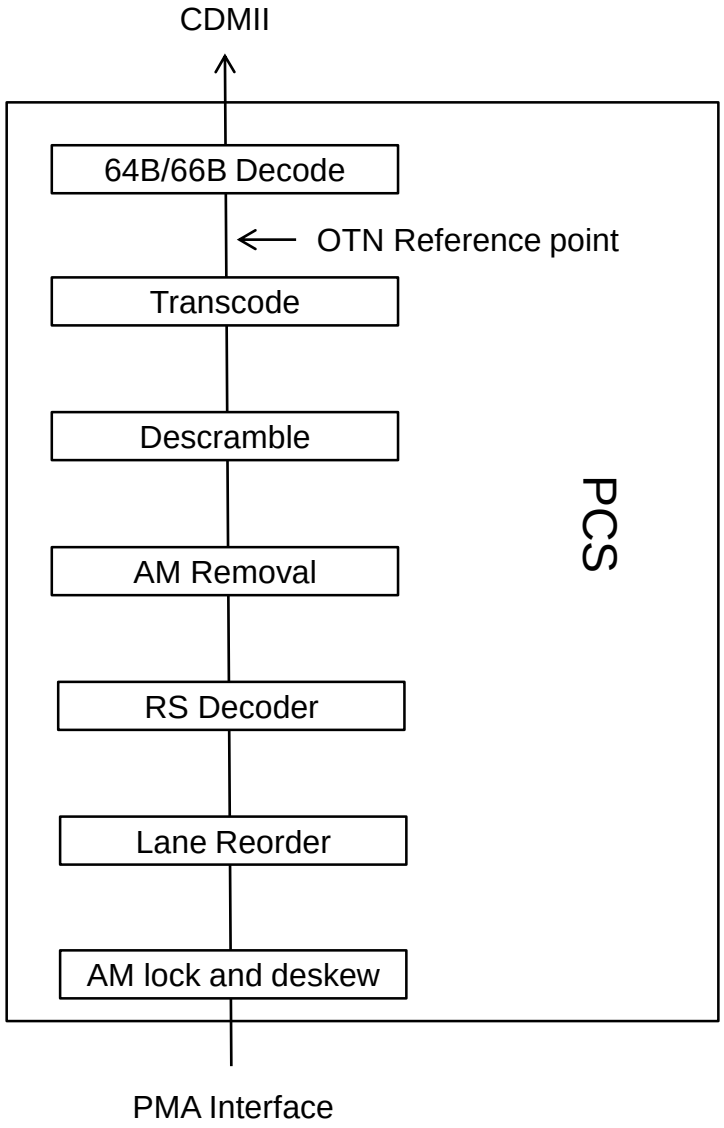
- 64B/66B encode based on clause 82
- Transcode to 256B/257B based on clause 91
- Scrambler is moved to after the Transcoding to simplify the flow
- FEC Encoder is RS(544,514,10), ~~in a 1x400G architecture~~
  - All FEC processing is as in clause 91, including error correction and detection modes
  - Method of forming PCS lanes from FEC codewords is TBD and dependent on further error analysis
- Location of the OTN reference point is as shown and adopted in the January meeting
- Support for any logical lane on any physical lane



Note: Updated and with more detail from what was adopted in dambrosia_3bs_02b_0115.pdf

