400 GbE Architectural Considerations

400 GbE Study Group

Ali Ghiasi, Eric Baden, and Zhongfeng Wang

Broadcom Corporation

Sept 2, 2013

York, UK

Overview

- Investigated 400 GbE virtual lane choices and implication to support dual mode 100G/400GbE ports
  - 5G virtual lanes
  - 25G virtual lanes
- FEC requirements and complications
  - Due to evolving PMDs
  - Dual mode operation

400 GbE Architectural Consideration and How it Can Address 5 Criteria

- Broad Market Potential
  - A key consideration in the 400 GbE architecture should be compatibility with 100 GbE and MLG, which will be the key driver for 400G volume for any years to come
- Economic feasibility
  - Increase overall investment available and lower the cost
- Technical feasibility
  - Proposed architecture can even be based on existing 100 GbE PMDs and FEC
- Distinct identity
  - 400 GbE is distinct
- Compatibility
  - Compatibility with 100 GbE PMDs and FEC will amortize the investment and will accelerate the market deployment.

400 GbE Needs to follow Foot Step of 10/40 GbE

- 10/40 GbE are example of great market success
  - 10 GbE serial optical PMD were based on single lane of 64/66B encoded “10Gbase-R PCS”
    - Volume fiber deployment are 10Gbase-SR/LR
    - Volume Cu deployment is 10GSFP+ Cu “DAC” defined in SFF-8431 project
    - Electrical signaling for serial PMD implementation is “SFI” and was defined in SFF-8431 project
    - Electrical swing for 10GSFP+ for Cu and optical PMDs are identical
  - 40 GbE volume PMD deployment are based on 4-lanes of 64/66B encoded with MLD “40Gbase-R PCS”
    - Volume fiber deployment are 40Gbase-SR4/LR4
    - Volume Cu deployment is 40Gbase-CR4
    - Electrical signaling for 40Gbase-SR4/LR4 is very similar to 10G SFI
    - Electrical signaling for 40Gbase-CR4 leverages larger amplitude similar to 10GBase-KR with 800 mV swing a minor nuisance compare to 10GSFP+DAC and CL86 nPPI
- The key to phenomenal success of 10/40GbE is the ease of implementing dual mode ports and the availability of fiber and Cu break out to go from QSFP+ to 4 SFP+ modules.

