Scalable 400 GbE Architecture

400 GbE Study Group

Ali Ghiasi, Rishi Chugh, Eric Baden, and Zhongfeng Wang

Broadcom Corporation

July 16, 2013

Geneva

How the Scalable 400 GbE Architecture Can Address 5 Criteria

- Broad Market Potential
  - An scalable 100/400G architecture will accelerate 400 GbE port deployment
- 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.

What can we learn from 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.

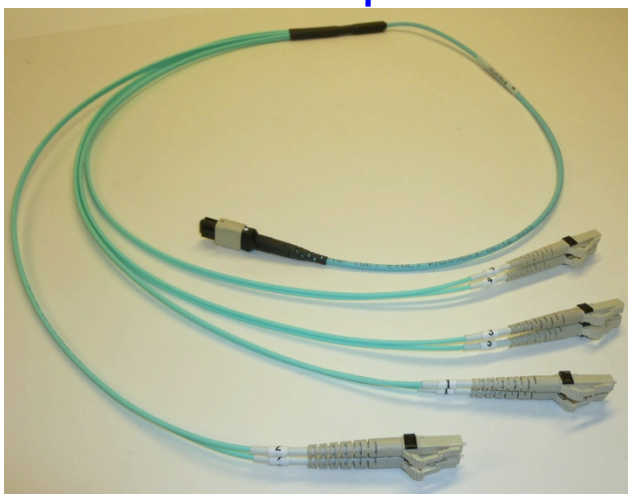
How Dual 10/40GbE Has Served the Marketplace

- 2nd generation QSFP+ module can support 40Gbase-SR4, interoperate with 10GBase-SR, and support 300 m reach
- QSFP+ ports can support 40Gbase-CR4, interoperate with 10GSFP +Cu DAC, and support 5 m Cu reach
- A 19” rack can support 36 QSFP+ stacked ports of 40GbE or 144 ports of 10 GbE
  - In the early days of 40 GbE, 10 GbE volume drove dual ports development
  - QSFP+ ports serve as 10/40 GbE flexible ports for uplinks as well as connecting to large number of servers with availability of hybrid jumper as shown below

