400GbE PCS Baseline Proposal

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

January 2015 Atlanta

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

Introduction

- This looks at a baseline PCS proposal, there are still many open issues (FEC etc.)

References

- 400G PCS and FEC options:

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

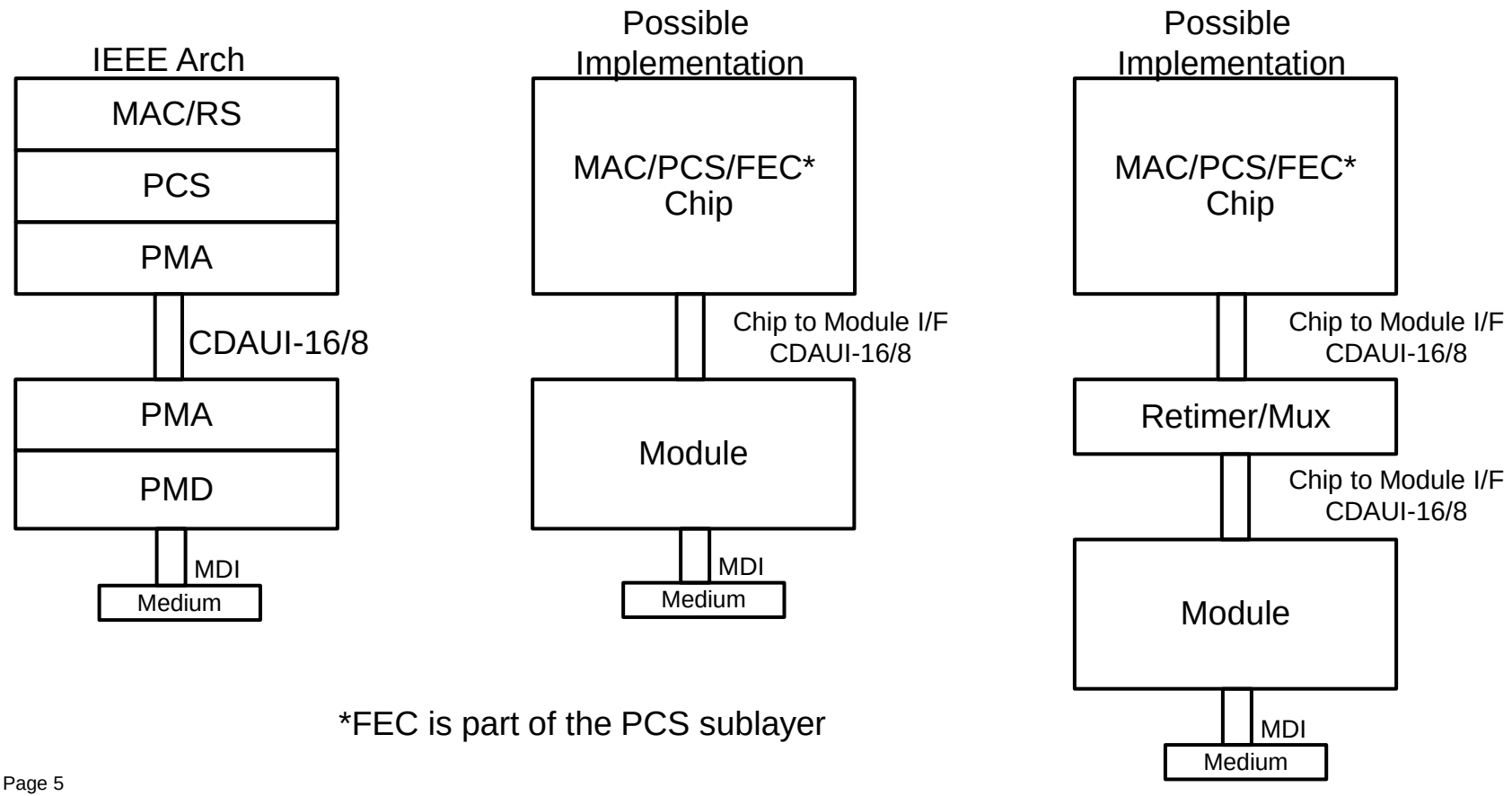