PCS Consideration for 400 GbE

- Two primary choice for PCS

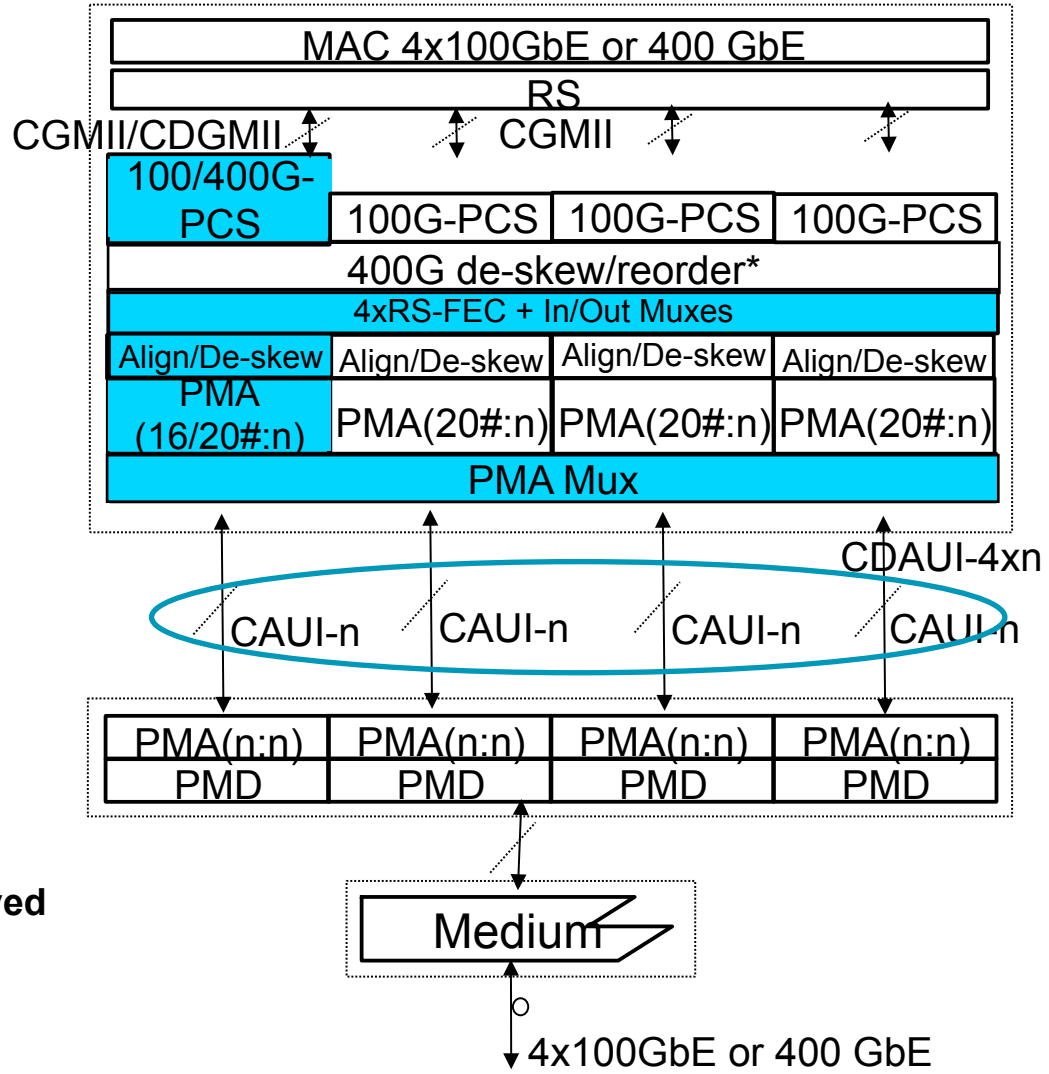
http://www.ieee802.org/3/400GSG/public/adhoc/logic/aug20_13/begin_01_0813_logic.pdf

  - 80 PCS lanes (5G VLs)
    - Compatibility with 100 GbE and MLG
    - Lane re-order complexity for 400 GbE
      - If re-order is required across 4 slices of 100G # 2:1 muxes=41712
      - But if re-order is within each slice of 100G # 2:1 muxes=10032*
  - 16 PCS lanes (25G VLs)
    - Reduce complexity of 400 GbE PCS and the PMA mux eliminates the MTTPFA issue
    - Lane re-order complexity for 400 GbE
      - If re-order is required across 4 slice of 100G # 2:1 muxes=8448
      - But if re-order is within each slice of 100G # 2:1 muxes=2112*
      - To mux FEC and PMA # 2:1 muxes required = 2112

- There is no doubt 16 PCS lanes is a better choice for 400 GbE if 100 GbE/MLG were not a factor!

*If re-order is within each slice then FEC[0:3] need to be connected to PMD[0:3] in order. Also future PMDs based on serial 200G or 400G would require re-order be implemented in the PMD if it is not in the PCS.

