Impact of Going Beyond 400G

Brad Booth, Microsoft

IEEE Beyond 400G Study Group
March 1, 2021

Supporters

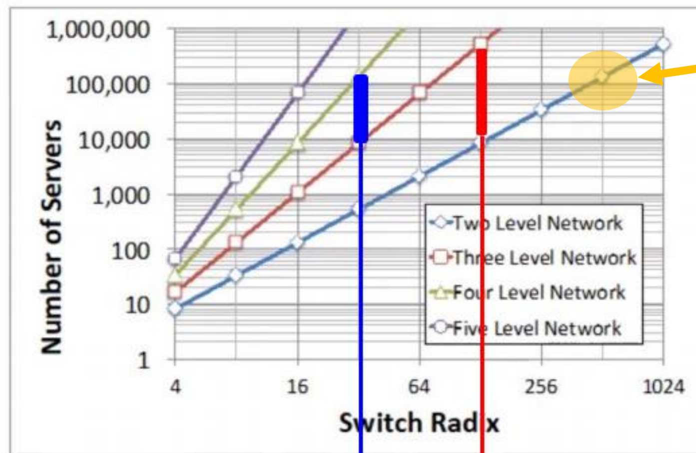
- Rob Stone, Facebook
- Mark Filer, Microsoft
- Renjit Mathew, Arista
- Jeff Hutchins, Ranovus
- Mark Nowell, Cisco
- Ali Ghiasi, Ghiasi Quantum
- Weiqiang Cheng, China Mobile
- Rang-Chen Yu, SiFotonics

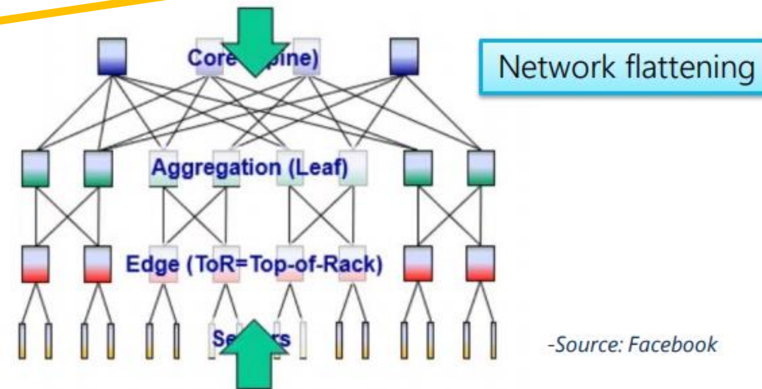
Market for Beyond 400G

- Not all data centers are the same
  - Microsoft uses a regional design – speed match between DCI & intra-DC
  - Geographic presence
  - T-shirt size design
  - Customers and offerings
  - Growth strategy, module preference, etc.
- Network topologies and applications differ
  - Distributed vs. single-point VMs
  - xAAS (x As A Service – IAAS, SAAS, PAAS, etc.)
  - AI/ML, general compute, high-performance compute, storage
  - Radix structured vs. flexible

Topology Difference

MSFT: Radix 512 networks
100K servers (32MW DC)





Network flattening

-Source: Facebook

Switch Generation	Radix = 32	Radix = 64	Radix = 128
12.8T	400G	200G	100G
25.6T	800G	400G	200G
51.2T	1.6T	800G	400G

