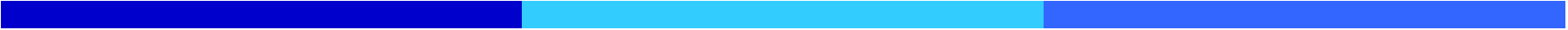


Background for today's call

- Andy Bechtolsheim presented at an NEA call in 2017, suggesting that we plan for 100G SerDes in the 400G over MMF CFI, but it was not considered technically feasible at that time.
- Ali Ghiasi reminded the 400G over MMF SG in Jan 2018 Interim at Geneva that server-switch lengths may grow from 2-3m to 20 to 30m due to future trends in switch radix (growing) and server count-per-rack (decreasing)
<Ref>
- By mid-2019 VCSEL-makers are ready to support 100G technical feasibility
- Some large cloud players are hearing that a lower-cost SMF server interconnect may be available in future
- In discussions in Vancouver & SLC, most people I discussed this with were quite interested, but many questions were raised.
- There seem to be sufficient interest & good questions for a SG to tackle



Lower cost, shorter reach, optical PHYs using 100 Gb/s wavelengths

In-progress CFI Consensus Presentation Draft

Robert Lingle Jr., OFS

July 30, 2019

NEA Ad Hoc Teleconference

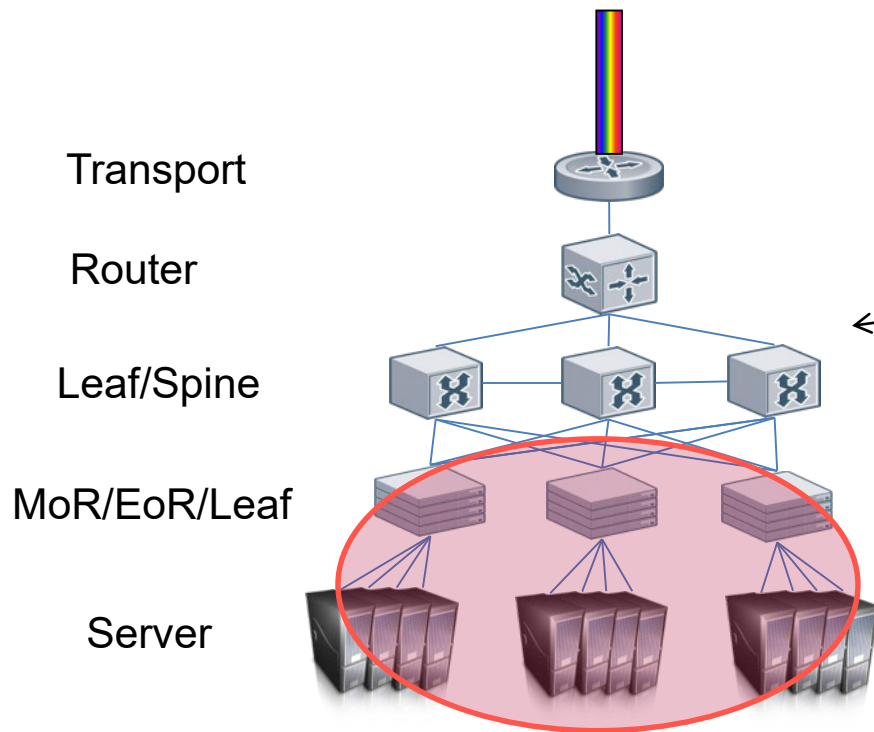
CFI objectives

- To measure the interest in addressing:
 - lower cost, shorter reach, optical PHYs using 100 Gb/s wavelengths
- We do not need to:
 - Fully explore the problem
 - Debate strengths and weaknesses of solutions
 - Choose a solution
 - Create a PAR or 5 Criteria
 - Create a standard
- Anyone in the room may vote or speak
- RESPECT ... give it, get it

Motivation

- It is attractive to consider shifting from ToR to MoR/EoR architectures, requiring longer server-attachment links, sometimes including breakout
 - Server attachment speeds are increasing from 25 and 50 GbE to 100GbE, while number of servers per rack are decreasing due to higher power dissipation and more auxiliary functions in server trays.
 - Meanwhile, drive to higher switch ASIC throughput and SerDes rates is making more ports per switch available and economical.
- This proposed study group would look to develop ~30m (TBD) SMF and/or MMF PHYs using 100G wavelengths to match emergin 100G SerDes
- The motivation is to leverage technology to address the ongoing cost pressures on optical interconnects in the web-scale datacenter market.
- Lower cost solutions occur due to reduced lane/component count or through enabling higher density solutions.