100/400 GbE Dual Mode Implementation Based on 16 or 80 VLs



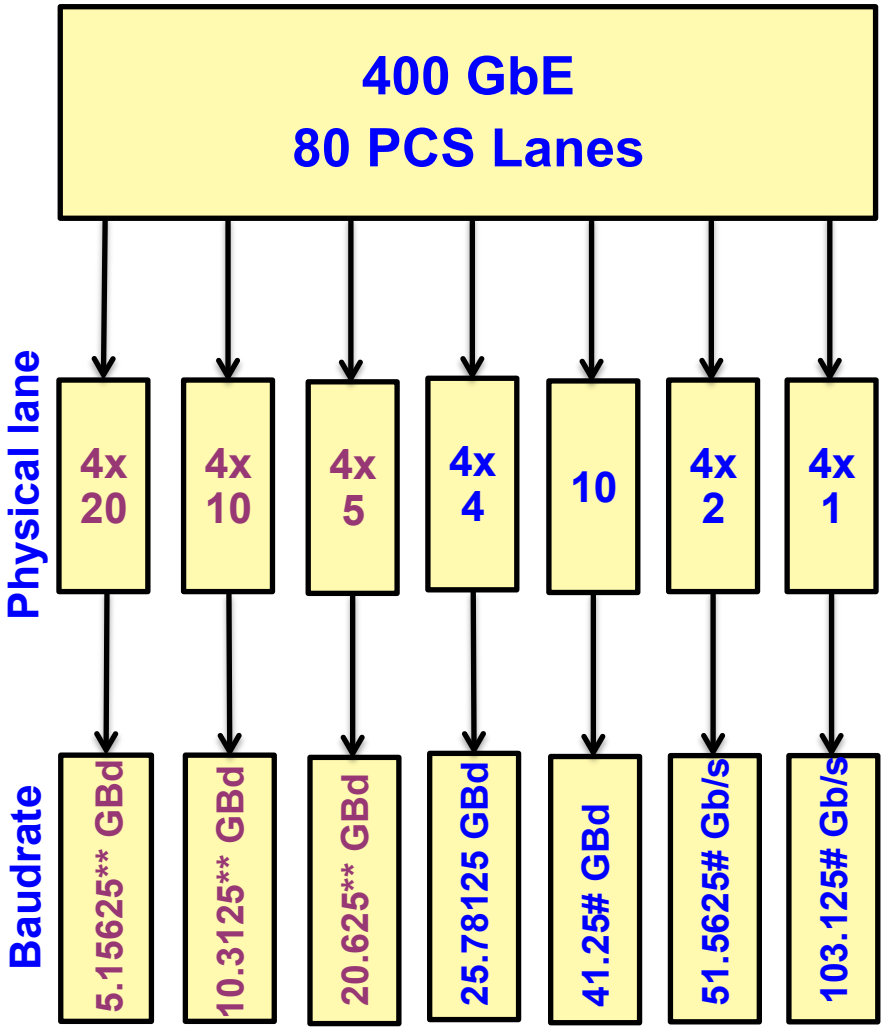
*** Reorder may not be required**

if FEC to PMD lanes are preserved

 # If 100G RS-FEC enabled then PMA will be (4:n)

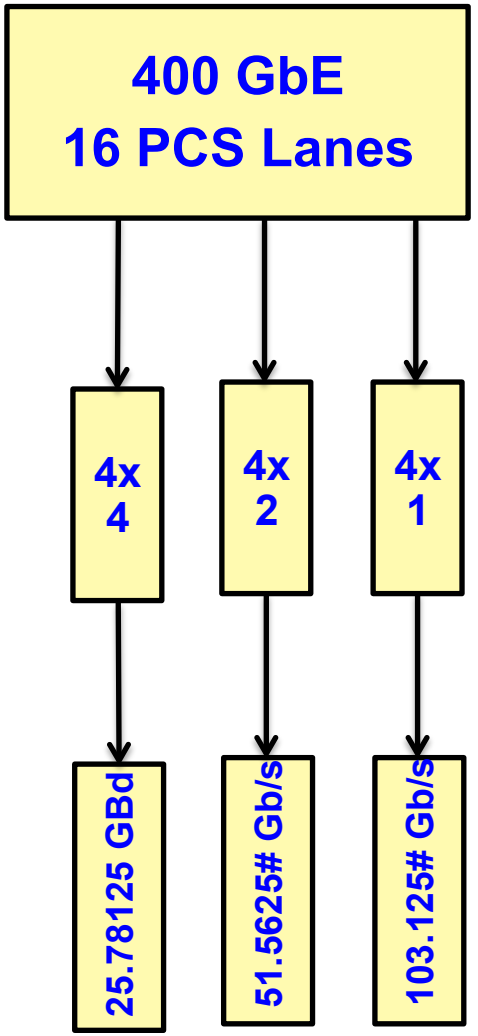
400 GbE PCS and Possible Physical Lanes

- Advantage of 80 PCS lanes
  - Integration with 100 GbE PCS is more stream line and better flow
  - Reuse the BJ RS-FEC (528,514) without needing to adjust AM distance
  - Compatible with MLG
  - Can support more easily 10x40 GbE
- Disadvantage of 80 PCS lanes
  - If re-order across 80 PCS lanes then 2:1 muxes increases by 5x but negligible compare to FEC size
  - MTTPFA will also exist for 16x25G PMA with DFE receiver



