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

*In-progress CFI Consensus Presentation Draft*

Robert Lingle Jr., OFS

September 30, 2019

NEA Ad Hoc Teleconference

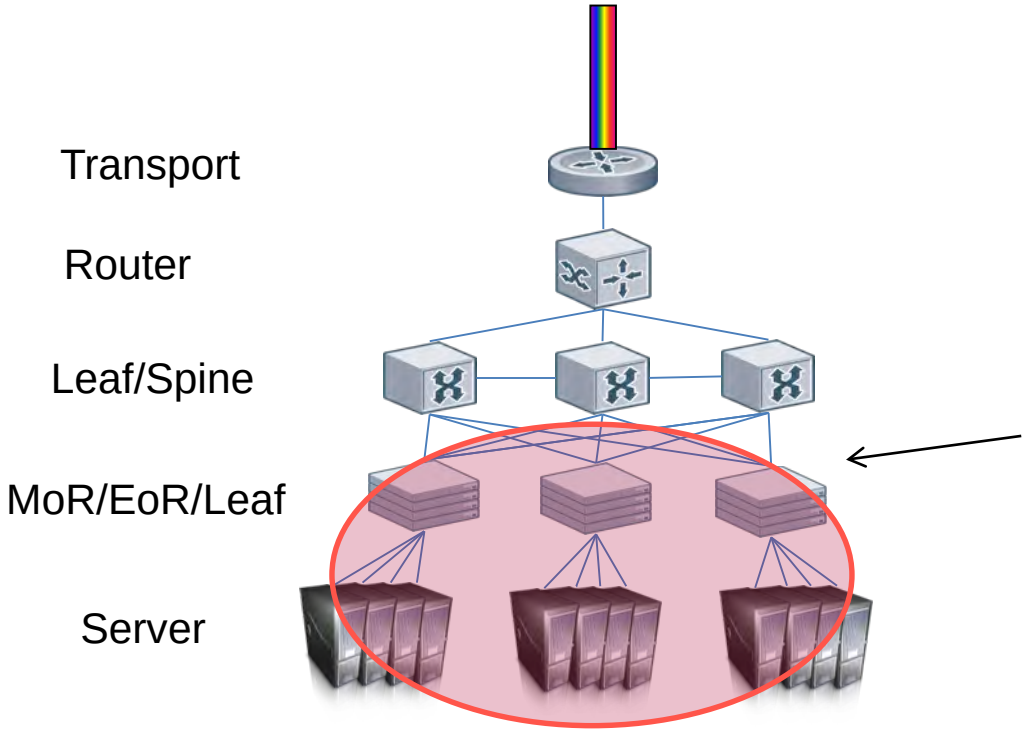
# CFI objectives

- To measure the interest in addressing:
  - lower cost, short 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, the drive to higher switch ASIC throughput and SerDes rates is increasing the number of ports per switch.
- This proposed study group would look at short reach (TBD) MMF and/or SMF PHYs using 100G per wavelength to match emerging 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 short-reach 100G PMDs include connectivity in Cloud Service Provider data centers for:

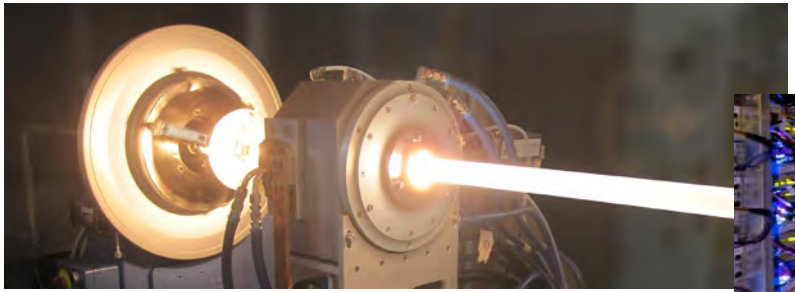
- **Server-to-switch**
- **Accelerator-to-switch**
- **Switch-to-switch**

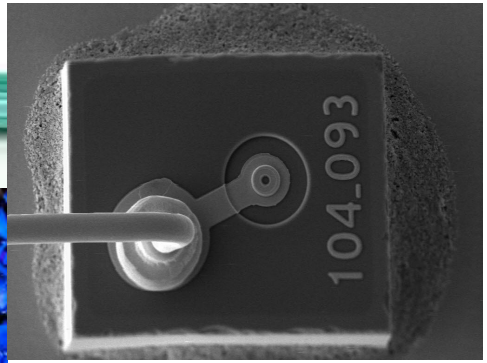
# Agenda

- **Presentations** (tentative)
  - **Market Drivers**
  - **Technical Feasibility**
    - Ramana Murty (Broadcom)
    - Vipul Bhatt (Finisar)
  - **Why Now?**