Review of the Sublayer Functions

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

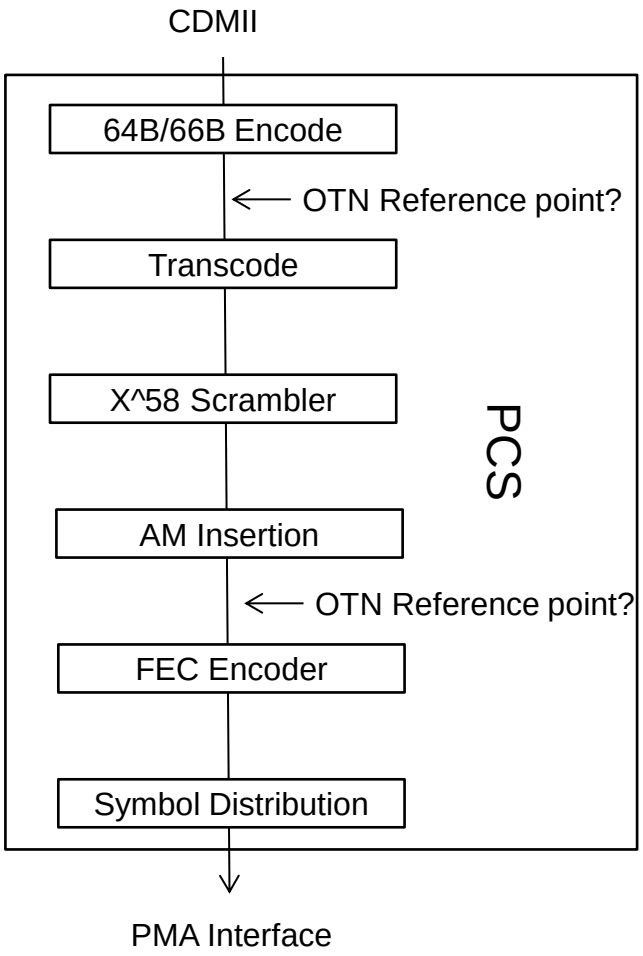
- Likely implementation options, 16 lane CDAUI interfaces first and then 8 lane interfaces later
- You can mix the two, just PMA Muxing to go back and forth
- In this instance a single FEC is used, across up to 5 interfaces (in the PCS sublayer)
- Assuming a single FEC covers up to 5 interfaces