400 GbE PCS and Possible Physical Lanes

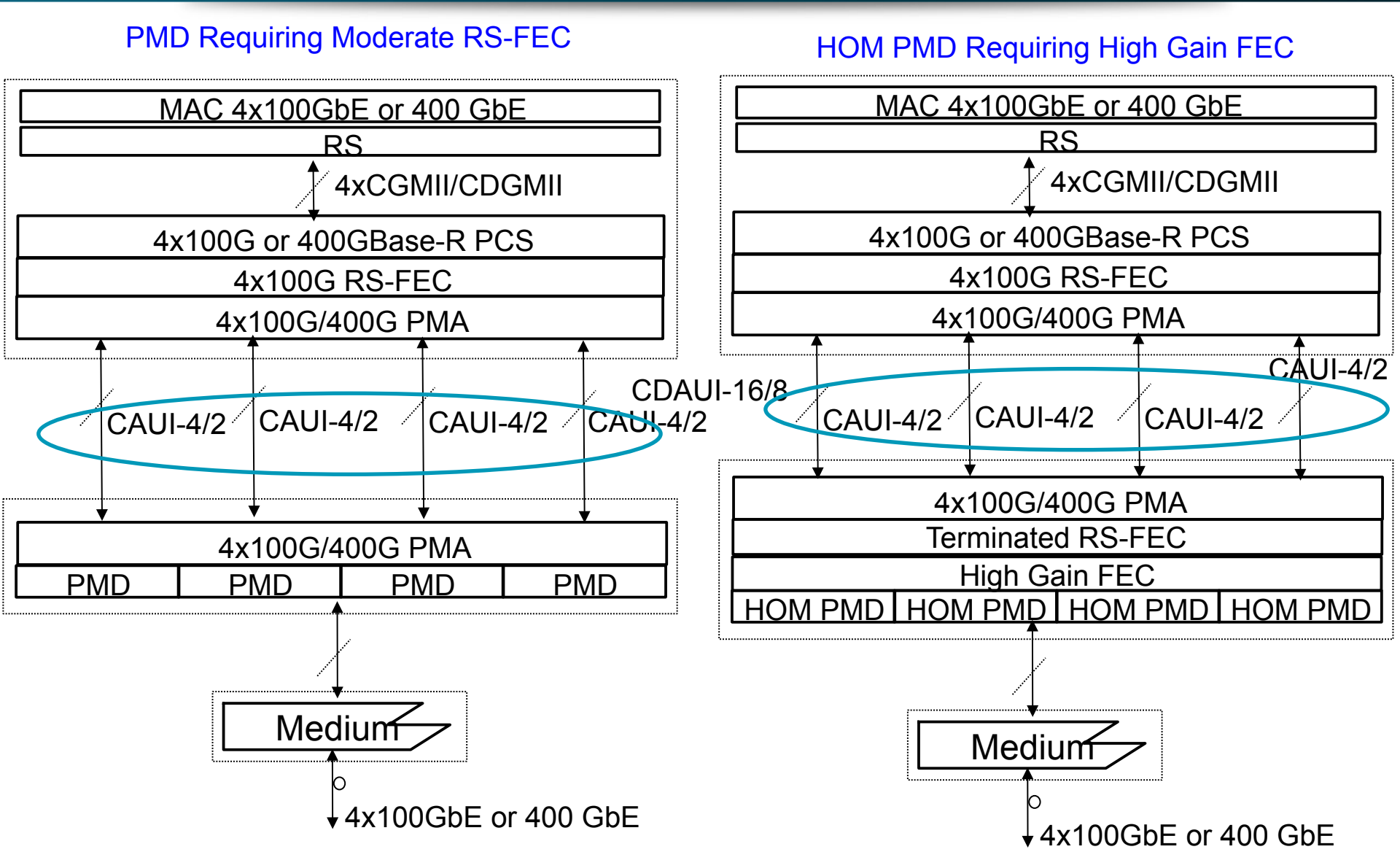
- Advantage of 16 PCS lanes
  - Re-order is 5x less complex
  - With simple PMA mux the MTTPFA issue goes away for DFE receivers without FEC
  - Assuming RS-FEC is salvaged by changing the distance between alignment markers (AMs) then biggest drawback of using 16 PCS lanes is addressed
- Disadvantage of 16 PCS lanes
  - All the logic to support 80 VLs would be there for 100 GbE and MLG with exception of re-order/de-skew across 80 lanes
  - 40 GbE would require MLG inverse-gearbox or the port must operate at 8x41.25 GBd
  - Dual mode architecture require PMA muxes
  - Possibility another RS-FEC is defined to be better compatible with 16 PCS lanes.