- **Straw Polls**

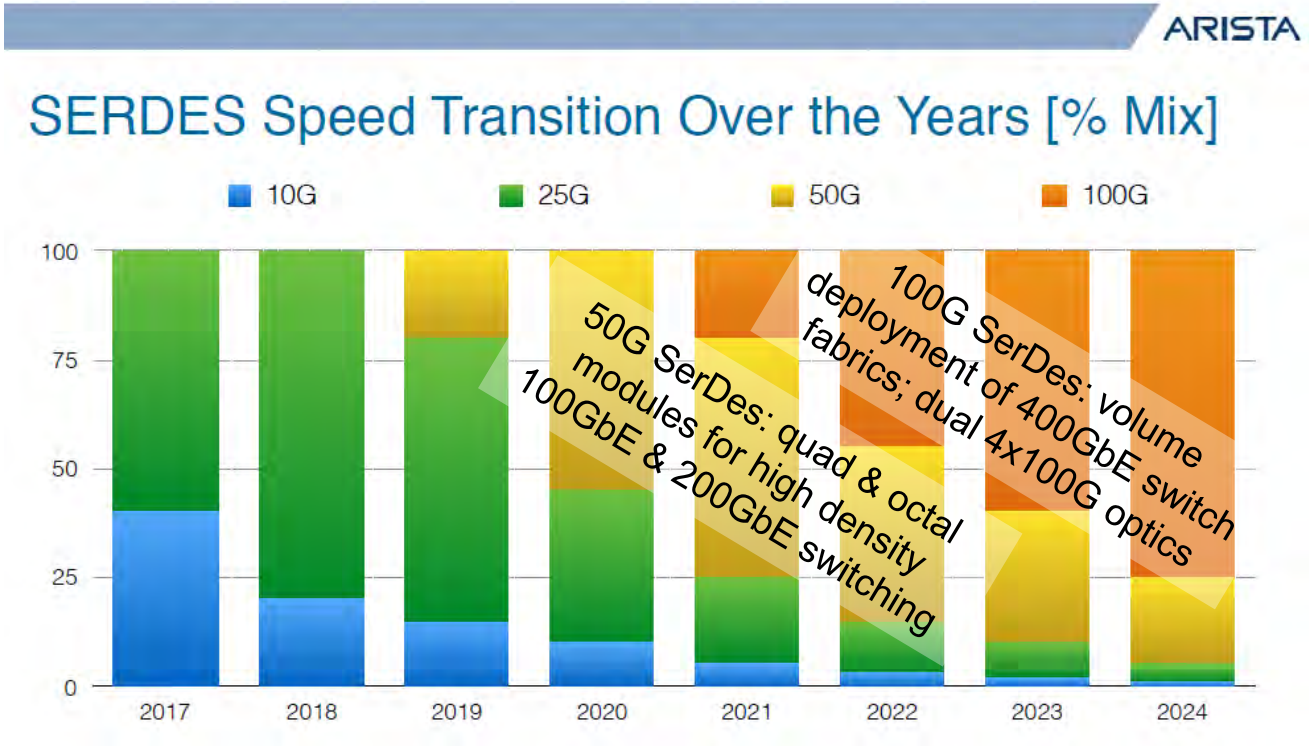
# Market drivers





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

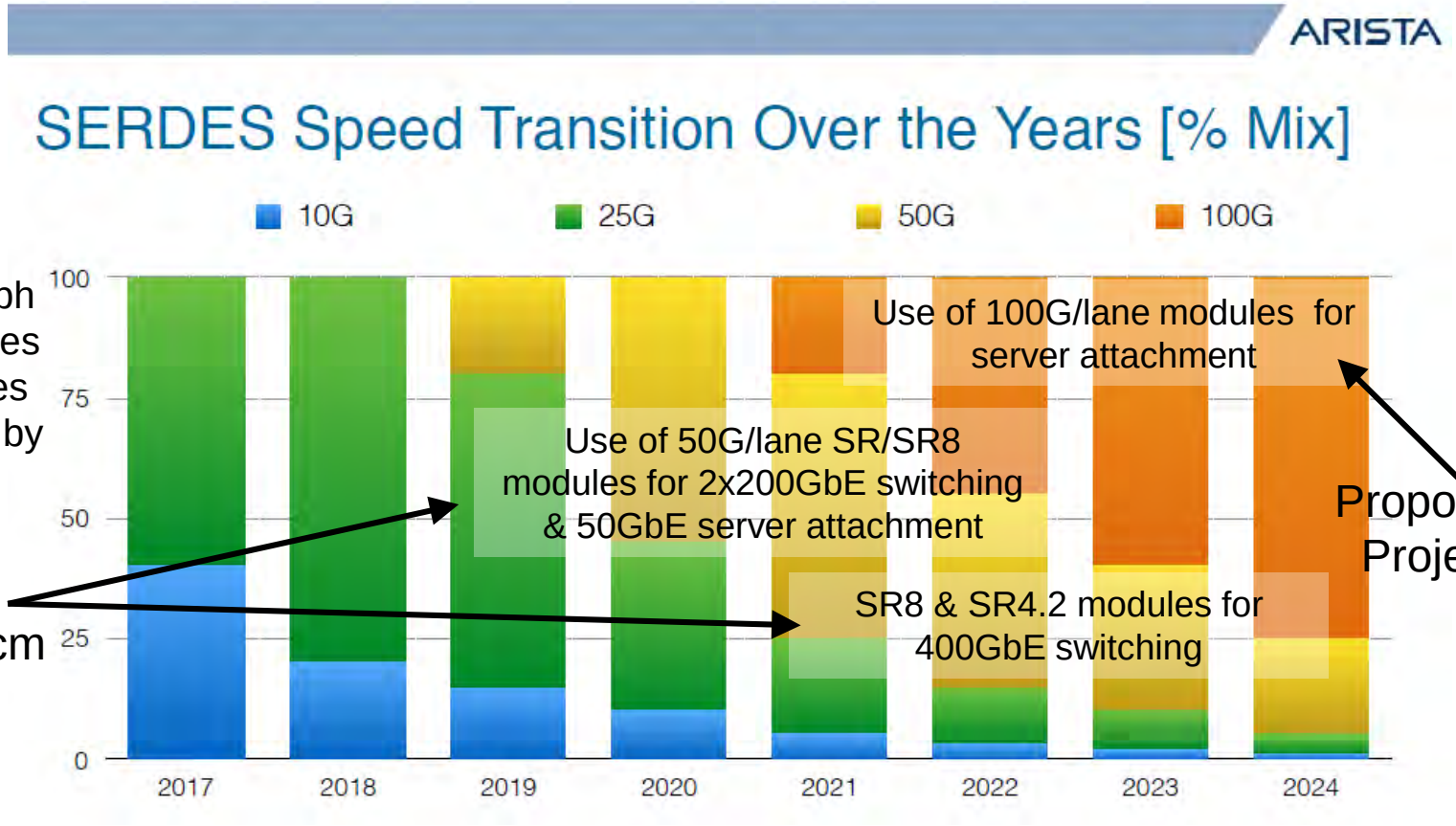
Y-axis of graph estimates sales mix of SerDes speeds by % by year




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

(Annotations on graph by Lingle)

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



# Servers also move to higher speeds over time

Placeholder for slide on evolution of SerDes speeds in server NIC cards

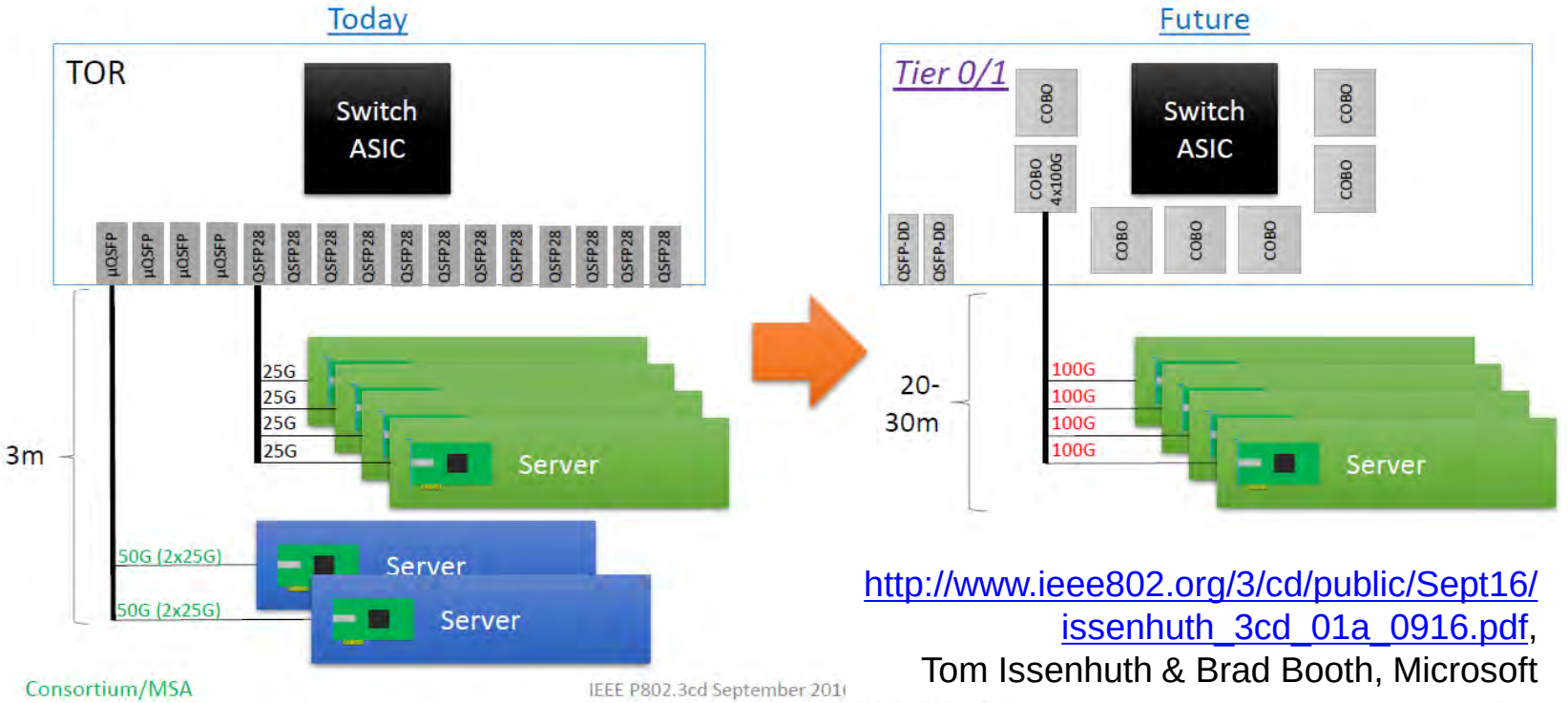
# Optimized server architectures evolve with server and switch technology