Proposed RX PCS Data Flow

- Reverse of TX
- Allows for arbitrary lane arrival



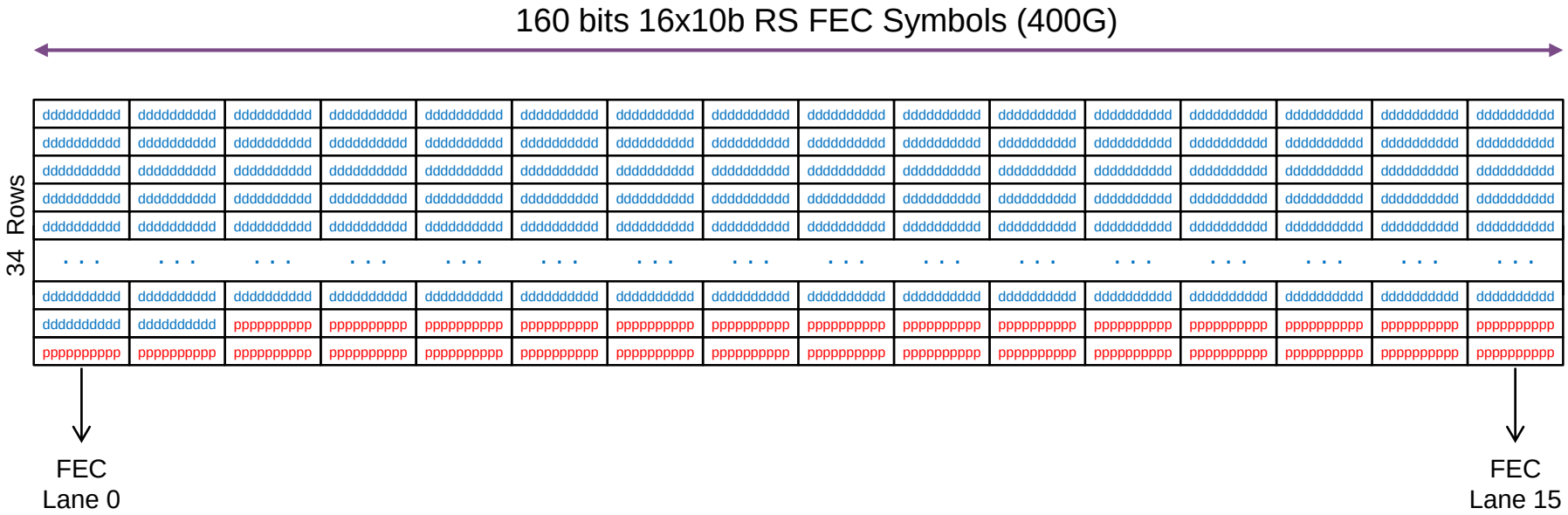
Scrambling

- 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

400GbE Data Distribution – 1x400G

- Below the RS-FEC sublayer, using 1x802.3bj KP4 FEC (400G single FEC instance), you would naturally have 16 FEC lanes
  - ddddddddddd = protected data (5140 bits total)
  - ppppppppppp = FEC Parity addition (300 bits total)
  - d + p = 5440 bits total