What are we talking about?



Applications for early adoption of next-generation MMF PMDs include connectivity in Big Cloud data centers for

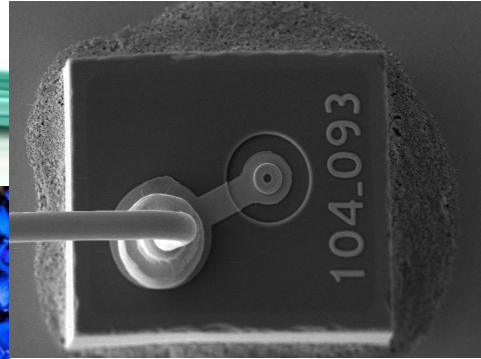
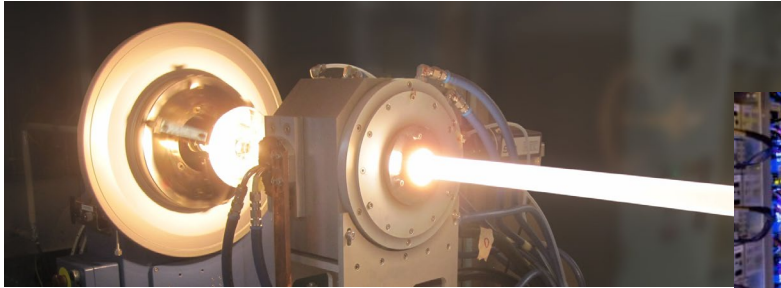
- **Switch-to-server**
- **GPU clusters**

Agenda

- **Presentations (tentative)**
 - **Market Drivers**
 - Several candidates
 - **Technical Feasibility**
 - Vipul Bhatt (Finisar)
 - Ramana Murty (Broadcom)
 - Other
 - **Why Now?**
 - Robert Lingle, Jr. (OFS)
- **Straw Polls**



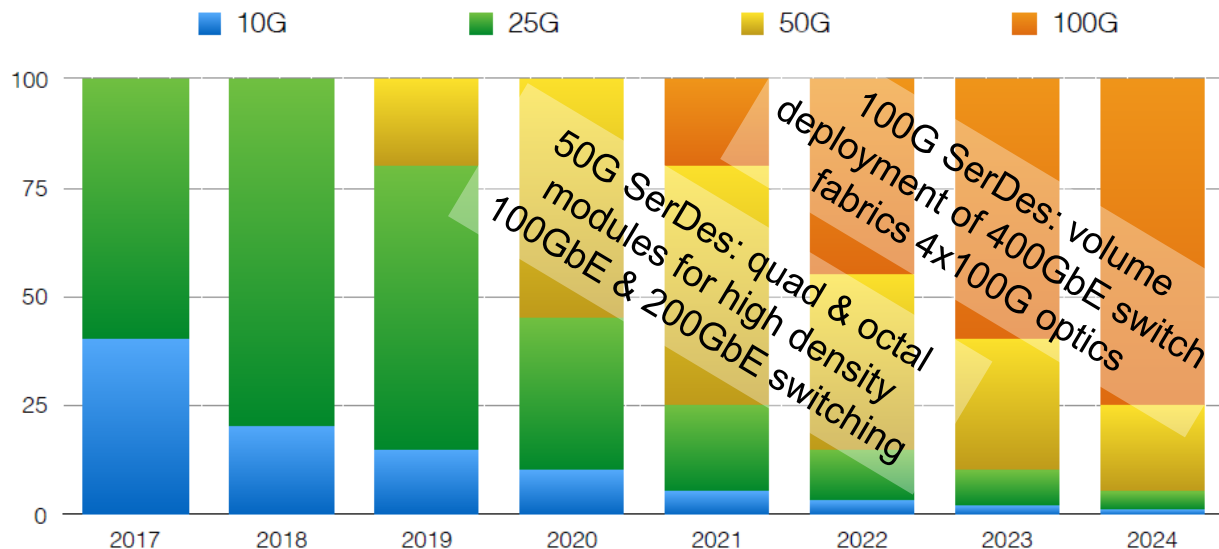
Market drivers



The progress of SERDES speeds drives the economics & evolution of switch fabric speeds and port counts