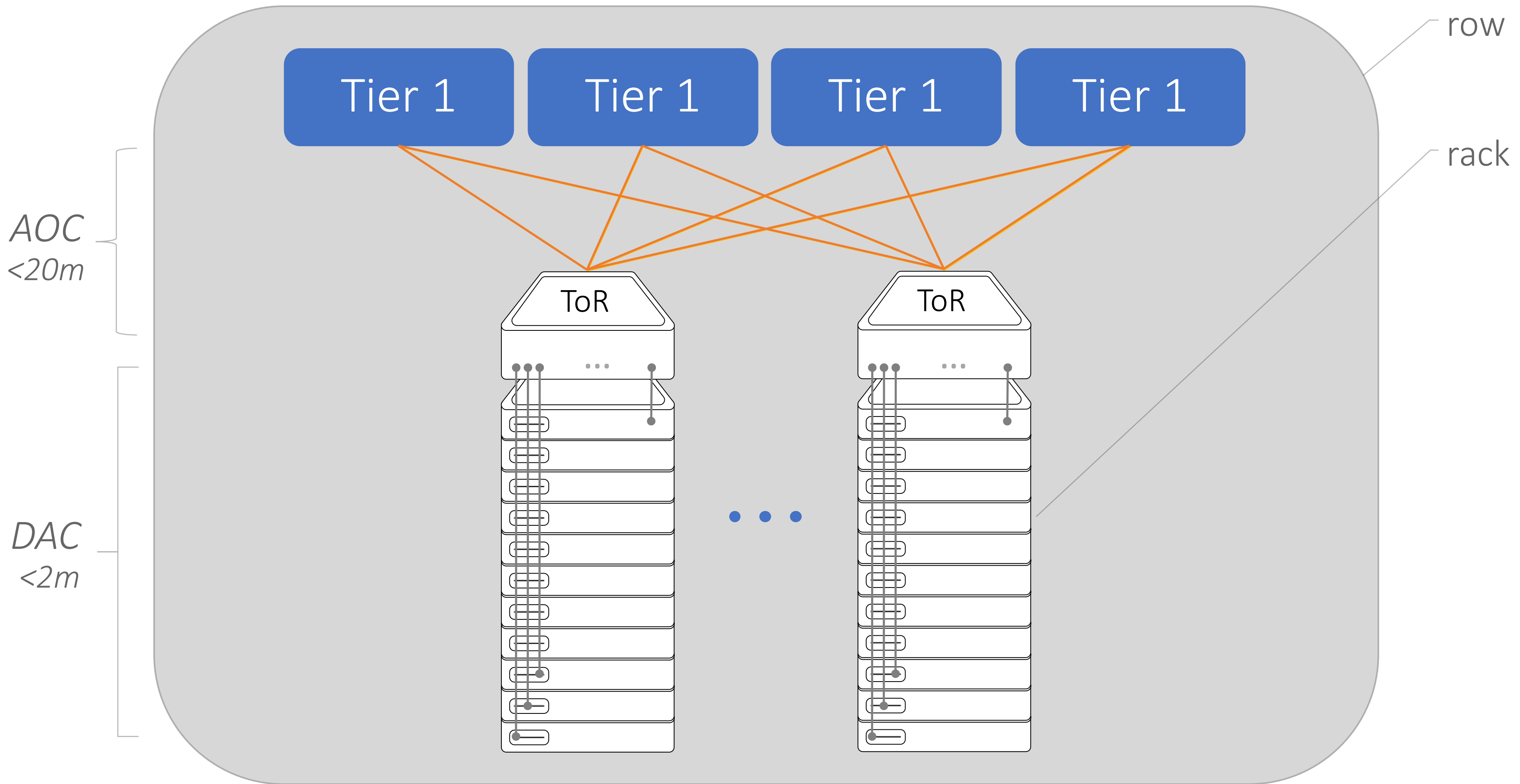
Optical interconnects

Inphi R. Nagarajan, Ilya Lyubomirsky, "Next-Gen Data Center Interconnects: The Race to 800G" 3 Inphi