400GbE 257b Block Mapping

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

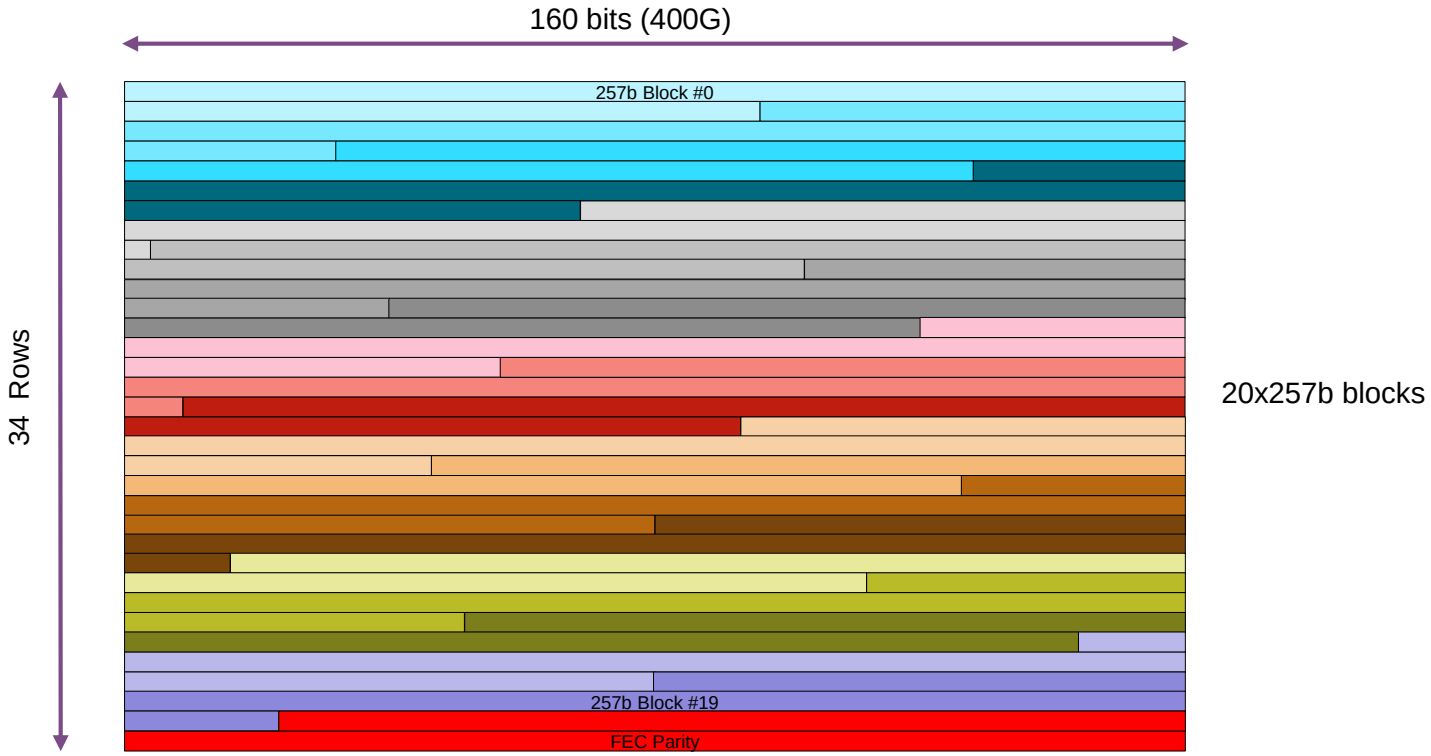
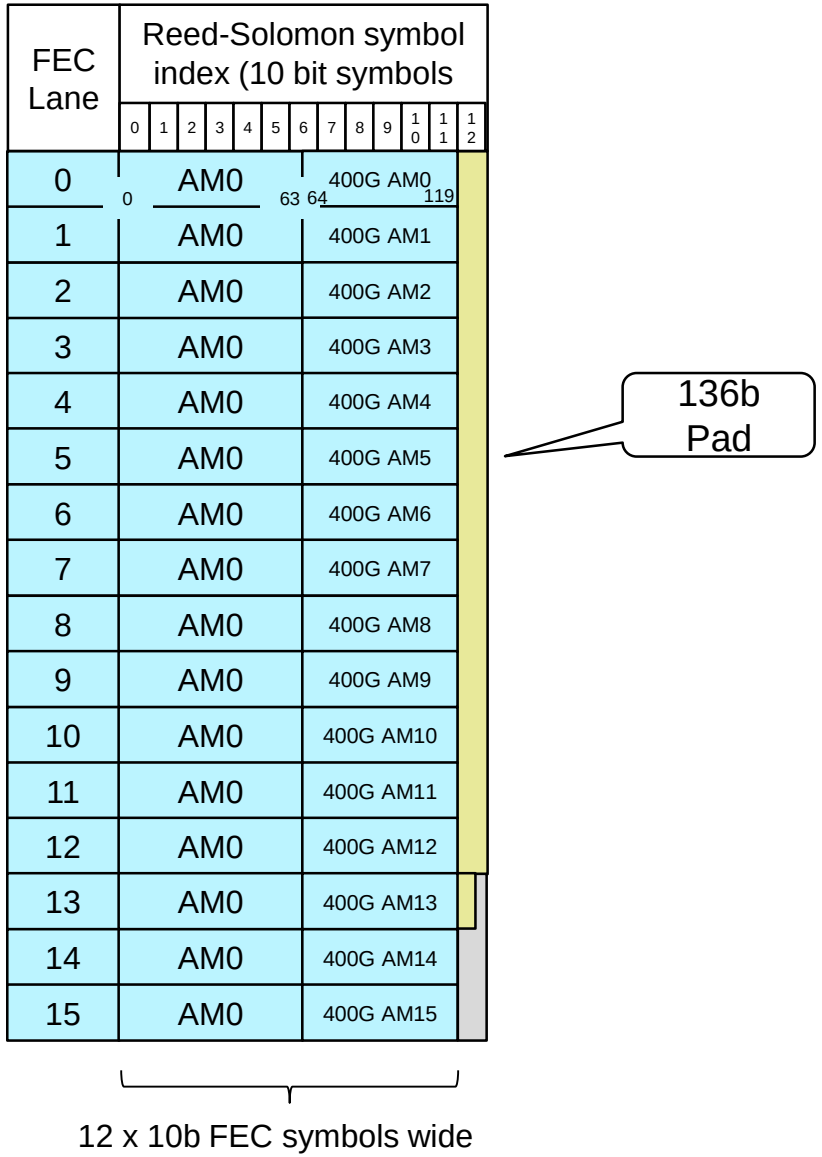