ARISTA

SERDES Speed Transition Over the Years [% Mix]



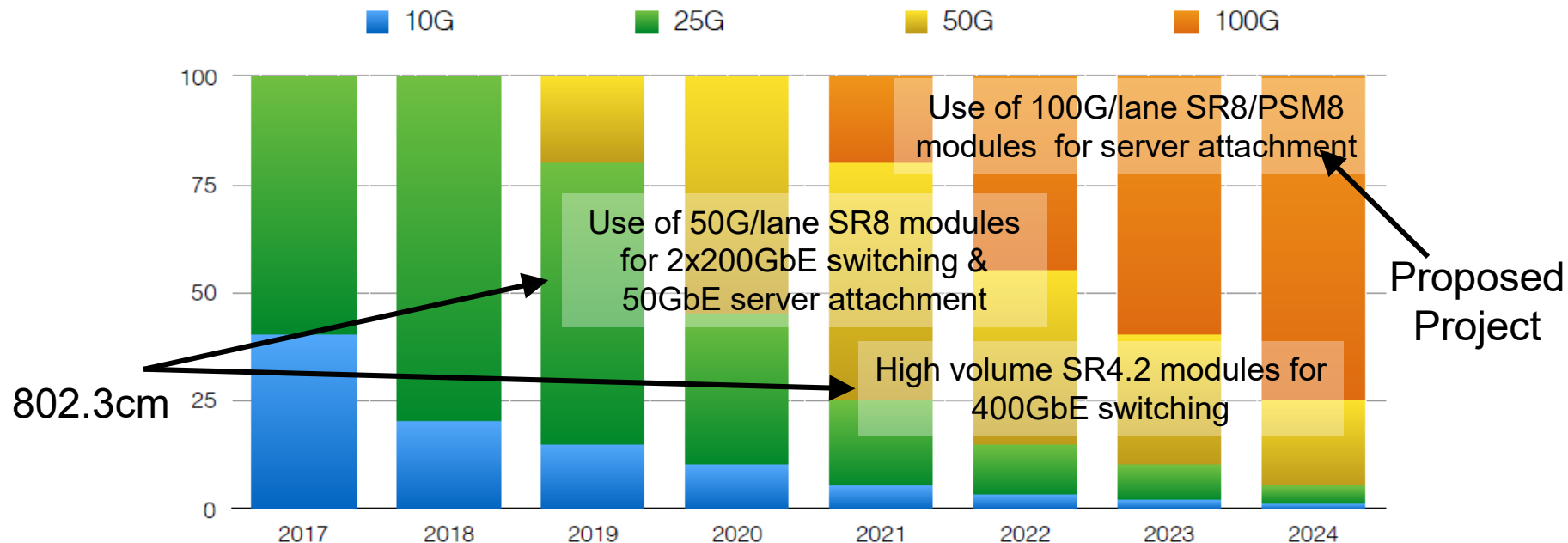
https://pc.nanog.org/static/published/meetings/NANOG75/1954/20190220_Martin_Building_The_400G_v1.pdf

(Annotations by author of this document)

Chart shows possible timing of some applications of existing (100m) and proposed (~30m) short reach PMDs

ARISTA

SERDES Speed Transition Over the Years [% Mix]