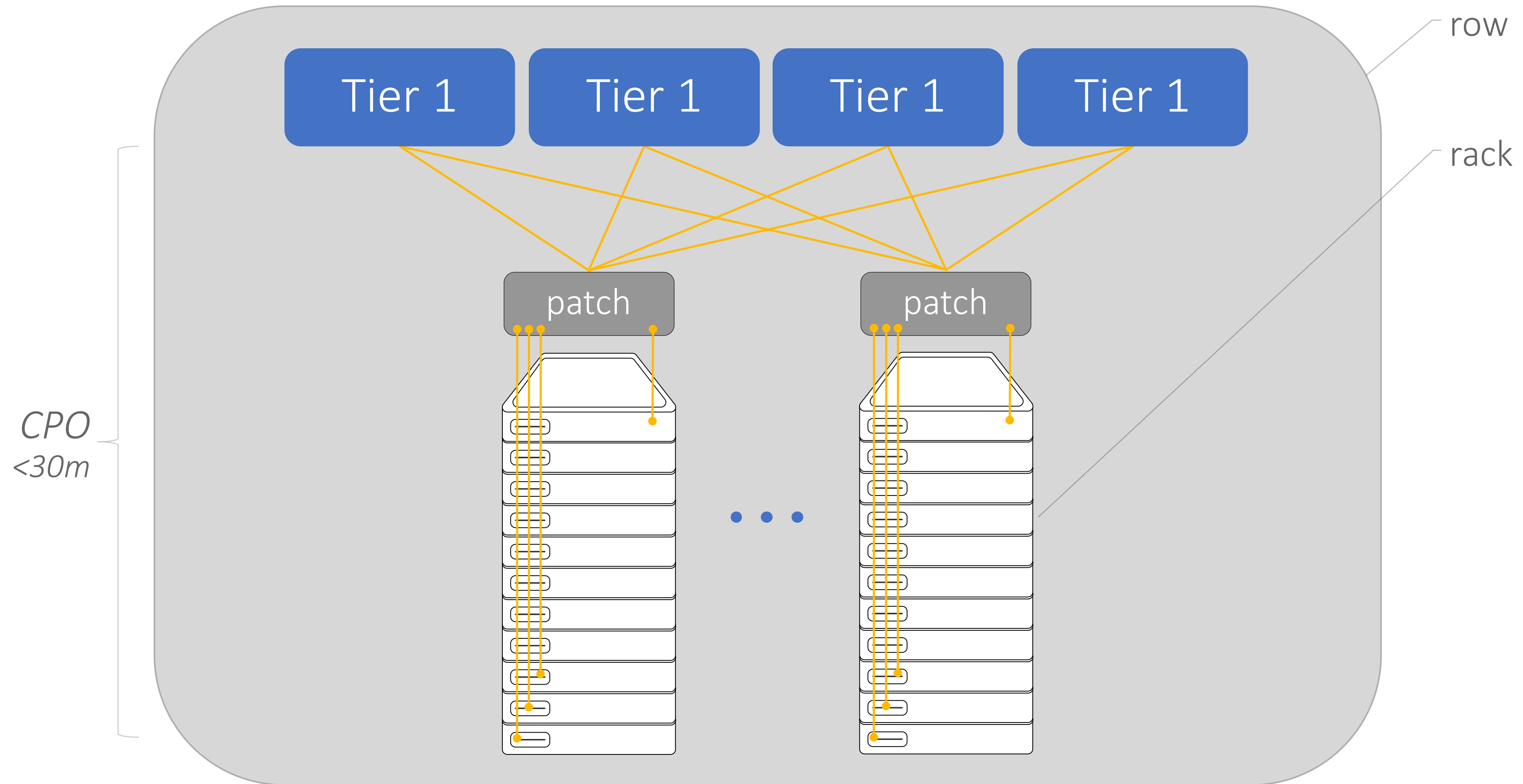
- Fewer tiers = decreased latency, lower power
- CPU bandwidth ~ 1G/core
- Volume of servers vs. power grid