*FEC is part of the PCS sublayer

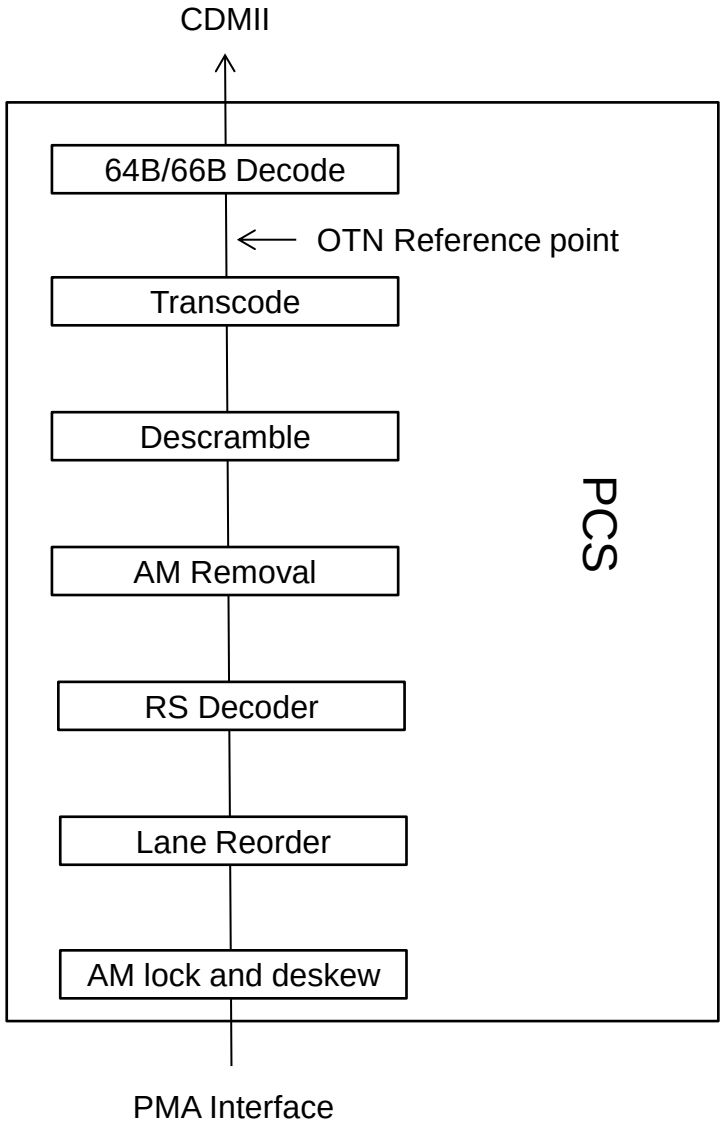
Proposed TX PCS Data Flow

- 64B/66B encode based on clause 82
- Propose that we do not direct encode to 256B/257B from a standards point of view, but implementations are free to do that
- Scrambler is moved to after the Transcoding to simplify the flow
- FEC Encoder could support KP4 FEC RS(544,514,10) or some other stronger FEC
- Open question if we should have a single 400G FEC or 4x100G
  - 4x100G has obvious re-use for 100GbE
  - Would the KP4 FEC be useful for mainstream 100G interfaces?
  - KP4 and KR4 FEC can share and encoder/decoder
- 16 PMA lanes (similar to PCS/FEC lanes)
- Location of the OTN reference point is TBD, some options are shown to the right, this is further explored in another presentation



Proposed RX PCs Data Flow

- Reverse of TX
- Allowing for arbitrary lane arrival

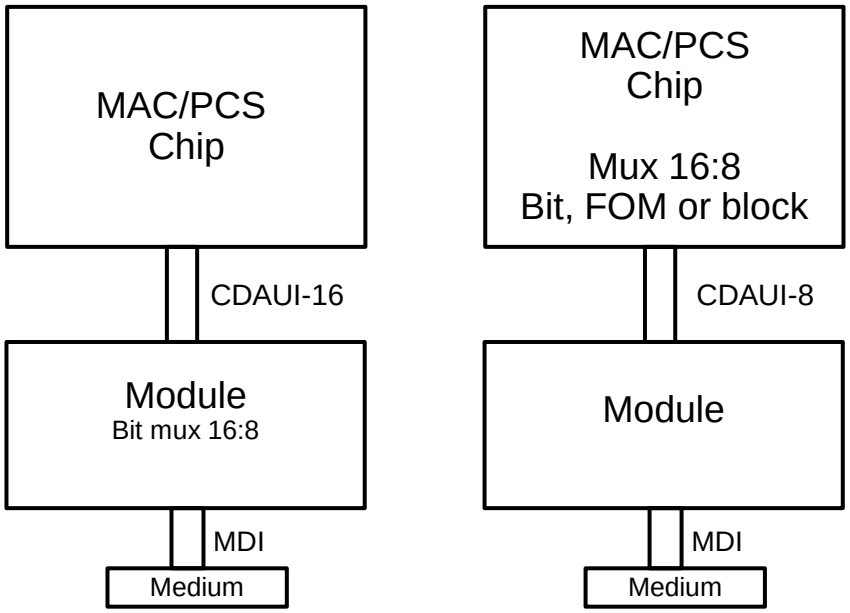


PMA Functions

- The following are the functions performed by the PMA sublayer
  - Provide appropriate multiplexing
  - Provide appropriate modulation (PAM4 for instance if required)
  - 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

PMA Multiplexing

- Multiplexing will be need 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
- Another presentation explores if PAM4 at 25GBaud would have issues with bit muxing



EEE

- It is assumed that the basis for the EEE implementation is based on 802.3bm, fast wake only

Work Items

- What FEC will be used, or even possibly multiple FECs
- 4x100G vs. 1x400G FEC
- What do AMs look like
- Details of the scrambling process, exactly what is scrambled and how
- What muxing is used for each PMA instance
- Details around EEE operation

Conclusion

- Having a single FEC will greatly simplify systems, but the right tradeoff between FEC gain and complexity/power needs to be made so that we can possibly include a single FEC in the large ASIC/FPGA/ASSP
- Other presentations at this meeting further explore FEC sizing concerns

Thanks!