Optimized server architectures evolve with server and switch technology

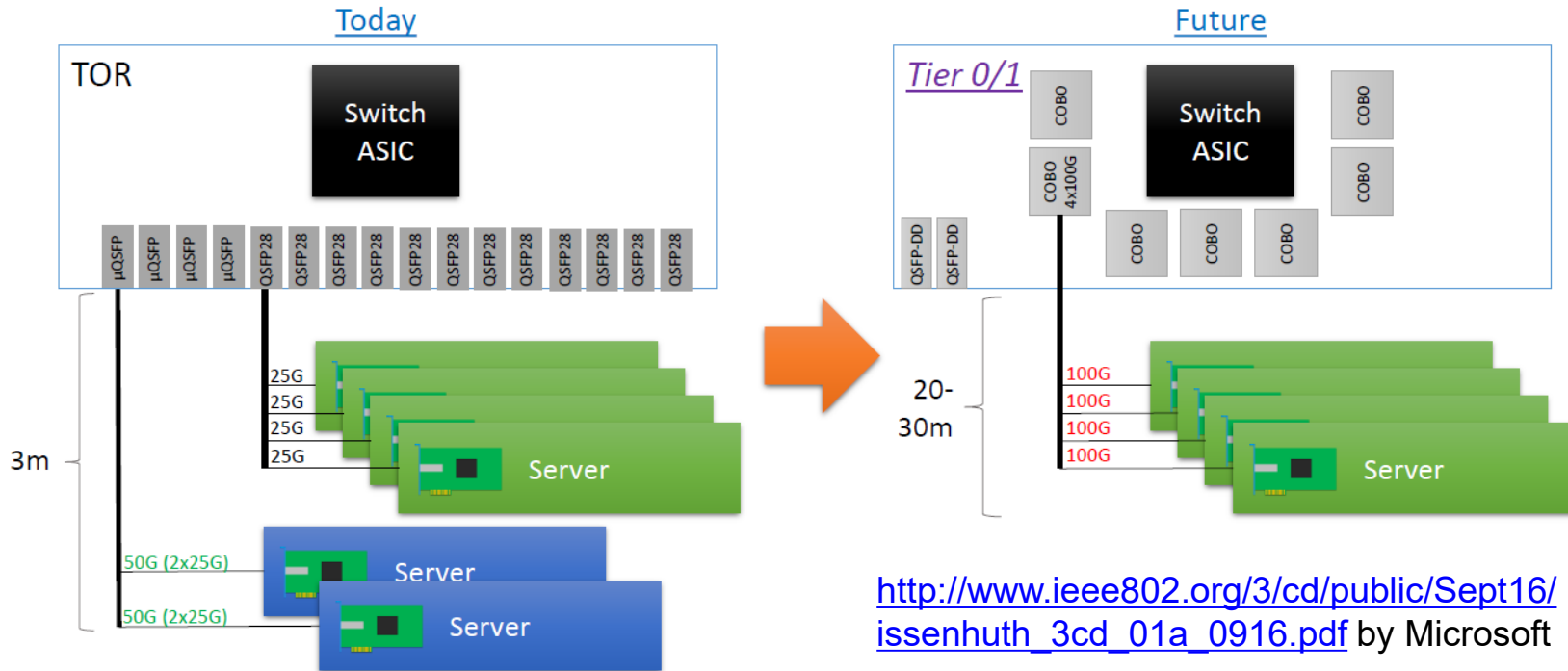
- Servers are using 25GbE and 50GbE links, with 100GbE on the horizon
- As each server becomes more powerful,
 - The number of servers per rack is decreasing
 - For example, some designs will move to 24 per rack and even as low as 6 per rack with GPU accelerators
 - Some architectures prefer to connect each server to two switches for redundancy
- Moving the first switch from ToR to MoR/EoR may allow higher utilization of switch ports and lower cost deployment of redundancy

Drive towards higher switch capacity for 400GbE provides higher port density for server attachment

- A 12.8T ASIC with 256x50G SerDes supports 32x400GbE, 64x200GbE, or 128x100GbE links.
- Technologies developed for 400GbE using 50Gb SerDes are primarily being used today to support higher radix 100GbE and 200GbE switch fabrics.
 - An 8x50G pluggable support 1x400GbE, 2x200GbE, or 4x100GbE links
 - Reverse gear boxes are even being used to connect 2x50G lanes to 4x25G QSFP28
 - [Volume deployment of 400GbE switch fabrics probably tracks 100G SerDes]
- As pointed out by Ghiasi in NGMMF Study Group (Jan 2018), the trends of increasing switch radix and decreasing server count-per-rack may combine to favor MoR/EoR architectures over ToR.

http://www.ieee802.org/3/NGMMF/public/Jan18/ghiasi_NGMMF_01_jan18.pdf

Emerging MoR/EoR architectures will benefit from compact optical cable and ease of breakout over 20-30m



http://www.ieee802.org/3/cd/public/Sept16/issenhuth_3cd_01a_0916.pdf by Microsoft

Comments on market need for low cost, short reach, server interconnects, from Zuowei Shen (Google Cloud)

- To be provided mid-August

Comments on a server & switch on market need for 100G/λ short-reach interconnects, from David Piehler (Dell EMC)