Other CSPs use lower radix

ToR Elimination Example



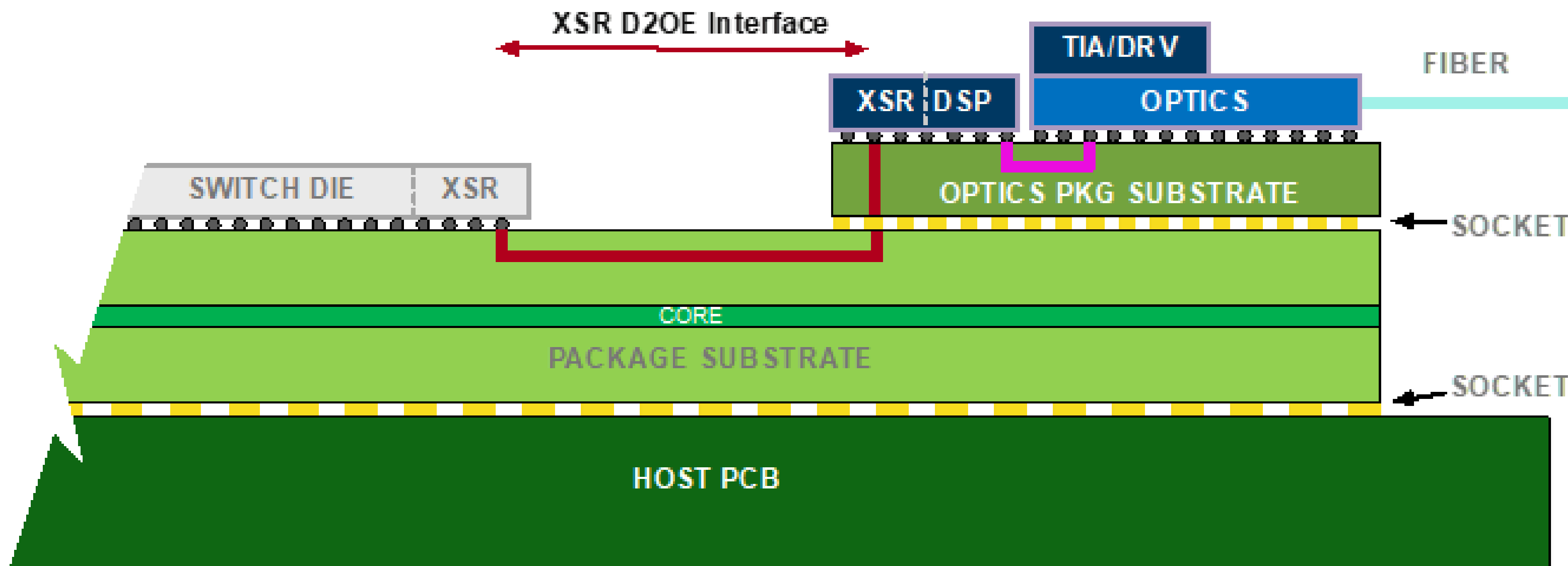
ToR Elimination Example (cont)



Co-packaged Optics (CPO) Intro

- Facebook & Microsoft project under the Joint Development Foundation (JDF)
  - www.copackagedoptics.com
- OIF CPO Framework project
  - www.oiforum.com/technical-work/current-work/#co-packaging
- COBO CPO working group
  - www.onboardoptics.org