- Servers are moving from 25 to 50 to 100 GbE links
- Passive copper reach decreases with increasing lane speed
- 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 server connection from ToR to MoR/EoR may allow higher utilization of switch ports & lower cost deployment of redundancy

# Drive towards higher switch capacity provides higher port density for server attachment

- A 12.8T ASIC with 256x50G SerDes can support 32x400GbE, 64x200GbE, or 128x100GbE links.
- Technologies developed for supporting 400GbE using 50Gb SerDes are primarily being used today to support higher radix 100GbE and 200GbE switch fabrics.
  - An 8x50G pluggable also supports 1x400GbE, 2x200GbE, or 4x100GbE links
  - Reverse gear boxes are used to connect 2x50G lanes to 4x25G QSFP28 at Facebook
  - Volume deployment of 400GbE switch fabrics probably awaits 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 require compact optical cable and ease of breakout over 20-30m




http://www.ieee802.org/3/cd/public/Sept16/issenhuth_3cd_01a_0916.pdf,
Tom Issenhuth & Brad Booth, Microsoft

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

- Market need
  - Low-cost interconnect between switches with 32 × 800G-capacity ports (expected in 2020).
    - Passive copper cable limited to (1-2?) m. Active copper cable limited to ~ 5m.
    - A possible 16x50G/λ PMD would have twice the lane count & an unusual higher fiber-count connector.
    - An 8x100G/λ module would be useful even if maximum distance is 30 m.
  - Low-cost interconnect for 100G (serial) servers (2021+)

- Use cases
  - 100GBASE-SR (nomenclature TBD)
    - SFP112 connections to for next-generation servers.
  - 400GBASE-SR4 in quad module
    - Lowest-cost, low-fiber count point-to-point connection for 400G QSFP56-DD ports
    - Breakout to 4 × 100GBASE-SR
  - Dual 400GBASE-SR4 in octal module
    - Lowest-cost, low-fiber count point-to-point connection for 2 × 400G QSFP112-DD or OSFP112 ports
    - Breakout to 8 × 100GBASE-SR

# Comments on need for short reach 100G/wavelength, from Chongjin Xie (Alibaba)

- Applications:
  - AOC (today) for server to TOR connections
  - Transceivers for TOR to leaf switch connections
- Distances:
  - 100 meters desired
  - 50 meters required for transceivers; covers 80% of TOR-LEAF switch links
  - 30 meters is currently a space for AOCs at Alibaba
  - Server connections will be longer than 2-3m in the future
- Configurations:
  - Use of breakout in the future could favor transceivers over AOC in some cases
  - Need for breakout depends on network architecture
- Cost & power concerns
  - Cost < 50% of DR
  - Power consumption ~ 50% of DR

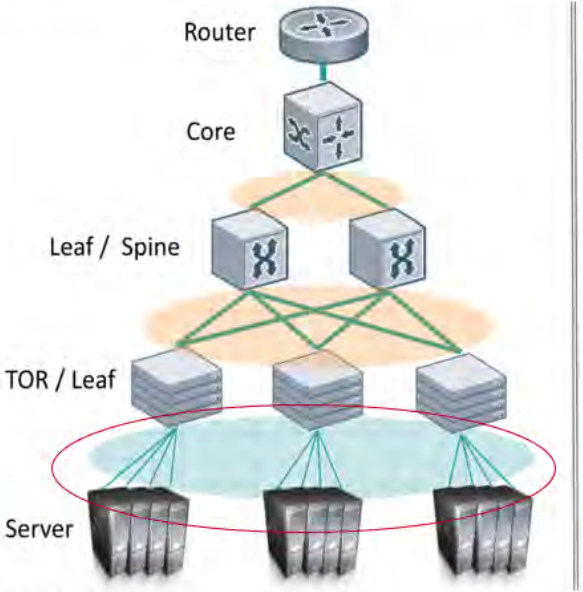
# Switch-to-switch applications benefit from 50m reach target

- Several large data center operators in China propose that a 50m reach would cover a large percentage of their TOR to T1 switch links and demonstrate a positive ROI based on their current deployment of MMF and switching topologies
- Company A: 80%
- Company B: 40%
- Company C: 100%
- Company D: 100%

# The ODCC recently completed a study of “Next Generation Data Center Connections in China” and shares these excerpts (1 of 2)

## Evolution of server & GPU Interconnects

开放数据中心委员会
Open Data Center Committee




### 1. Bandwidth

- Datacom companies started deploying 25G servers in 2017 with 100G ToR switches as the uplinks. Enterprise customers delay for 2~3 years.
- As for now, 25G access still dominates the market; however 100G servers are expected to be deployed early next year.
- It is expected that the scale will be increased around 21Q4.
- The main application will be GPU clusters.
- 50G is not being considered.

### 2. Connection (server to ToR)

- 5m within cabinet; a small number of cross-rack interconnects up to 20m.
- DAC interconnect is currently used for 25G access.
- For 100G access, due to constraints of distance and deployment (the diameter becomes thicker, the degree of buckling and the compatibility interoperating testing between vendors become complicated), server connections may turn to AOC or multi-mode transceivers.