- Market need
 - Low-cost interconnect for $32 \times$ "800G" switches (expected in 2020).
 - Passive copper cable limited to (1-2?) m. Active copper cable limited to ~ 5m.
 - "SR16" (50G/λ) has $2 \times$ lane count + unusual higher fiber-count connector.
 - Useful even if maximum distance is 30 m.
 - Low-cost interconnect for 100G (serial) servers (2021+)
- Use cases
 - 100GBASE-SR(1)
 - SFP112 connections to for next-generation servers.
 - 400GBASE-SR4
 - Lowest-cost, low-fiber count point-to-point connection for 400G QSFP56-DD ports
 - Breakout to $4 \times$ 100GBASE-SR(1)
 - $2 \times$ 400GBASE-SR4 (**don't know what to call "SR8" since there is no 800GbE defined**)
 - Lowest-cost, low-fiber count point-to-point connection for $2 \times$ 400G QSFP112-DD (or OSFP112 ports)
 - Breakout to $8 \times$ 100GBASE-SR(1)

Comments on need for short reach 100G/wavelength, by Chongjin Xie, Sr. Director at Alibaba

- Applications:
 - AOC for server to TOR connections
 - Transceivers for TOR to leaf switch connections
- Distances:
 - 100 meters desired
 - ≤ 50 meters required for transceivers
 - ≤ 30 meters is space for AOCs at Alibaba
 - Server connections will be longer than 2-3m in the future
- Configurations:
 - Breakout support required
 - May or may not be breakout, depending on network architecture
- Cost & power concerns
 - Cost < 50% of DR
 - Power consumption \sim 50% of DR

Additional market input from end-users and OEMs is expected

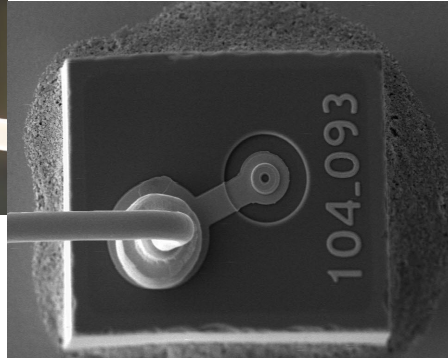
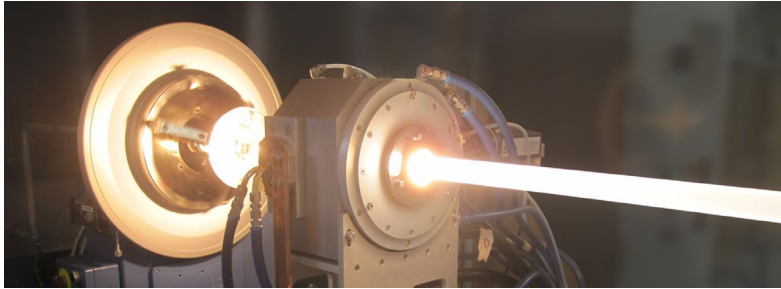
- Prior to September Interim

Cost and density benefits accrue from higher lane speeds

- For ~100m (MMF) and ~500m (SMF) applications, 400Gb Ethernet has been defined as:
 - 400GBASE-SR16 (16x25G) with 16 fiber pair (fp) cable
 - 400GBASE-SR8 (8x50G) with 8 fp cable
 - 400GBASE-SR4.2 (8x50G) with 4 fp cable
 - 400GBASE-DR4 (4x100G) with 4 fp cable(100GBASE-DR and 50GBASE-SR are also defined to match, respectively)
- It is proposed here to study ~30m links optimized for lower cost, lower power, such as:
 - SR, SR4, and/or SR8 style PMDs, based on 100G VCSEL and MMF
 - Single-lane and/or parallel PMDs using lowest-cost 100G/wavelength over SMF
- For MMF links, higher speed lanes leads to reduced lane counts, reduced component counts, reduced complexity, and lower cost than previously standardized PMDs
- Some clearly prefer pluggable transceivers, but lower cost of AOC is attractive to others.