QSFP+ to 4 SFP+ Cu DAC

Picture Courtesy of TE

MTP to 4 Duplex LC



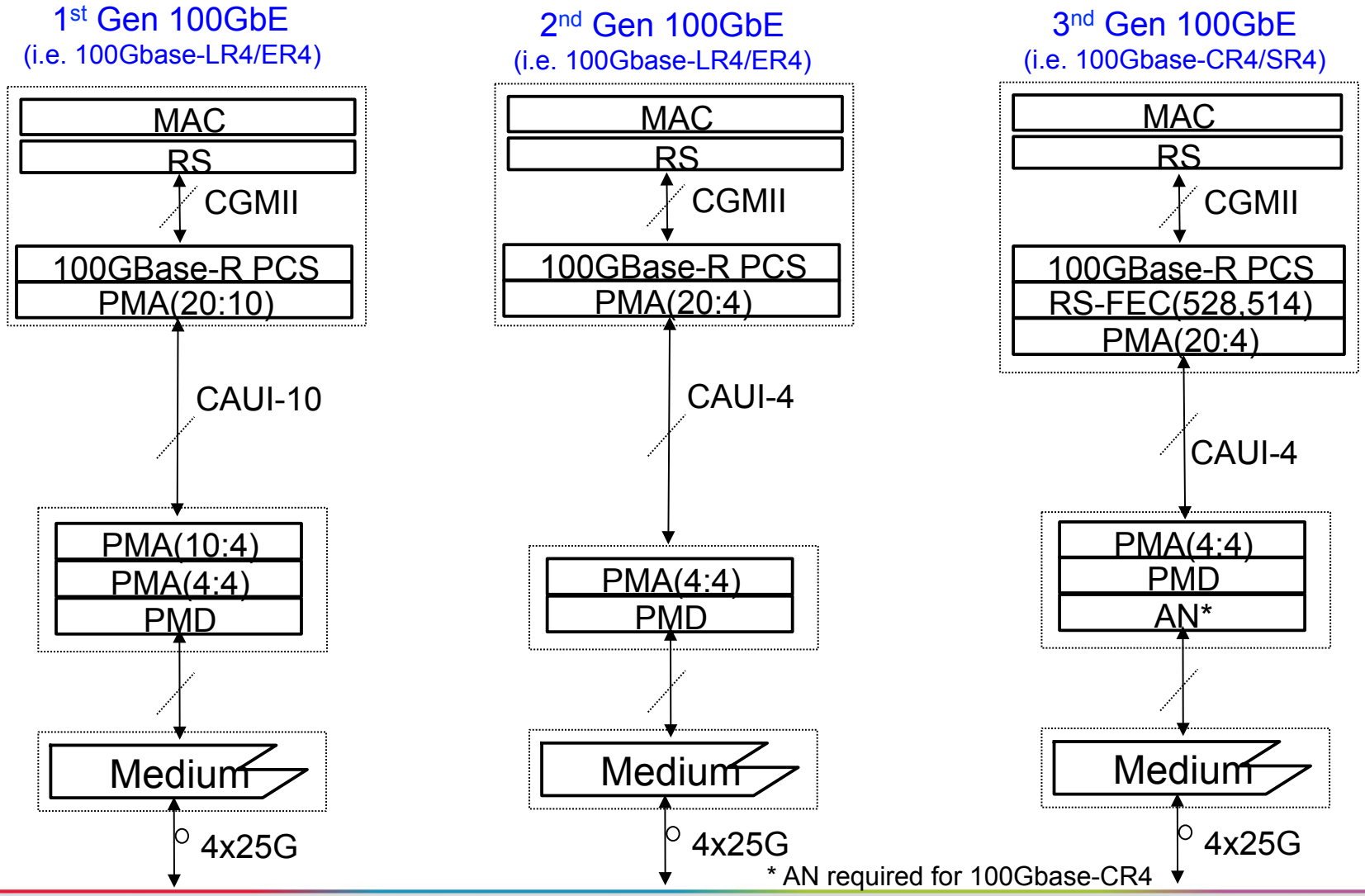
Picture Courtesy of TE

Options for Starting with Clean Sheet of Paper

- Change PMA from simple bit multiplexing to block multiplexing to mitigate MTTFPA as result of DFE burst error if there is no FEC http://www.ieee802.org/3/bj/public/may12/cideciyan_01_0512.pdf
  - FEC is mandatory for all PMD defined in 802.3bj/bm so the issue is moot
  - A 64/66B DFE link operating without FEC would require link BER ~1E-17 to have MTTFPA equal to life of universe
    - Changing PMA definition will make it incompatible with the only 100G links defined without FEC and DFE, 100Gbase-LR4/ER4, so it is moot
  - If the PMA definition is really broken then it needs to be fixed in CL80
- The most important decision for us is: reuse/instantiated RS-FEC 4 times or alternatively create 400 GbE specific FEC, these options are explored in http://www.ieee802.org/3/400GSG/public/13_07/wang_400_01_0713.pdf
  - The expected initial 400 GbE PMDs are likely CDAUI and SR16 where BJ RS-FEC (528,514) with gain of ~5.8 dB is sufficient
  - Higher order modulation (HOM) such as PAM/DMT likely require PMD specific FEC due to complexity and gain required
  - Stripping FEC block across 400G vs 100G reduces RS-FEC (528,514) latency from ~90 ns to ~45 likely too little force every 100/400G switch/ASIC to integrate two dedicated FEC!