Courtesy of Guo Liang and Jie Li from CAICT

http://www.odcc.org.cn/introduction-en.html 12

# The ODCC recently completed a study of “Next generation Data Center Connections in China” and shares these excerpts (2 of 2)

## Conclusion

开放数据中心委员会
Open Data Center Committee

- The industry needs to be ready in advance for next generation 100G server access and 800G interconnect technologies to support the evolution of cost-efficient connections for data center networks in the next 5~6 years, allowing them to cope with the development of internet services, HPC, AI and distributed storage applications.
- Due to its large market share, MMF might still have a cost advantage and is expected to evolve to 100G per wavelength.
  - 30m is a reasonable first goal for MMF links for AOC or transceivers in server interconnects.
  - Longer distance MMF transceivers are needed for switch interconnects with transceivers.
- Pluggable is preferred in 800G while co-package is an inevitable trend, however technical challenges should be resolved by the entire industry before adoption.
- With low complexity and low power consumption, IM-DD might be the choice for 800G below 2km, while Ethernet coherent might be a new possibility for data center inter-connects.

Courtesy of Guo Liang and Jie Li from CAICT

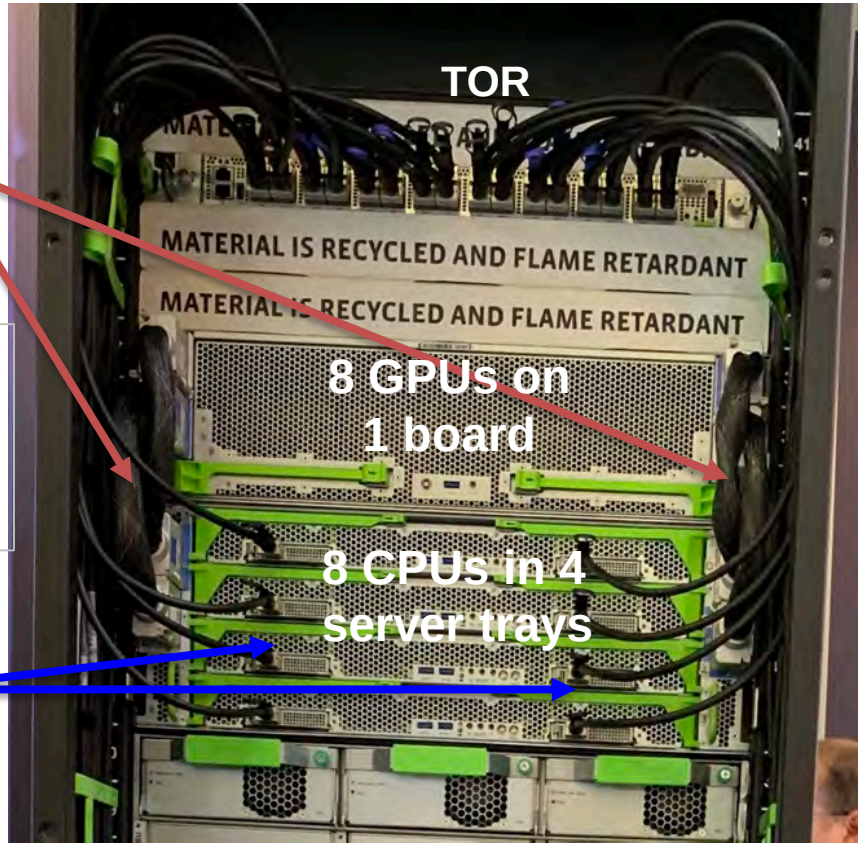
http://www.odcc.org.cn/introduction-en.html 18

# Increasing sizes of AI problems require accelerator scale-out over next 5 years, stressing interconnect speed & reach (1 of 2)

Comms between GPUs & servers over 16 lanes of PCIe 3.0 over twin-ax copper

*Zion is Facebook's large-memory unified training platform for AI workloads*

Eight 100G QSFP ports



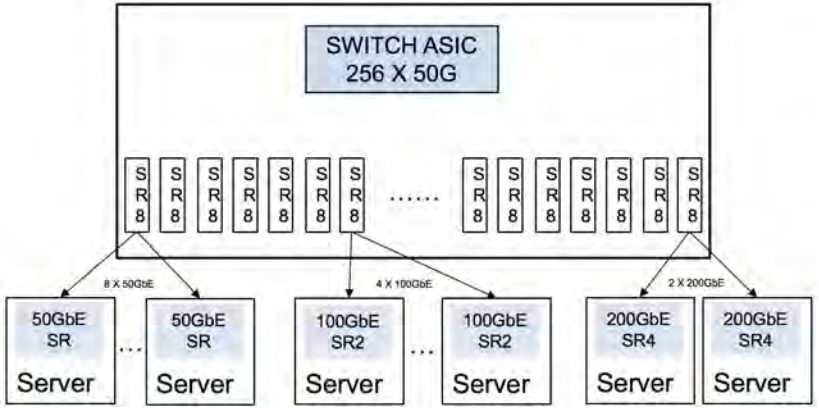

- Eight 100G connections from each cluster to TOR
- Four Zion clusters per rack, with 1.5 Tbyte memory each
- If > 1.5 Tbyte needed, problem can be spread across clusters
- Connections between racks occur through next layer of switch
- Ex. Baidu runs 10 Tbyte models now

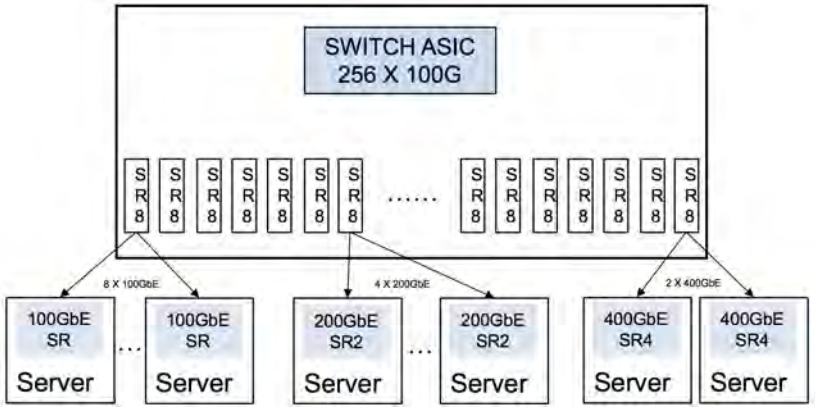
# Increasing sizes of AI problems requires accelerator scale-out over next 5 years, stressing interconnect speed & reach (2 of 2)

- Demand on compute, memory, and interconnect bandwidth have each grown by ~10x yr-over-yr recently, in the AI sector
- Ethernet is a good protocol for accelerator cluster interconnects
- In the FB example, 100GbE (4x25G) DAC cables connect each server/GPU pair to the ToR (OCP Wedge100 w/ 32x100GbE ports).
  - Scaling a problem out within the rack requires one switch hop
  - Scaling a problem outside the rack requires three or more switch hops
  - Future demands may require faster & more efficient interconnects
- Longer reach and higher speed optical Ethernet links will support larger clusters with flatter interconnects to support larger machine learning problems of the future

# Why use transceivers? (p1 of 2)

- Server attachment rates can be selected by grouping a number of SR8 ports together as required with structured cabling
- Reusable as lane rates increase