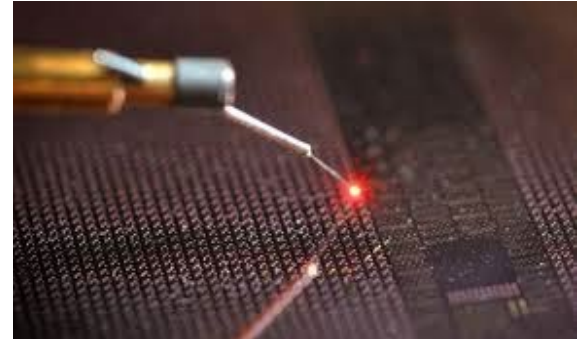
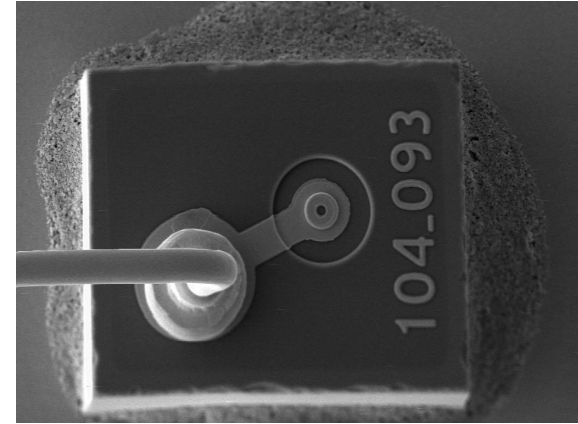


Technical feasibility



Historically VCSEL-MMF links have advantages for lower cost and lower power short-reach interconnects

- Relaxed alignment tolerances
 - Several microns vs. sub-micron
 - Allows passive alignment in module
 - Better cost/loss trade-off for connectors
- Connectors more resilient to dirt
 - Cleaning SMF connectors is common issue
- Lower drive currents
 - 5-10mA vs. 50-60mA
- On-wafer testing
- 802.3cd and .3cm standardized 50G links
- Ethernet does not yet address 100G VCSELs



Could innovations in silicon photonics allow lower cost, lower power short-reach interconnects than DR / DR4?

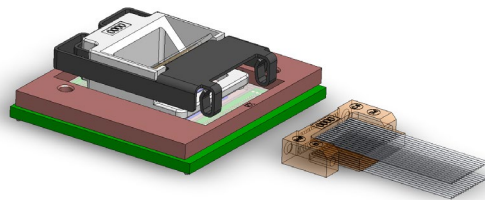
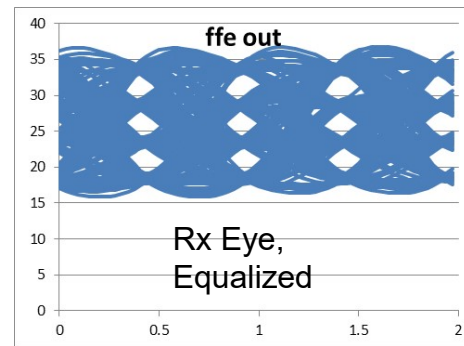
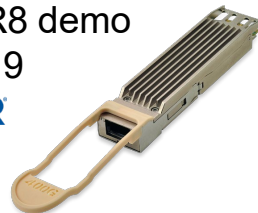
- <Invitation for someone to provide a slide here>

Slide on probable re-use of existing PCS & FEC

Technical Feasibility: 100G Multimode PMD (1 of 2)