3.2T Optical Chiplet Concept per CPO JDF

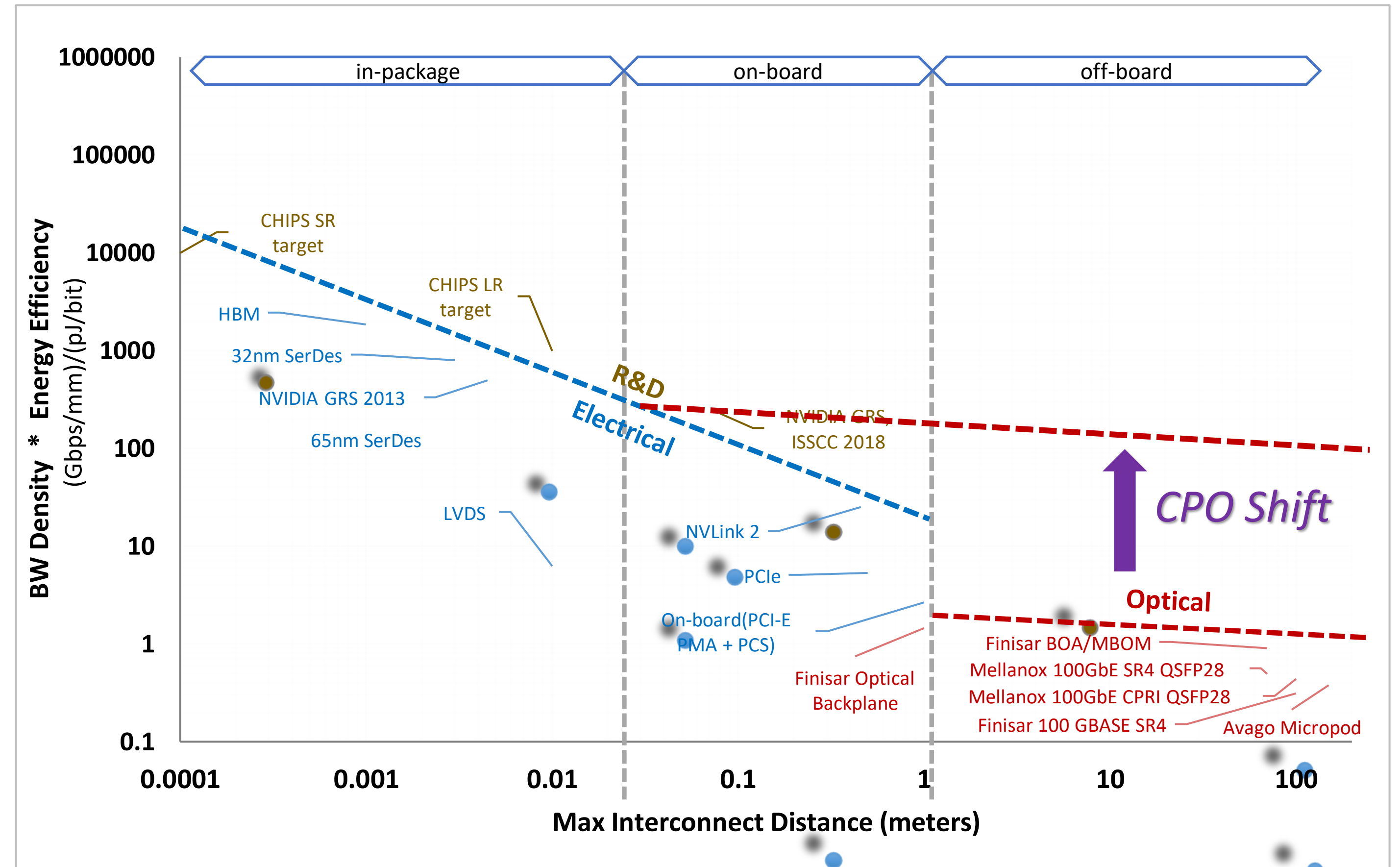


Interconnect Figure of Merit (FoM)

Source: DARPA Photonics in the Package for Extreme Scalability (PIPES)

Primary Metrics

- Reach (Distance)
- Bandwidth Density
- Energy Efficiency
- Cost



CPO Potential

- FoM comparable to chip-to-chip electrical interconnects
- Benefits:
  1) Photonic integration
  2) Standard manufacturing
  3) Multi-vendor ecosystem
  4) Broad range of use cases

Impact to Ethernet Standards (C2M)

- Multi-lane interfaces are not new to 802.3
  - 400G has 16x25G, 8x50G & 4x100G
  - 4, 8 and 16 lane modules exist or are in development (i.e. QSFP112, QSFP-DD, OBOx16)
- Systems can take advantage of multi-lane PHYs
  - Modules have a broad range of bandwidth support
  - CPO chiplet initially targeting 32-lane
- Provides flexibility for varying topologies and applications

Thoughts

- 802.3 higher speed projects
  - Up to 10G, MAC and PHY rates “matched”
  - Beyond 10G, MAC exceeded PHY → created a “fill-in” approach
  - Next generation optics (CPO, OBO, FPP) breaks this approach
- Time to modify 802.3’s approach to growth
  - Many other standards bodies have already shifted their approach
  - Make the PHY per lane the building block (n, where n= 100G, 200G, etc.)
  - Make the MAC flexible ($n*2^m$, where m={0, 1, 2, 3, 4, 5, 6})
    - Permits radix flexibility for differing topologies & applications
- Enables a smoother growth path w/ fewer “fill-in” projects (avoids gaps)

Flexible MAC Concept

- Objectives
  - 800G & 1.6T are close, but…
  - Specify a flexible MAC to support 2^m PHY layers
  - Specify the PHY to permit an aggregation up to m lanes
- SG only needs to determine the range of “m”
  - Recommend a max m of 4

Thank you.