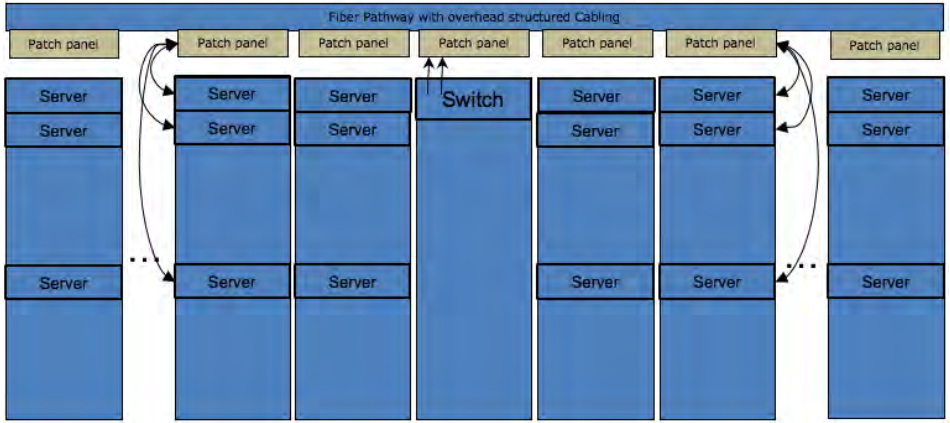




http://grouper.ieee.org/groups/802/3/NGMMF/public/Jan18/shen_NGMMF_01_jan18.pdf

# Why use transceivers? (p2 of 2)

**Supports server-row cabling objectives**

- Move switch from TOR to MOR to better consume radix (example 192 potential server connections with a 3:1 contention ratio) http://grouper.ieee.org/groups//802/3/NGMMF/public/Jan18/ghiasi_NGMMF_01_jan18.pdf
- Enable pre-installed overhead cabling that supports multiple line rate generations (50/100G)
  - Attach to overhead cabling with short cords
  - Repeat installation pattern for all server racks for installation efficiency of ≤ 5 hours for a server row - *Rich Baca (Microsoft)*
  - Allow breakouts in structured cabling to support various server data rates (50/100/200G)



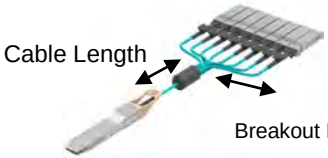

- Typical server row 16 – 20 cabinets
- Cabinets arrive on site with servers installed
- Overhead cable is pre-installed with pathway
- Simple patching from server to overhead patch panel

# Why not use DACs or AOCs?

**Inherent limitations of DACs and AOCs**

- DACs are bulky
  - Congest pathways
  - Difficult to bend and route compared to fiber
- DACs too short for MoR switch
  - Reach limited to ~ToR switch placement
- AOCs require on-site installation
  - Must route transceiver ends thru pathways
  - Longer AOCs hinder deployment speed
- AOCs with breakouts even more difficult
  - Breakout involves routing multiple transceiver ends
  - Endpoint location diversity becomes challenging


AOC/DAC OFC 2019




Number of Servers / Speed

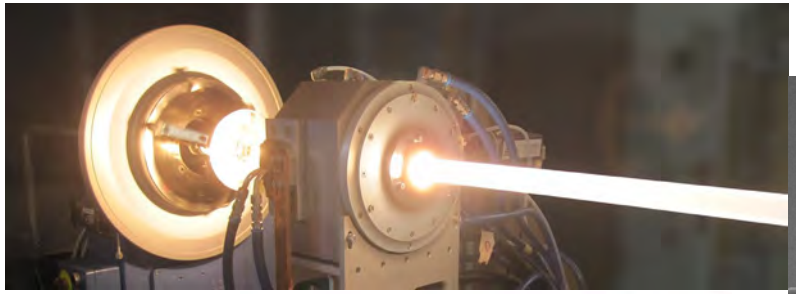
DAC Reach/Size

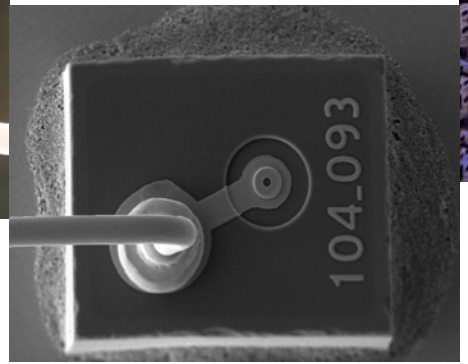
# Cost & density benefits accrue from higher lane speeds

- 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 short reach links (target length is TBD, but likely not longer than 50m) optimized for lower cost, lower power, such as:
  - SR, SR2, SR4, and/or SR8-style PMDs, based on 100G VCSEL & MMF; BiDi also possible
  - Single-lane and/or parallel PMDs using lowest-cost 100G/wavelength over SMF
- For MMF links, higher speed lanes lead to reduced lane counts, reduced component counts, reduced complexity, and lower cost than previously standardized PMDs
- Some prefer pluggable transceivers, but lower cost of AOC is attractive to others. Further study is needed.

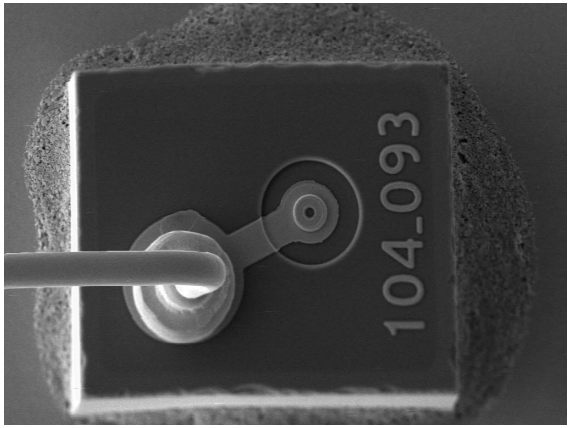
# Technical feasibility






# 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 & .3cm standardized 50G per lane links
- Ethernet does not yet address 100G VCSELs




# In the SMF case, can shorter reach permit development of lower cost, lower power interconnects compared to DR / DR4?

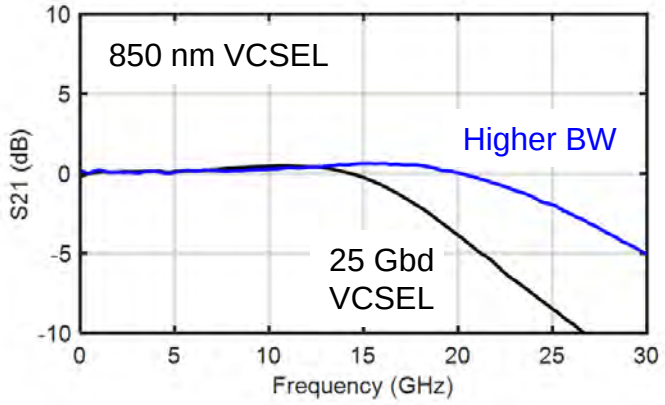
<Invitation for someone to provide a slide here.>

