Scalable 400 GbE Architecture

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

May 17, 2013

400 GbE Study Group

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.

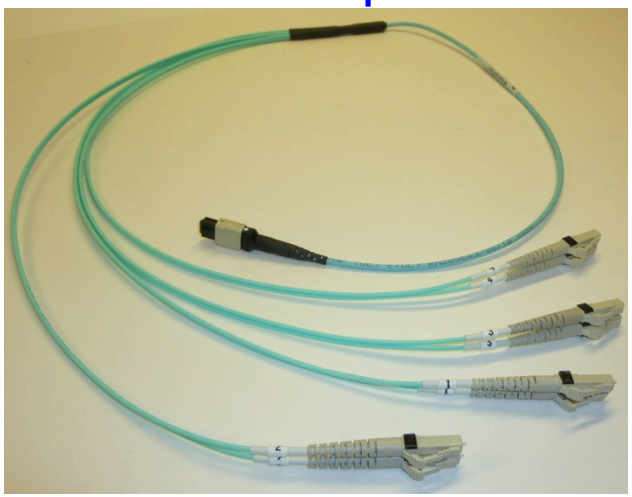
How Dual 10/40GbE Has Served the Marketplace

- 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

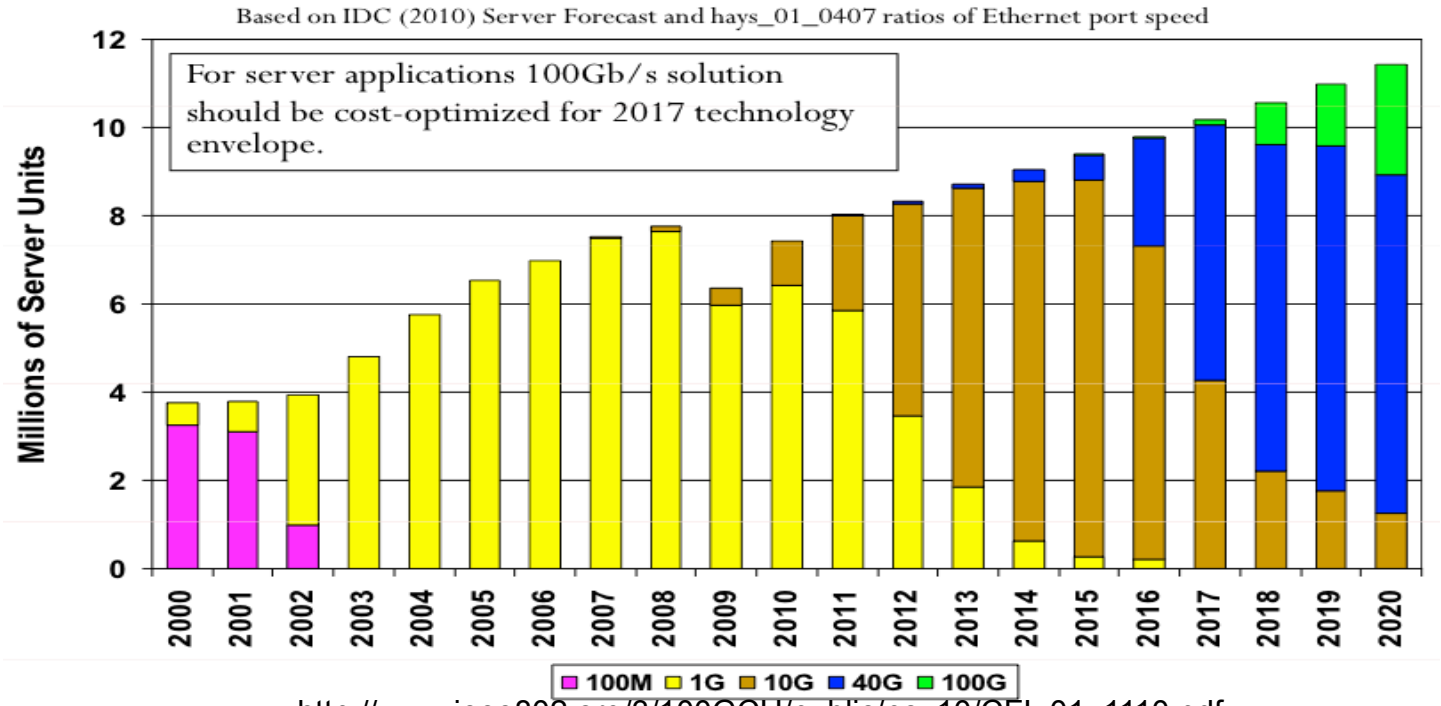
MTP to 4 Duplex LC



Picture Courtesy of TE

Market Reality

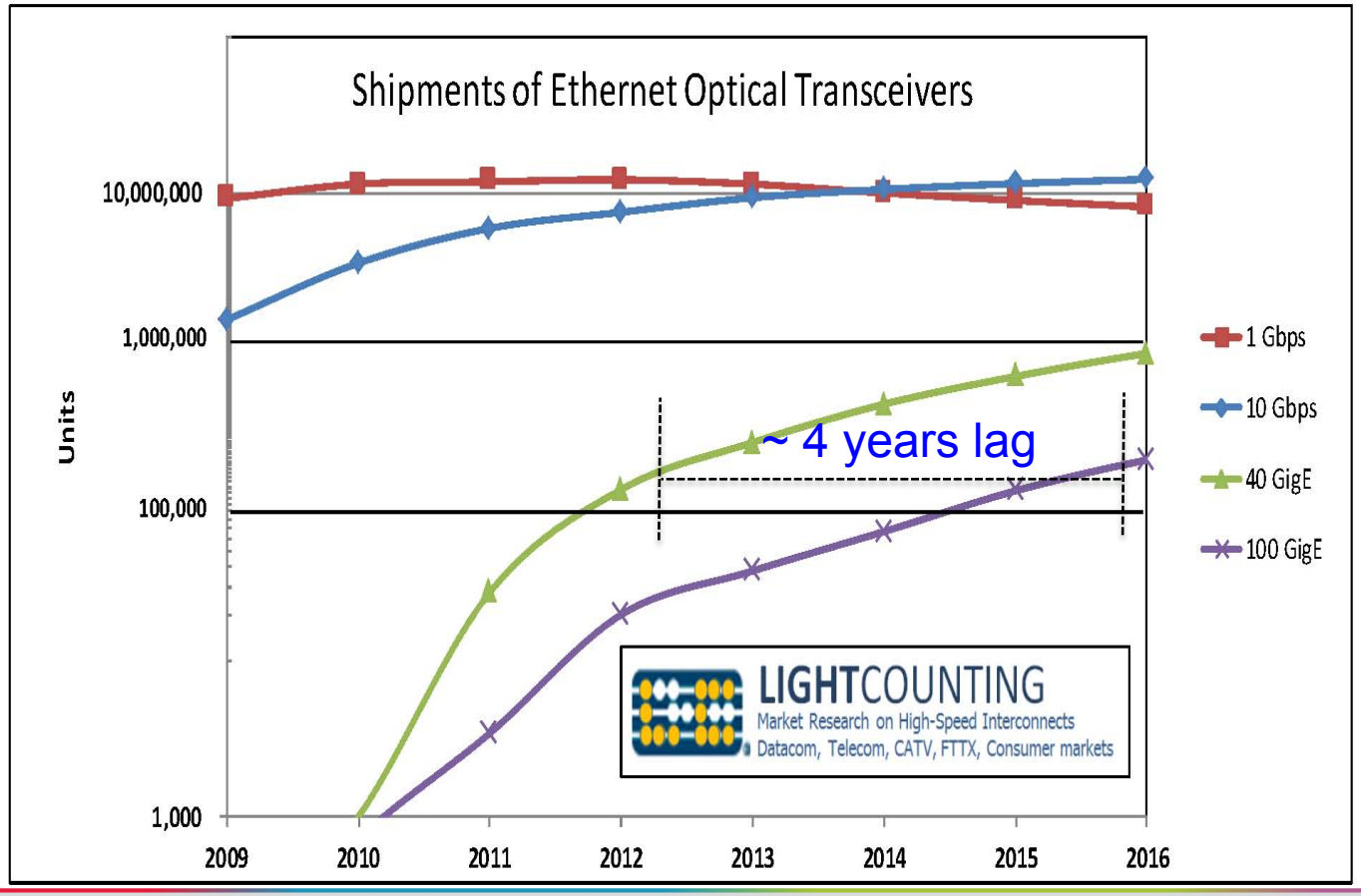
- 100 GbE still is a niche market and 400 GbE volume could be delayed 4-9 years compare to 100 GbE
  - If the cost of implementing 400 GbE ≤ 4x100GbE then volume for 100 GbE could drive dual mode port reducing 400 GbE market lag
- It was well stated by muller_01_0107 “A network technology that cannot connect to computers is (by definition) low volume”
  - We will not see volume deployment of 100 GbE on x86 servers till ~2020