100 GbE Layer Diagram Evolution

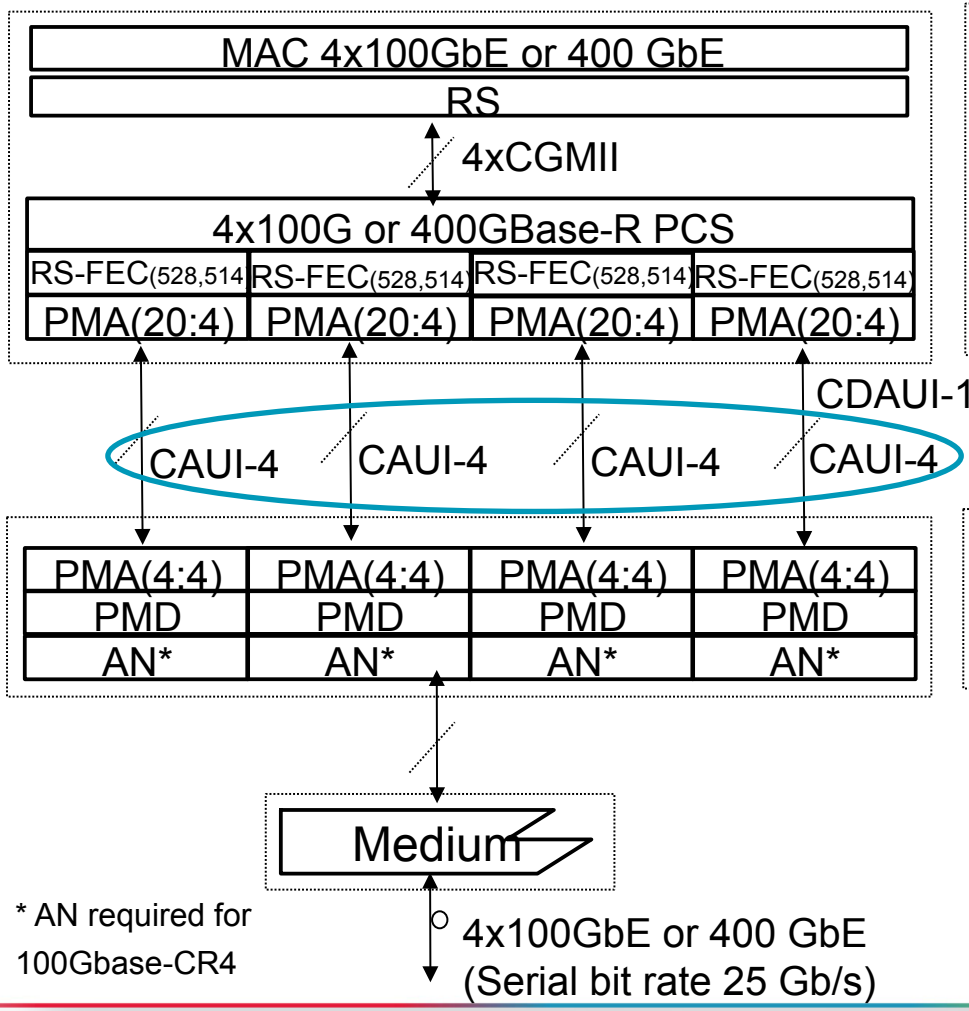
- Seamless evolution with introduction of 100 GbE PMDs