# PCS & FEC Re-use

- Goal is to re-use existing PCS/FEC/PMAs for this project
  - C2M support from 803.3ck, which reuses PMA/FEC/PMAs from 802.3cd and 802.3bs
  - Uses RS(544,514) FEC with 4:1 bit multiplexing to support 100G per lane interfaces
  - This supports a BER of $2.4 \times 10^{-4}$ on the PMD

# Technical Feasibility: 100G Multimode Fiber Link

- Development of 50 Gbd VCSEL is in progress.
  - Substantial increase in bandwidth over 25Gbd VCSELs for 400G-SR8 & 400G-SR4.2 achieved.
  - Targets low RIN and k-factor to reduce modal noise & MPN penalties in the link.
- 50 Gbd VCSEL can lead to lower cost solution for next-gen short reach optical links.
  - 850 nm wavelength will extend the use of OM3 & OM4 multi-mode fibers.
  - Equalization (pre-emphasis and Rx FFE) expected to reduce VCSEL bandwidth requirement.
  - Link simulations suggest feasibility of a 50m OM4 link.

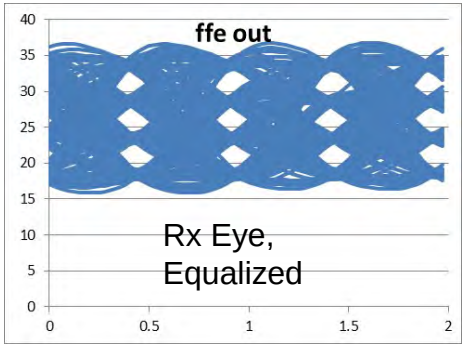


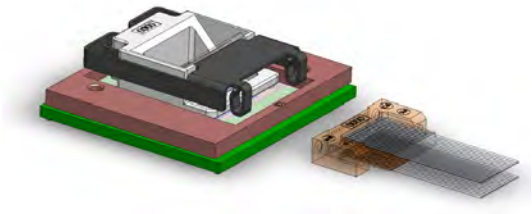

Broadcom 100G VCSEL under development

# Technical Feasibility: 100G Multimode PMD (1 of 3)

- 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 assumed for simulations
- Development of 50G NRZ multimode transceiver for High-Performance Computing is progressing well
  - Same signaling rate as 100G PAM4

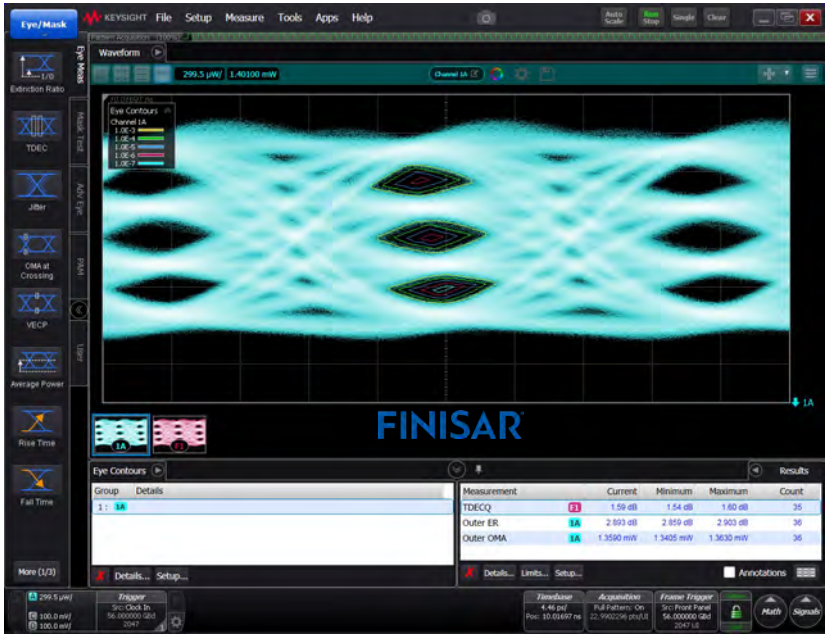
400G-SR8 demo
OFC 2019
FINISAR



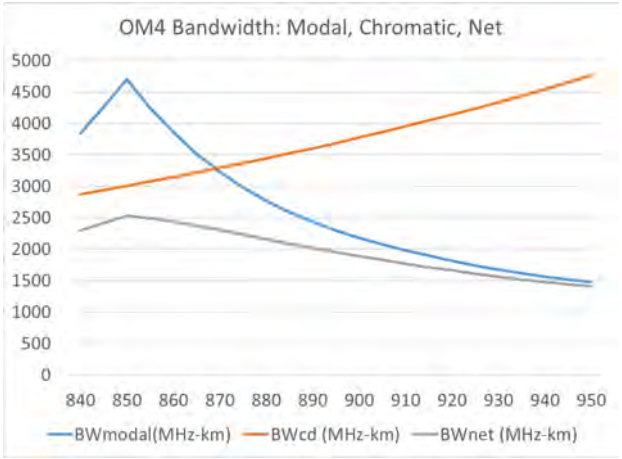
# Technical Feasibility: 100G Multimode PMD (2 of 3)

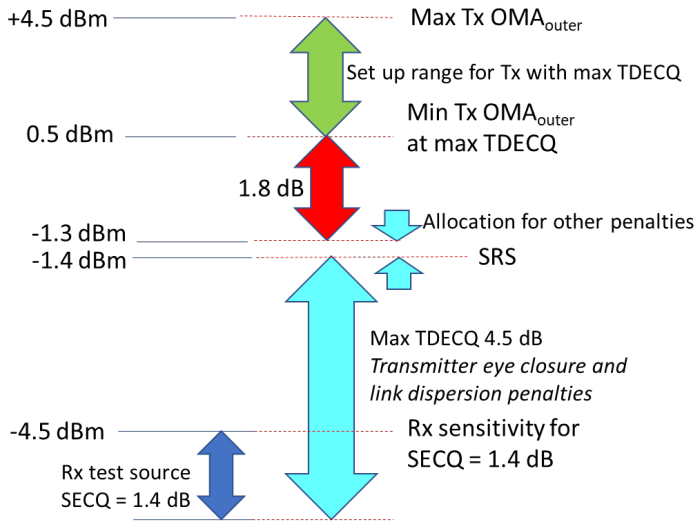
- 112G PAM4 VCSEL Tx Eye
- Early prototype demonstration
- 850 nm
- 56 Gbaud, with 39.8 GHz SIRC filter, PRBS11

# Technical Feasibility: 100G Multimode PMD (3 of 3)

- 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
  - Choice of wavelength: 850 or 940 nm; market vs. technical considerations