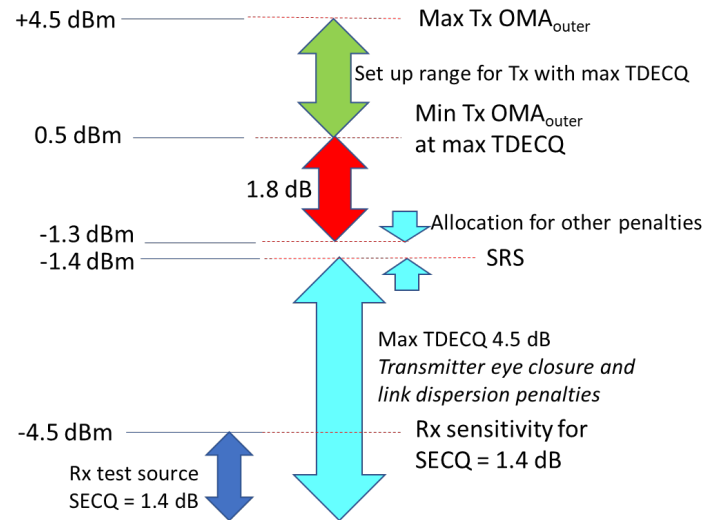
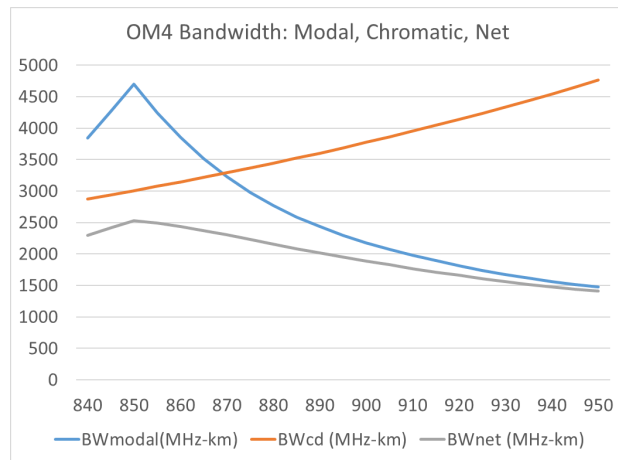
- 100G (50 GBaud PAM4) VCSEL-based multimode PMD cost-optimized for ~30 meters reach is technically feasible
- Development of 50G VCSELs for 400G-SR8 and 400G-SR4.2 transceiver modules have given the implementers a head start
- Development of 100G-capable VCSEL is ongoing
- Simulations from models based on early characterization data show feasibility to ~50 meters
 - Based on a well damped, well behaved VCSEL response, ~24 GHz, pre-emphasis (2-tap T-spaced FFE -0.5,1), 5-tap Rx FFE, 0.6 nm spectral width, 940 nm
- Measurements taken on early VCSEL prototypes look promising
- Development of 50G NRZ multimode transceiver for High-Performance Computing is progressing well
 - Same signaling rate as 100G PAM4

400G-SR8 demo
OFC 2019
FINISAR



Technical Feasibility: 100G Multimode PMD (2 of 2)

- From a link perspective, we have several options to reduce the burden on VCSEL bandwidth requirement
 - Stronger post-detection equalization
 - Improved fiber modal bandwidth
 - Spectral width management
 - Choice of wavelength: 940 nm may offer better performance, relaxed spectral width



Strawman link budget presented by J. King to INCITS T11.2 in June Doc # T11-2019-00161-V000.pdf

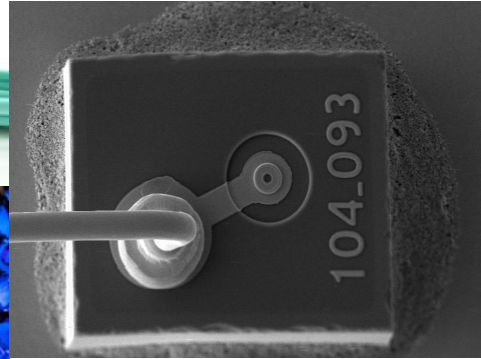
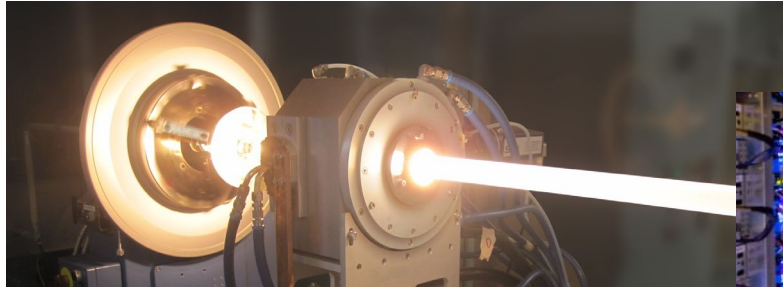
Summary: We can demonstrate that cost-effective 100G/lane multimode PMD for short reach is technically feasible.

Placeholder

- Ramana Murty of Broadcom will supply strong support for technical feasibility in August