Table Of Contents

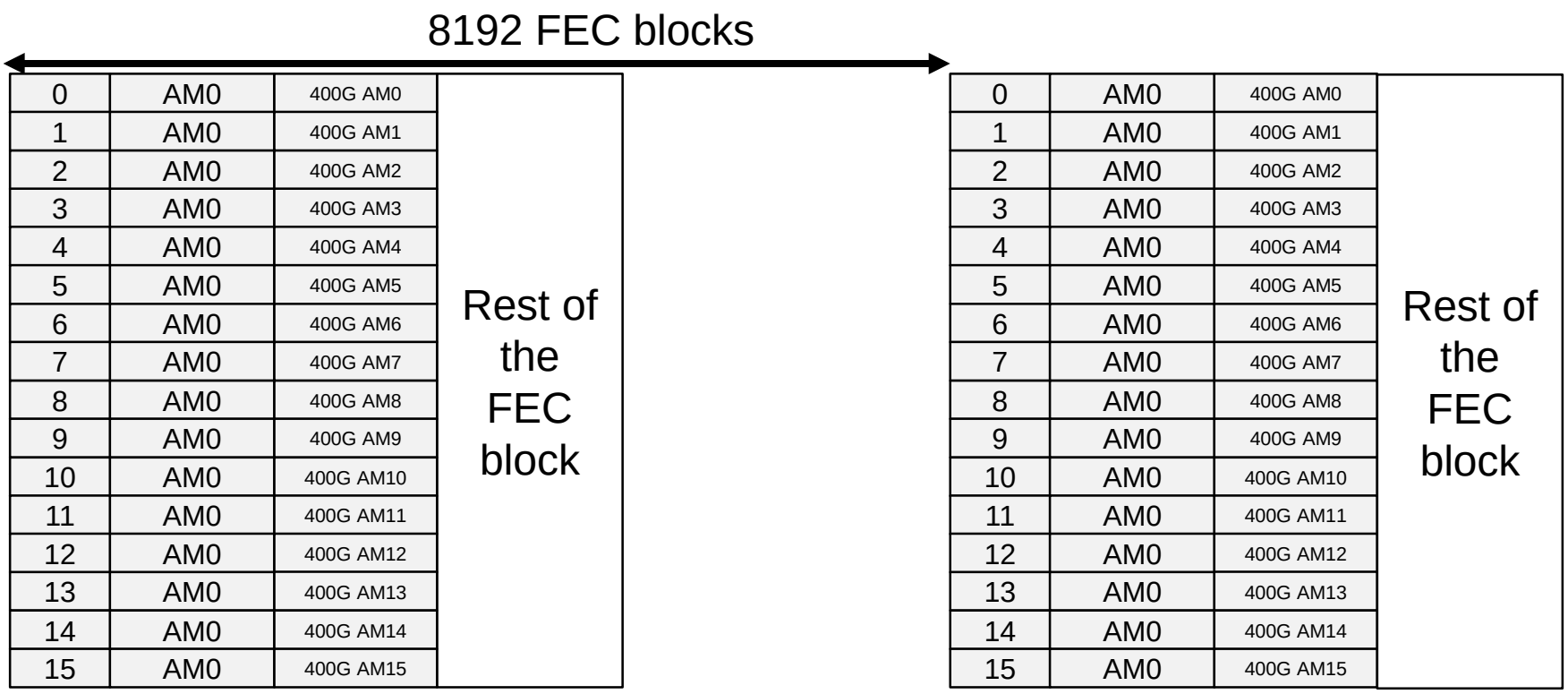
Proposed 400Gb/s AMs

- Re-use 100G 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, putting 120b on each FEC lane after RS symbol distribution requires 8x257b blocks
  - The pad allows us to fit evenly within 8x257b blocks to ease processing
  - Content of 400G AMx 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)
  - This works out to a little less overhead than we have at 100GbE (~49PPM vs. ~61PPM)



Proposed 400Gb/s AM Detail

- Original AM0 contents: 0xC1, 0x68, 0x21, BIP3, 0x3E, 0x97, 0xDE, BIP7
  - Put what we originally had instead of the BIP fields back in the 802.3ba days
  - 0xC1, 0x68, 0x21, 0xF4, 0x3E, 0x97, 0xDE, 0x0B
- 400G AM0, AM1 etc. contents (56b)
  - Create a 28b unique AM field for each marker
  - 2nd 28b is just the bit inversion of the first 28b to keep balance
  - Anything else we need to add in/carry?
- What goes in the 136b pad?
  - Fill in with free running PRBS9 pattern which continues running from frame to frame. X^9+x^5+1

Table Of Contents

PMA Functions

- The following are the functions performed by the PMA sublayer
  - Provide appropriate multiplexing
  - Provide appropriate modulation (NRZ/PAM4)
  - Provide appropriate coding as needed
    - Gray coding as appropriate (for the PAM4 electrical interface for instance)
  - Provide per input-lane clock and data recovery
  - Provide clock generation
  - Provide signal drivers when applicable
  - 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

Note: Updated list from what was adopted in dambrosia_3bs_02b_0115.pdf

PMA Multiplexing

- The PMA will support bit muxing, without regard to skew or PMA lane identity
- Total skew is handled in the RX PCS
  - Skew budgets are TBD (variation and total skew)
  - Skew variation must be handled by PMA instances that do bit muxing, same as in the 802.3ba architecture

PMA Data Rate

- With KP4 FEC the per lane signaling 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

Testing Concerns

- Propose to continue the use of scrambled idles as the PCS test pattern
  - Defined in clause 82.2.10 and 82.2.17

- Support PCS loopback, TX MII data is looped back to the RX MII and transmitted towards the PMA

Table Of Contents

Work Items

- Hi BER, use FEC thresholds?
- Sublayer delay constraints are TBD, same with skew limitations
- Define 400G AM fields
- Precoding?
- Exact criteria for achieving AM lock

Conclusion

- This proposal implements the adopted RS(544,514,10) as the FEC in the 802.3bs 400GbE PCS using a 1x400GbE architecture
- Additional PMA details are included, including a bit muxing architecture

Thanks!