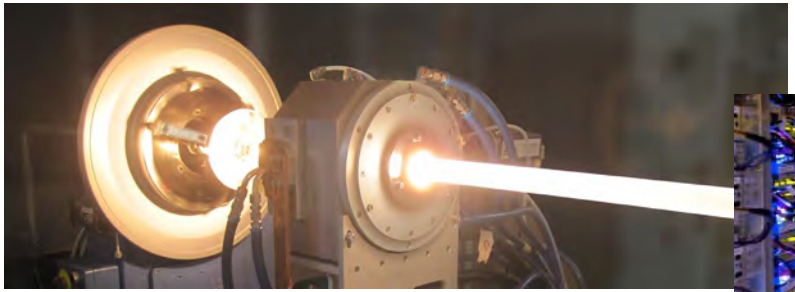

Strawman link budget presented by J. King to INCITS T11.2 in June Doc # T11-2019-00161-V000.pdf

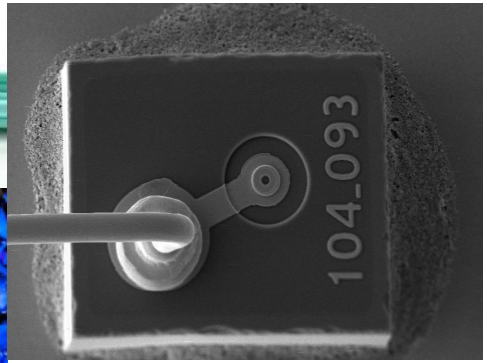
Summary: We can demonstrate that cost-effective 100G/lane multimode PMD for short reach is technically feasible.

# Re-use of Receiver Components

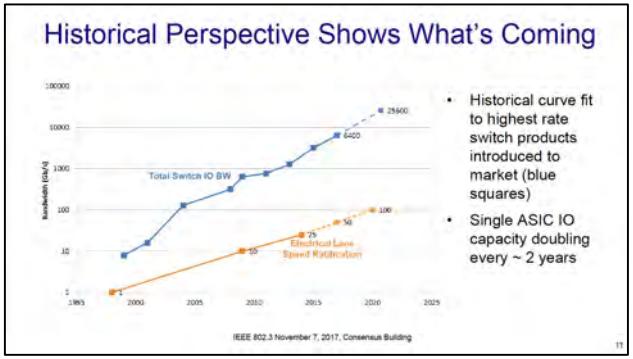
- Except for the low-cost photodiode, receiver components for 100G optics are not unique to multimode receivers
- Transimpedance amplifiers and downstream clock recovery and signal processing circuits will be the same functions as used in longwave receivers (802.3cu implementations, for example)
- In fact, such re-use will lead to further improvements in economies of volume