http://www.ieee802.org/3/100GCU/public/nov10/CFI_01_1110.pdf

10, 40, and 100 GbE Port Shipment

- Data courtesy of Light counting
- Shipment up to 2012 are actual volume
  - In 2012 ~50% of 100G shipments were CXP modules, but they do not contribute much to the 2013-2016 forecast.

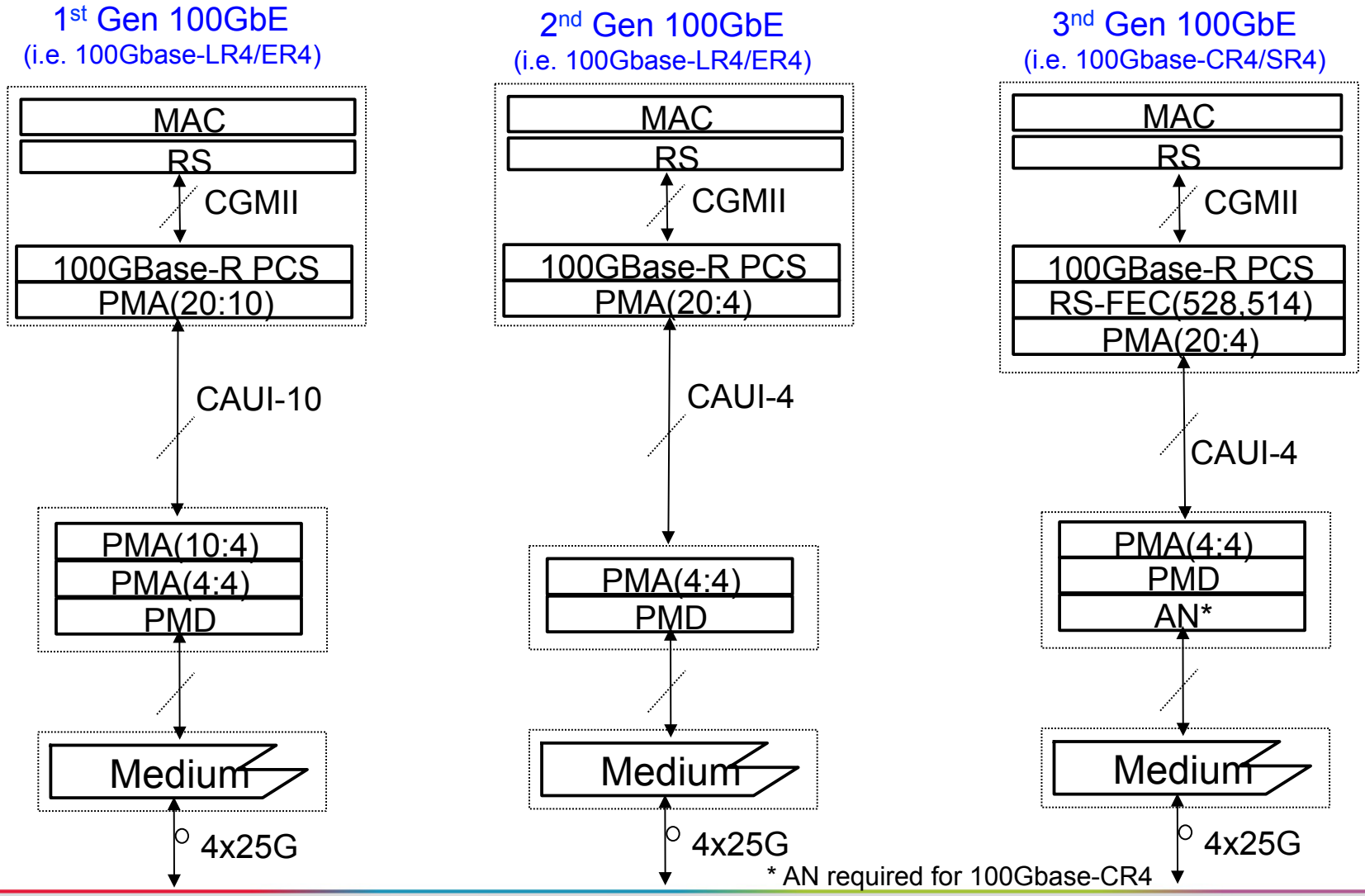


Further Market Analysis and the Reality

- Creehan Datacenter Research (1/2013) for 2016 reports ~30 M ports of 10GbE, ~5 M ports of 40 GbE, but comparable port count to Light Counting for 100 GbE
  - Indicative of significant Cu DAC ports shipment plus some unpopulated 10 GbE and 40 GbE ports
- 100 GbE volume lag 40 GbE by ~ 4 years, under the most optimistic scenario 400 GbE volume will lag 4 years from 100 GbE
  - Expect the volume for 400 GbE sometimes in the 2020-2025 to reach 100GbE volume shipment in 2016
  - But an scalable 400 GbE architecture compatible with 100 GbE will accelerate above date
  - Creating the most elegant 400 GbE architecture not synergistic with 100 GbE will significantly delay volume shipment of 400 GbE
- If we create a synergistic eco-system similar to 10/40 GbE then 100 GbE volume will drives high density dual mode 100/400 GbE
  - By 2020 100 GbE volume server will start and the success of 10/40 GbE dual mode port can be repeated!