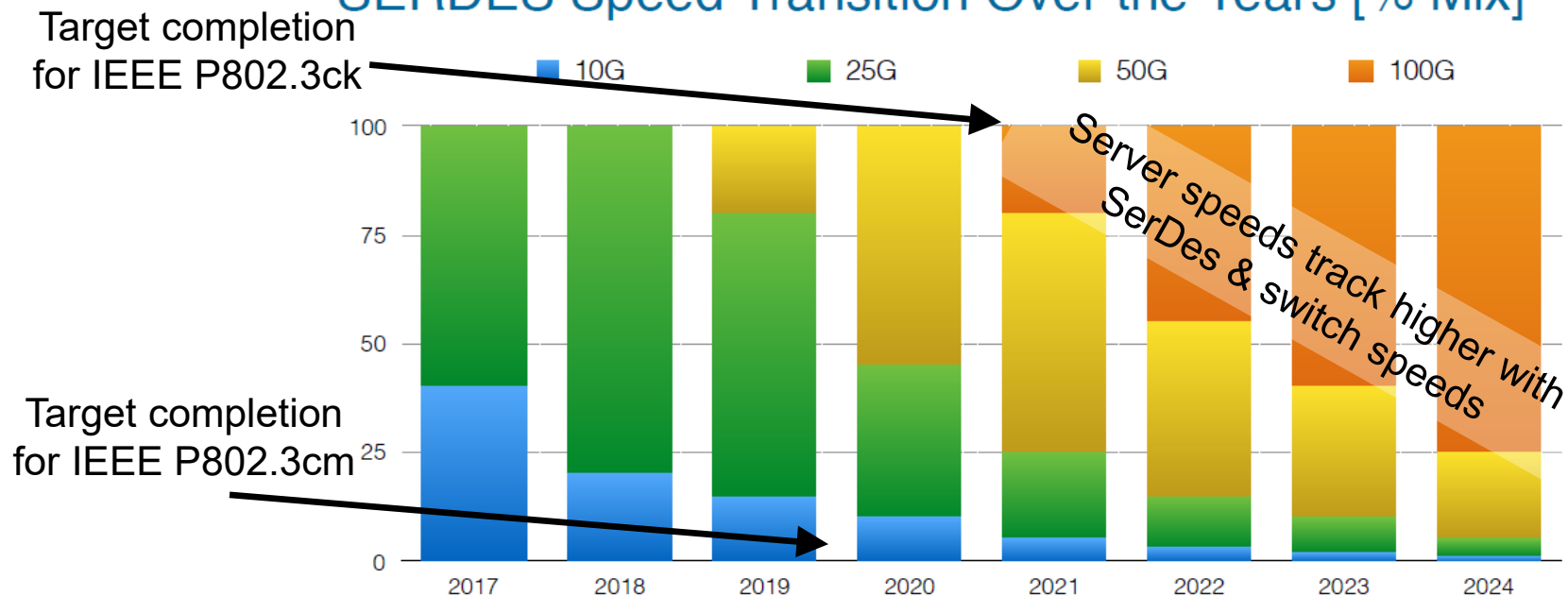
Why now?



Beginning a project now on 100G cost-optimized server interconnects roughly with the expected roll-out of 100G SerDes and higher switch fabric speeds and expected evolution of server speeds

ARISTA

SERDES Speed Transition Over the Years [% Mix]



Contributors

Chongjin Xie, Alibaba

David Piehler, Dell EMC

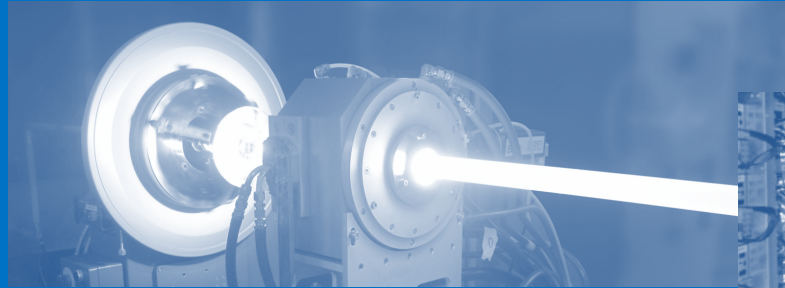
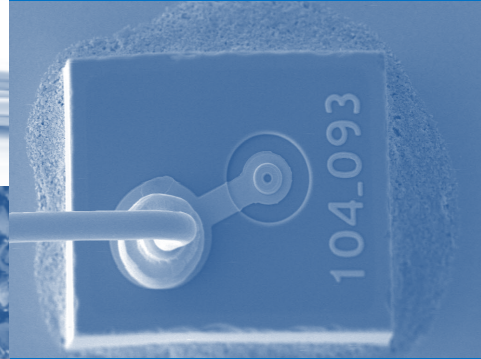
Jonathan King, Finisar

Vipul Bhatt, Finisar

Dale Murray, LightCounting

Supporters (x Individuals from y companies)

Straw Polls





Back Up