# Why now?

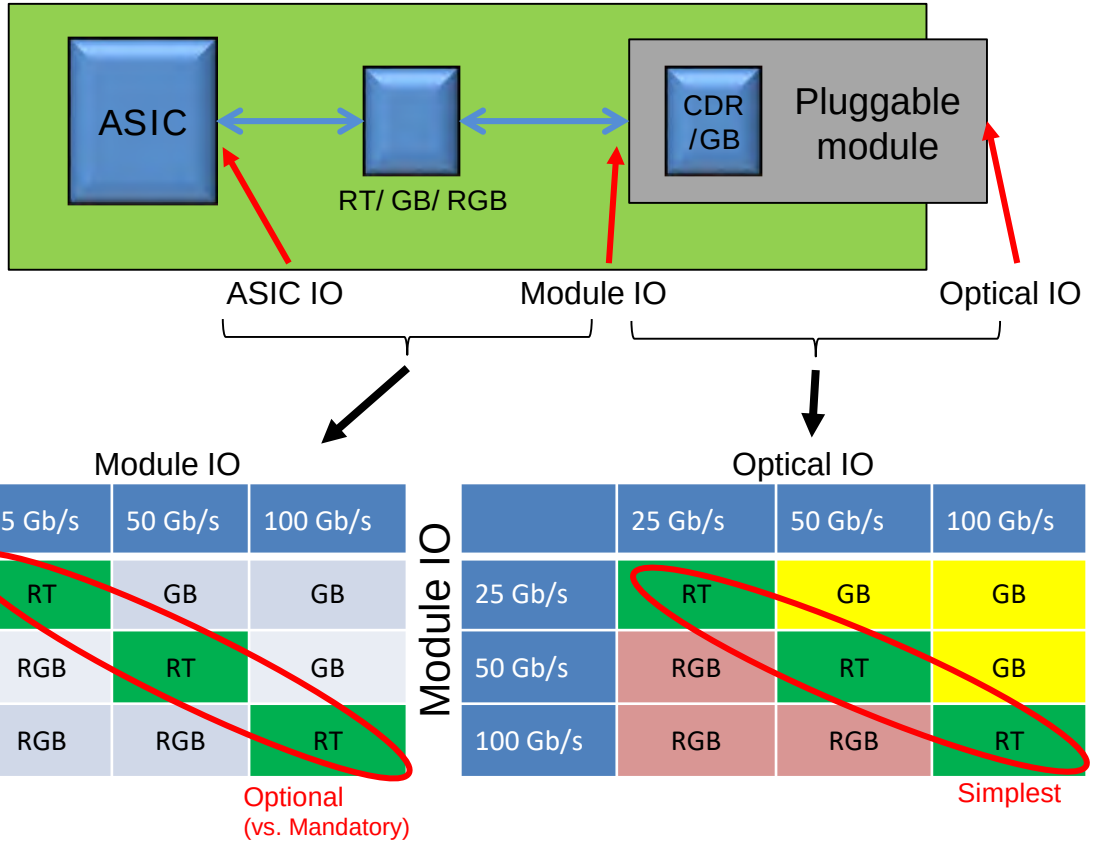




# There are strong reasons to match ASIC IO to module IO at 100G




IEEE P802.3ck's CFI:
http://www.ieee802.org/3/cfi/1117_3/CFI_03_1117.pdf

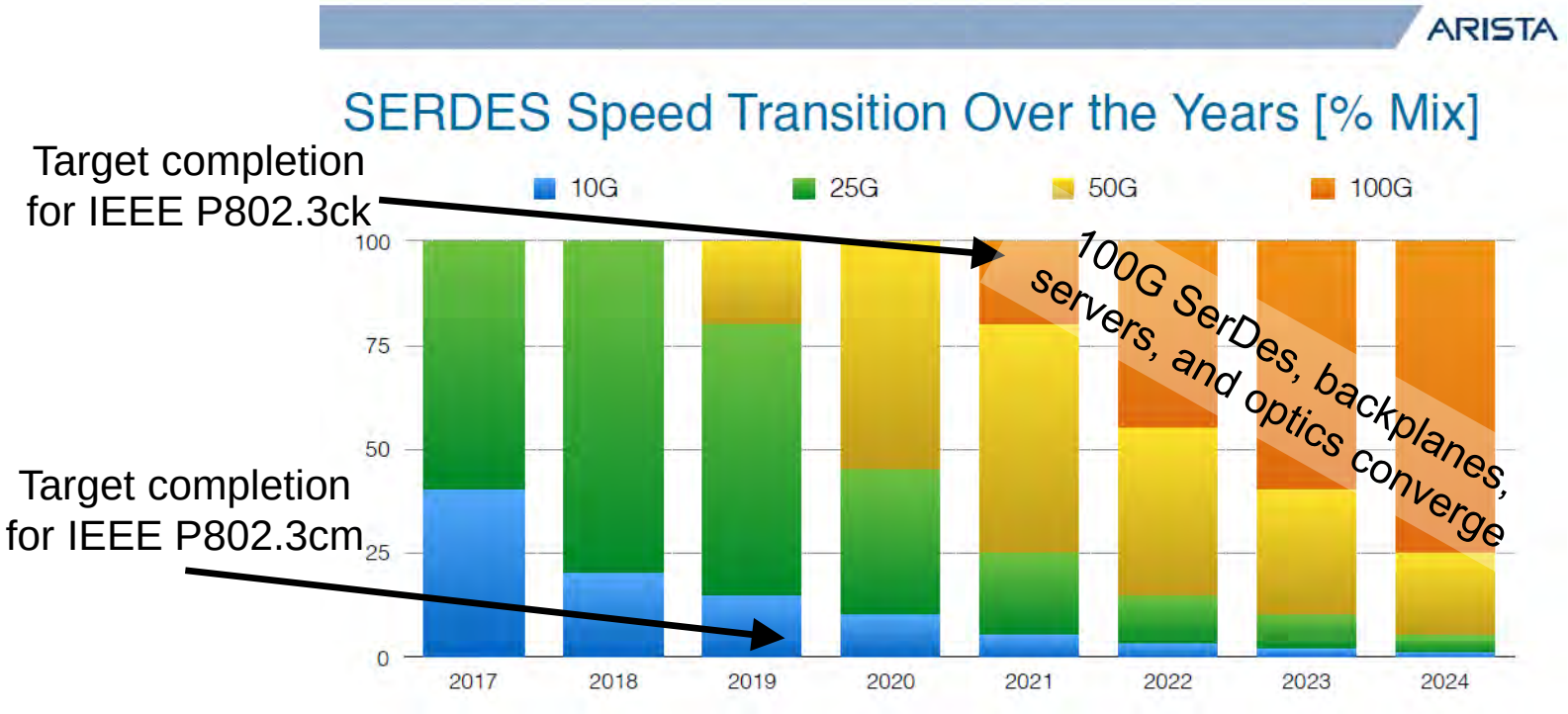



- ASIC IO "needs" to increase
- Module IO "advantage" to match ASIC IO (no mandatory extra host device)
- Optical module simplified when Optical IO matches Module IO

From M. Nowell, 100G Lambda CFI Consensus

RT: Retimer/CDR GB: Gearbox RGB: Reverse Gearbox

# Beginning a project now on 100G cost-optimized shorter-reach interconnects roughly coincides with expected roll-out of 100G SerDes, ramp of 100G servers, and higher switch fabric speeds



# It is the right time to study whether short-reach PMDs built with 100G wavelengths meet the IEEE 802.3 criteria for standardization

- It is technically feasible to build lower-cost 100G / wavelength transceivers with a reach target of possibly 50m on OM4 MMF (TBD)
- Promising applications, in datacenters built by cloud service providers, include:
  - Optics to the server / accelerator
  - Tier1 to Tier2 switch links
- The 100G per wavelength modules proposed for study could support 100, 200, and 400GbE links:
  - Using low cost & power single-mode or multi-mode technologies
  - In SFP112, QSFP112, and QSFP-DD800 form factors

# Contributors

Chongjin Xie, Alibaba
David Piehler, Dell EMC
Jonathan King, Finisar
Vipul Bhatt, II-VI
Dale Murray, LightCounting
Ali Ghiasi, Ghiasi Quantum
Mark Gustlin, Cisco
James Young, CommScope
Paul Kolesar, CommScope
Ramana Murty, Broadcom
Yan Zhuang, Huawei
Guo Liang, CAICT
Jie Li, CAICT
Peng Dong, Huawei

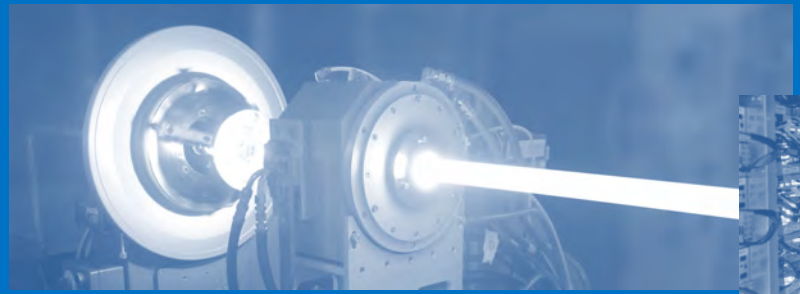
# Supporters

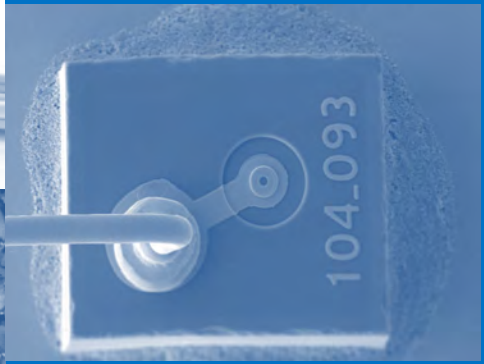
Jon Lewis – Dell EMC
Yan Zhuang - Huawei
Chongjin Xie – Alibaba
Mike Dudek – Marvell
David Piehler – Dell EMC
Steve Swanson – Corning
James Young – CommScope
Vipul Bhatt – II-VI
Ramana Murty – Broadcom
Paul Neveux – Superior Essex
Paul Kolesar – CommScope
Phong Pham – USConec
Flavio Marques – Furukawa LATAM
Guo Liang – CAICT
Jie Li – CAICT
Peng Dong – Huawei
Dale Murray – LightCounting
Kenneth Jackson – Sumitomo

Ali Ghiasi – Ghiasi Quantum
Rick Pimpinella – Panduit
Vince Ferretti – Corning
Greg LeCheminant – Keysight
Kobi Hasharoni – Dust Photonics
Pavel Zivny – Tektronix
Jose Castro - Panduit
Earl Parsons – CommScope
John Abbott – Corning
Mabud Choudhury – OFS
Ronald Nordin – Panduit
Chris Cole – II-VI
John Johnson – Broadcom
Masaru Terada – Furukawa Electric
Leon Bruckman – Huawei
David Malicoat – Senko
Jan Filip – Maxim Integrated
David Chen – AOI

Christophe Metivier - Arista
Jeff Maki – Juniper Networks
Jonathan Ingham – FIT
Piers Dawe – Mellanox

# Straw Polls





# Straw Poll 1: Call-For-Interest

Should a Study Group be formed to consider lower cost, short reach, optical PHYs using 100 Gb/s wavelengths?

Y: N: A:

Room Count:

# Participation

I would participate in the “100G Short Reach”* Study Group in IEEE 802.3.

Tally:

My company would support participation in the “100G Short Reach”* Study Group in IEEE 802.3.

Tally:

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

# Back Up