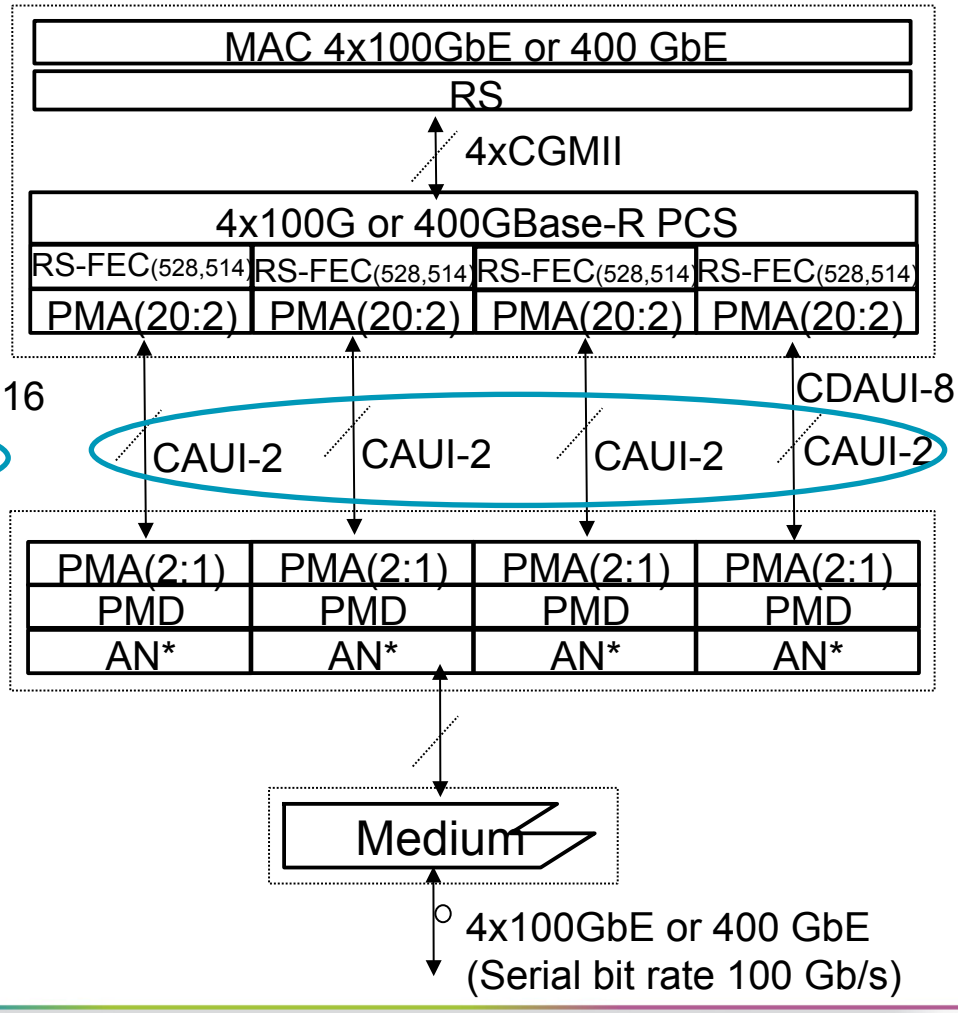
Scalable 100/400 GbE Implementation

- Many blocks will be common for 100 GbE and 400 GbE

1st Gen 4x100GbE/400 GbE

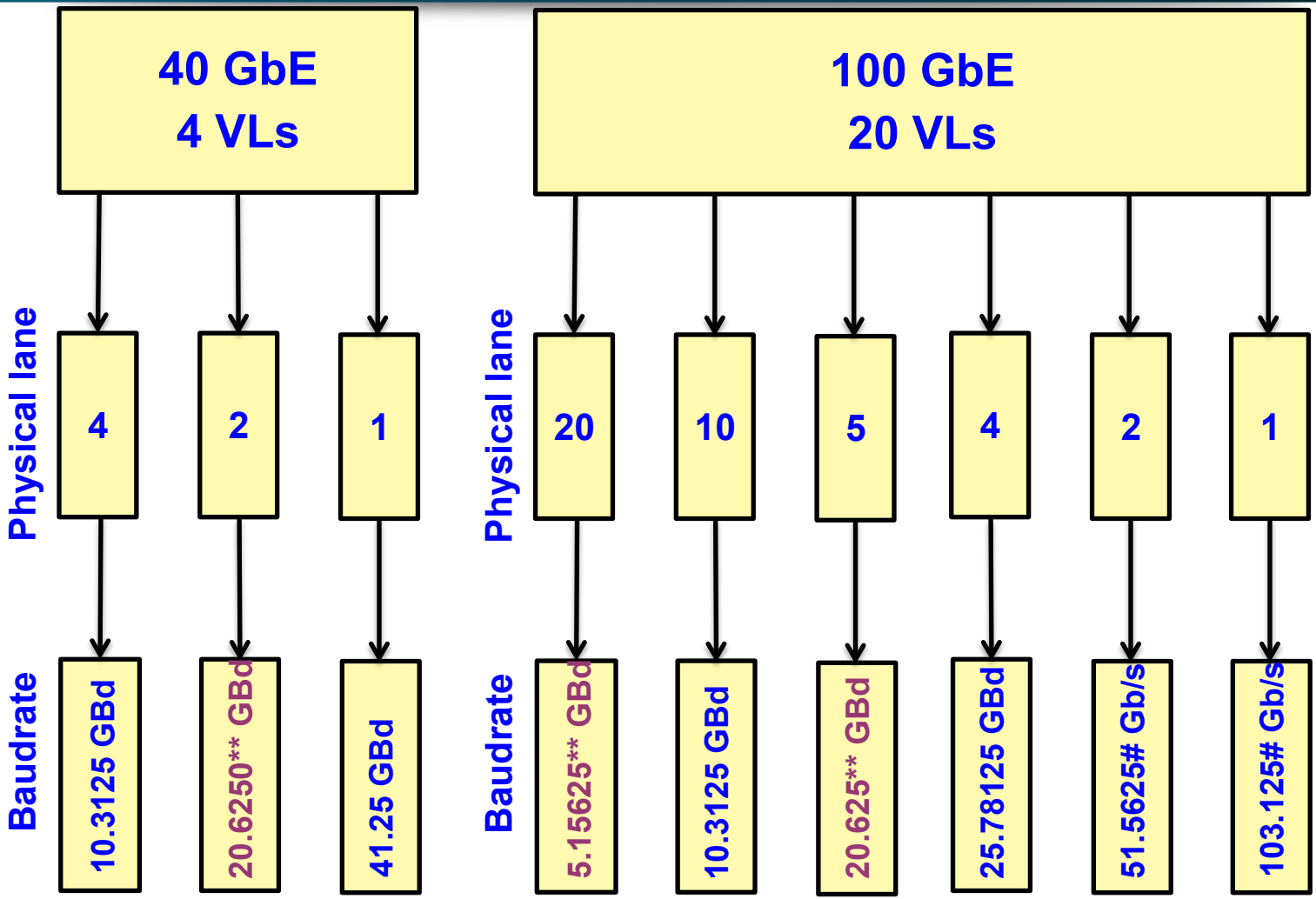


2nd Gen 4x100GbE/400 GbE



* AN required for 100Gbase-CR4

40/100 GbE PCS and Possible Physical Lanes



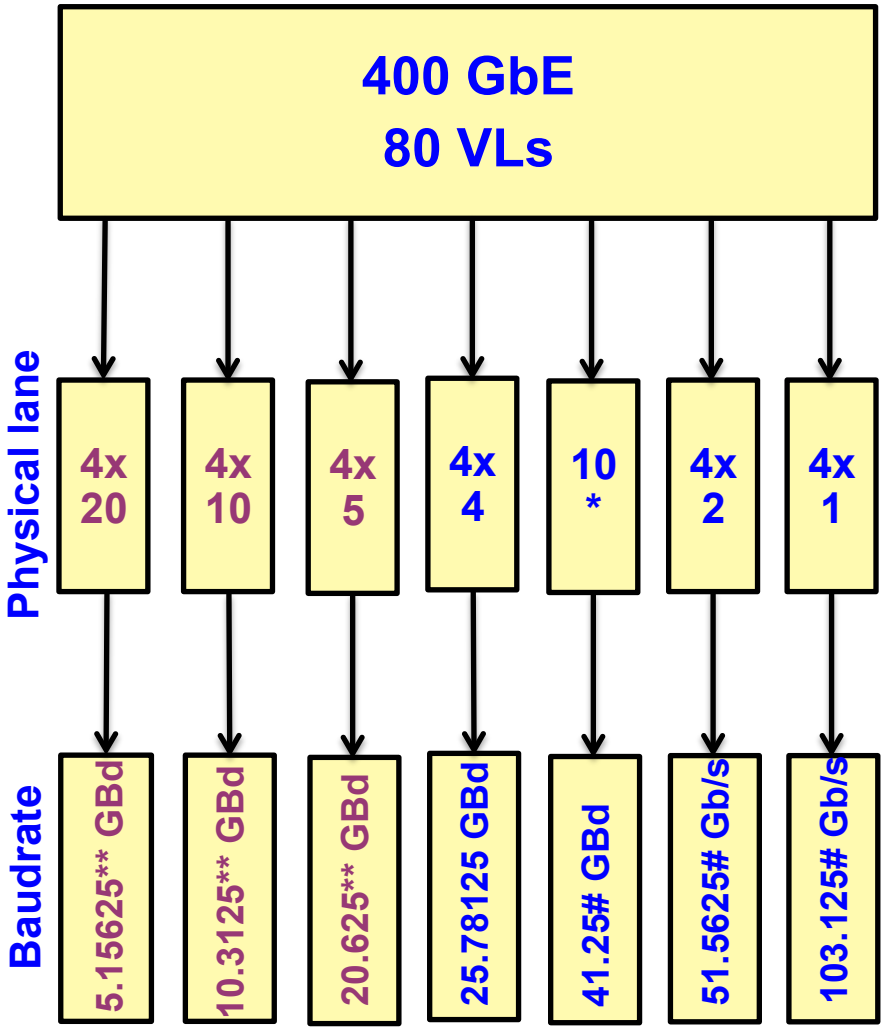
* Only ¼ of 400 GbE slice is shown

**Uncommon implementations

Likely future PMD implementation based on advance modulation

400 GbE PCS and Possible Physical Lanes

- 80 VLs provide maximum flexibility without eliminated specific PMD implementations at this early stage
  - Provides compatibility with 100 GBE PCS
  - It can reuse the BJ RS-FEC (528,514)
  - It will be compatible with MLG to transport up to 40 independent `10 GbE
- 400GbE based on 16 VL will not compatible with MLG
  - It can only transport 4x100GbE
  - No straight forward way to transport 10 GbE or 40 GbE
  - 16 VL will limit applications of 400 GBE!



*** Alternate approach to get high density 40 GbE without using MLG but not friendly to 100 GbE**

****Uncommon implementations**

 # Likely future PMD implementation based on advance modulation

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
  - 40Gbase-KR4/CR4 are two very successful standards that leveraged what was already defined in the 10Gbase-KR
  - Let borrow a page from 10/40 GbE success and let history repeat itself
- Lets not force an architecture which require duplicate FECs, PCS, or different VL lanes not consistent with 100 GbE
  - At this early stage we need to keep flexibility and not eliminate specific physical instantiation
  - Synergy and compatibility should have precedence over starting with clean sheet of paper specially when there is little to be gained
- The proposed architecture with common blocks for 100 GbE and 400 GbE amortizes the cost when 400 GbE volume is very small
  - Integrating 100 GbE RS FEC into a high port count ASIC already challenging
  - Integrating a dedicated 400 GbE FEC will more than double the above gate count
  - The problem even gets worse considering HOM like 100 Gb/s serial that require high gain PMD specific FEC, now you have to duplicate two very complex FEC!