What to do in regard to FEC?

- Any dual-mode 100/400GbE port will have an RS(528,514) FEC to support 100Gbase-SR4/CR4/KR4
- Is there a use case for RS(528,514) for 400 GbE?
  - Gen0 PMDs (25 Gb/s/lane) such as 400Gbase-SR16/CR16 (4x100Gbase-SR4/CR4) will need RS(528,514) FEC or another FEC with ~ 5.8 dB of gain
  - Gen1 PMDs (50 Gb/s/lane) such as 400Gbase-SR8/CDAUI-8 may also use the RS(528,514) FEC
  - Gen2 PMDs (100 Gb/s/lane) such as 400Gbase-LR4/PSM4 based on higher order modulation (HOM) such as PAM/DMT likely requires PMD specific high gain FEC
- How to proceed on generational FEC decision?
  - If Gen0 PMDs to be deployed then RS(528,514) FEC should be defined
    - If we go with 80 PCS lanes exact same RS-FEC can be used but if we go with16 PCS lanes then AM spacing need to be adjusted per method http://www.ieee802.org/3/400GSG/public/13_09/wang_400_01_0713.pdf
  - If Gen1 PMDs to be defined, RS(528, 514) likely will meet application needs
  - If Gen2 PMDs to be defined, high gain PMD specific FEC is required that needs to be defined as part of the PMD definition

Host FEC and PMD Specific FEC Architecture



Summary

- We have an opportunity in the 400 GbE project to define an scalable architecture which can address high density 100 GbE as well as 400 GbE applications
  - There is no doubt 16 PCS lanes is a better choice for 400 GbE if 100 GbE/ MLG were not a factor
  - 80 PCS lanes allow more stream line integration with 100G port but gate count for 80 PCS implementation is higher but negligible compare to FEC block
  - An scalable 400 GbE port can be created at reasonable cost for both 16 VLs and 80 VLs as long we can save the FEC
- Key decision for us to move forward is what to do in regard to FEC with evolving PMDs and with knowledge that RS(528,514) will be available in every port to support 100 GbE
  - If we are going to use some of the 100 GbE PMDs “Gen0” then the decision is clear we need RS-FEC with the same gain
  - The same RS-FEC likely will be enough for CDAUI-8 as well as optical PMDs at 50 Gb/s
  - We will likely require high gain PMD specific FEC for PMDs operating at 100 Gb/s.