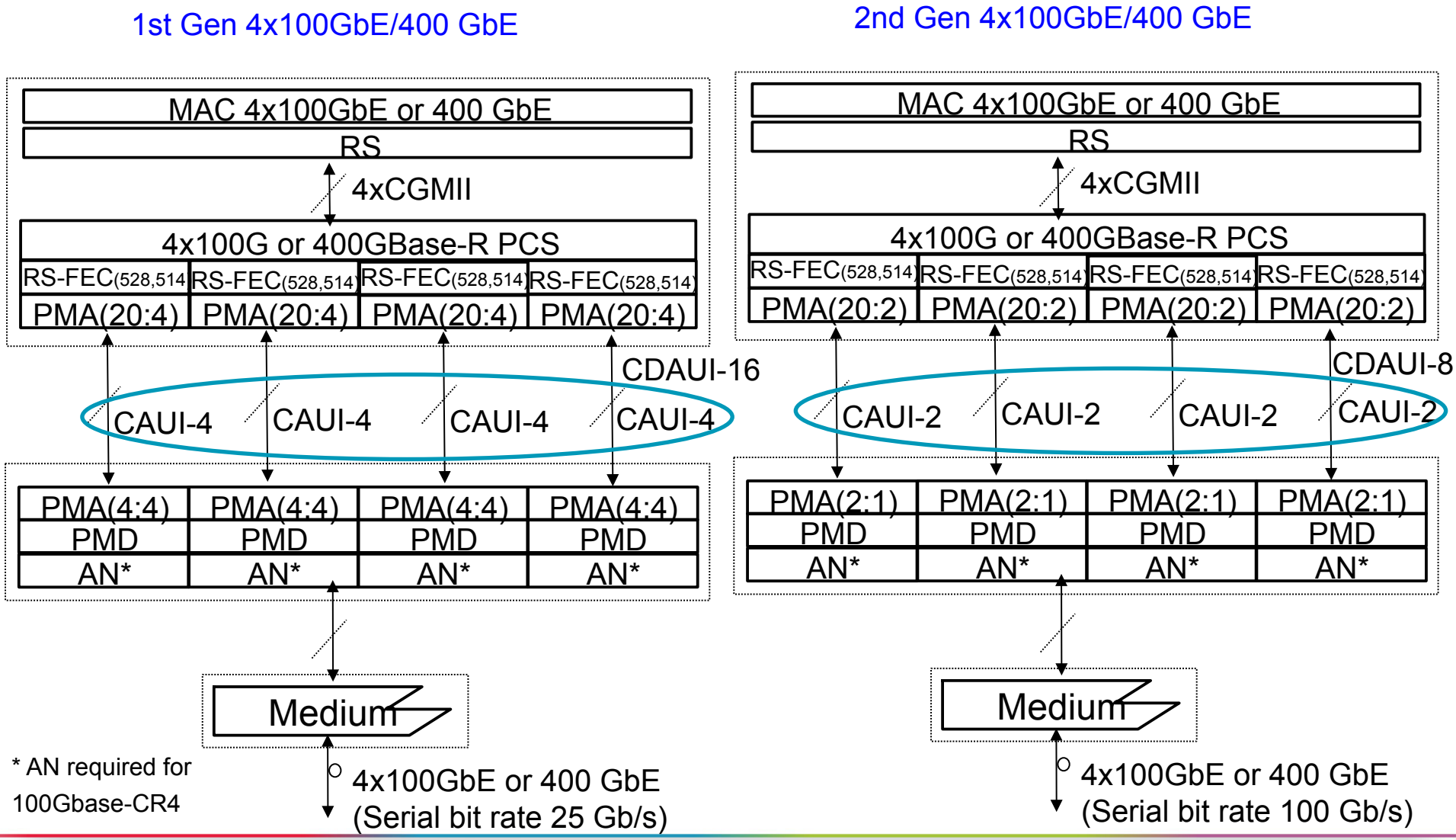
100 GbE Layer Diagram Evolution

- Seamless evolution with introduction of 100 GbE PMDs

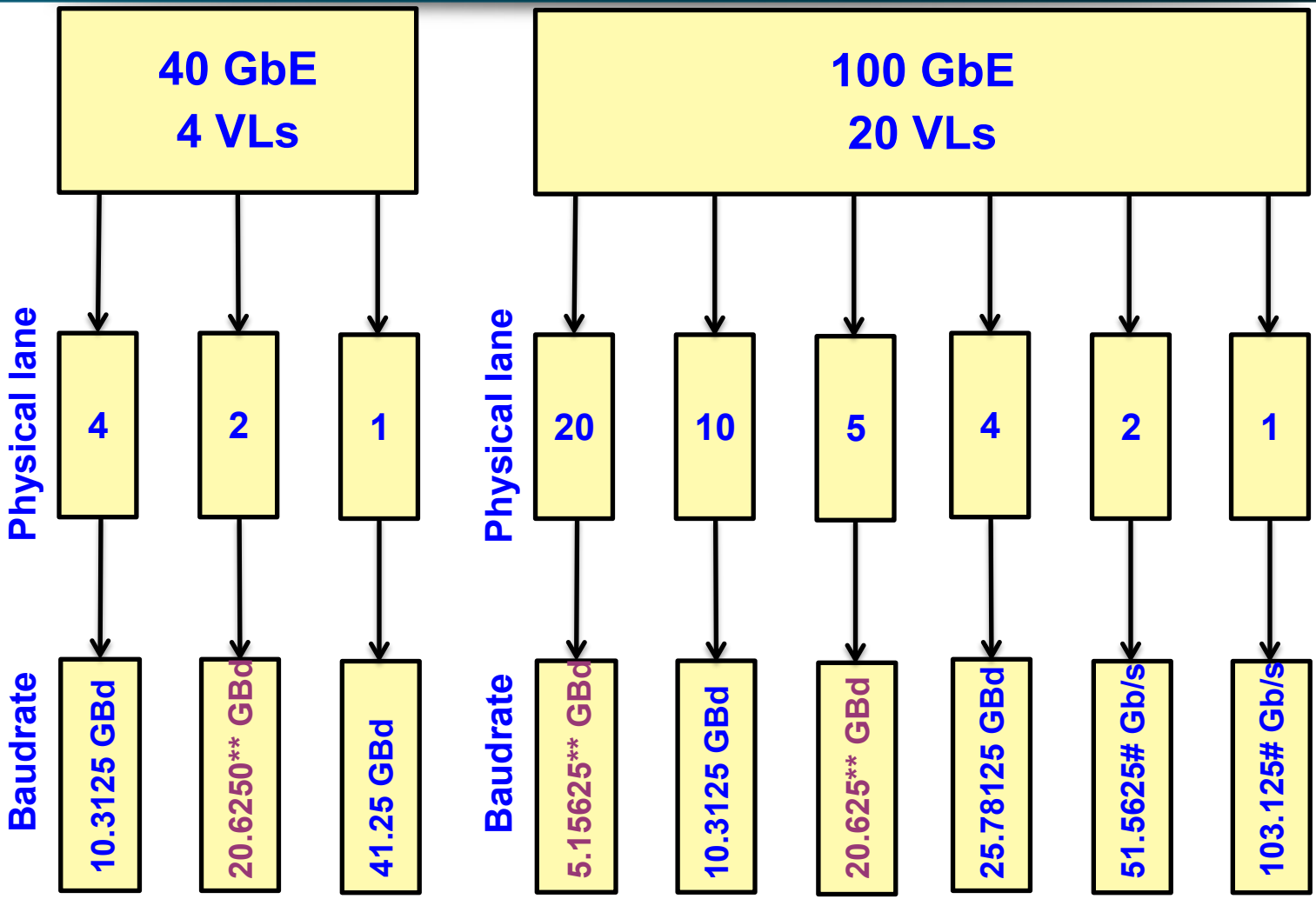


Scalable 100/400 GbE Implementation

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



40/100 GbE PCS and Possible Physical Lanes

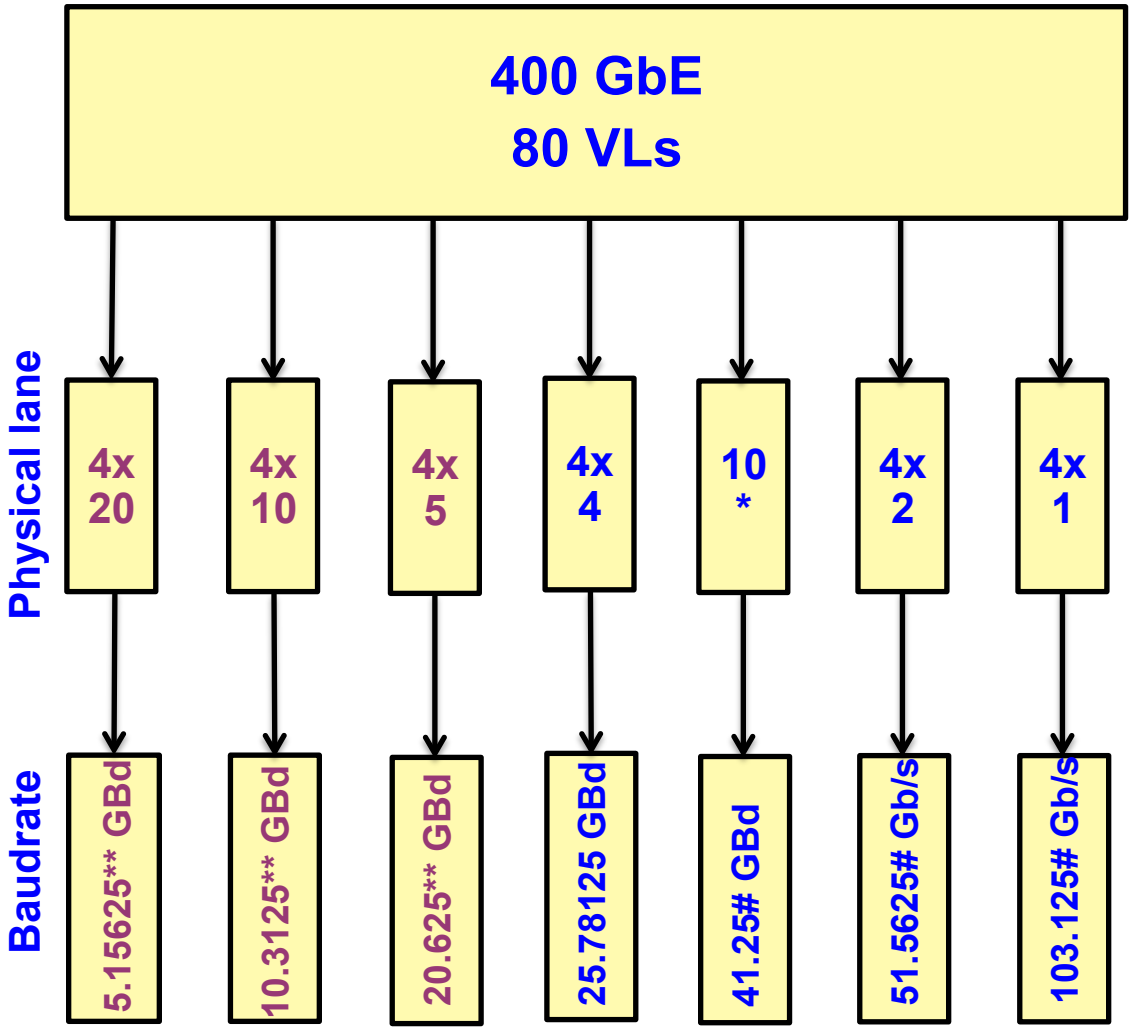


* Only ¼ of 400 GbE slice is shown

**Uncommon implementations

Likely future PMD implementation based on advance modulation

400 GbE PCS and Possible Physical Lanes



* 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
  - Let borrow a page from 10/40 GbE success and let history repeat itself
  - Lets not create an architecture which require duplicate FECs, PCS, or different VL lanes not consistent with 100 GbE
- The proposed architecture with common blocks for 100 GbE and 400 GbE amortizes the cost when 400 GbE volume is very small
- Early product could be implemented by 4xCAUI-4/CDAUI-16
- With OIF 56G VSR currently underway 4xCAUI-2/CDAUI-8 is natural next step
- 802.3bm already has investigated feasibility of 100 Gb/s with advance modulation on fiber and is natural for 400 GbE
- In the next few meetings the scope of the project need to be define and if key new technology blocks such as CDAUI-8 and/or serial 100 Gb/s will be defined in this project.