



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

CFI Consensus Presentation

Robert Lingle Jr., OFS

Mabud Choudhury, OFS

November 12, 2019

IEEE802 November Plenary

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

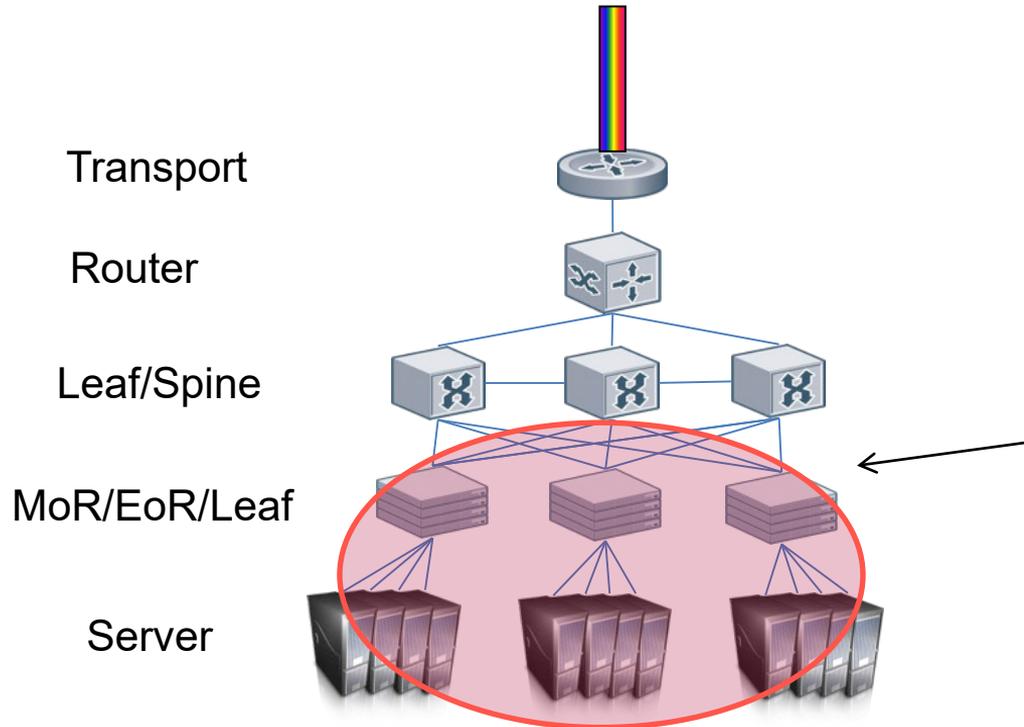
Agenda

- **Introduction**
 - Mabud Choudhury (OFS)
- **Motivation**
- **Market Drivers**
 - Robert Lingle (OFS)
 - James Young (CommScope)
- **Technical Feasibility**
 - Ramana Murty (Broadcom)
- **Why Now?**
 - Robert Lingle
- **Straw Polls**
 - Mabud Choudhury

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 supporting the server.
 - Meanwhile, the drive to higher switch ASIC throughput and SerDes rates is increasing the number of ports per switch.
- The first switch-to-switch links for early adopters are short in some regions
- 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?

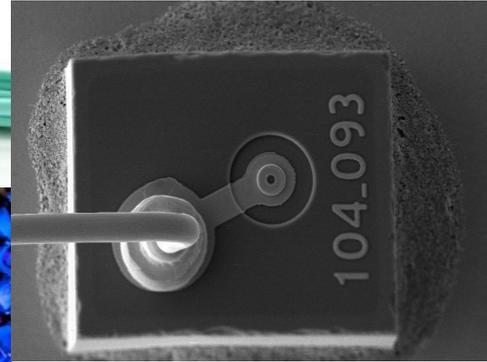
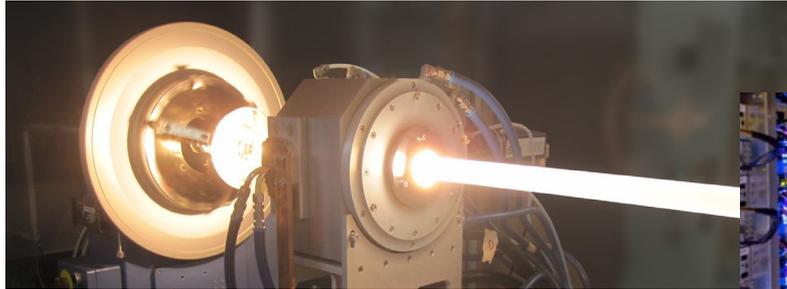


Applications for early adoption of short-reach 100G PMDs include connectivity in Cloud Service Provider data centers for:

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



Market drivers

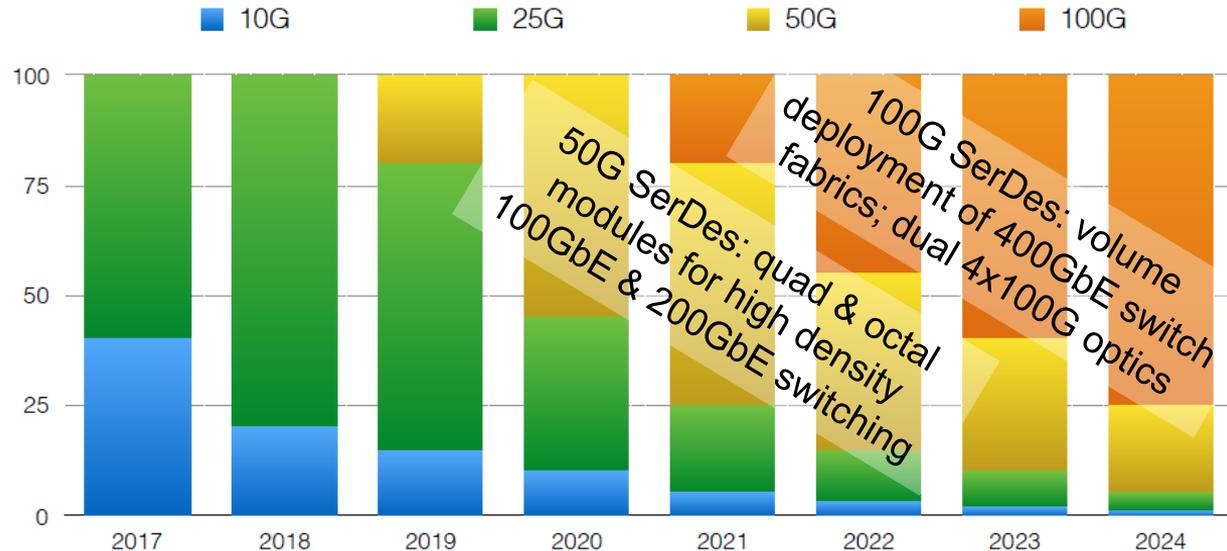


The progress of SerDes technology drives the economics & evolution of switch fabric speeds & port counts

ARISTA

SERDES Speed Transition Over the Years [% Mix]

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

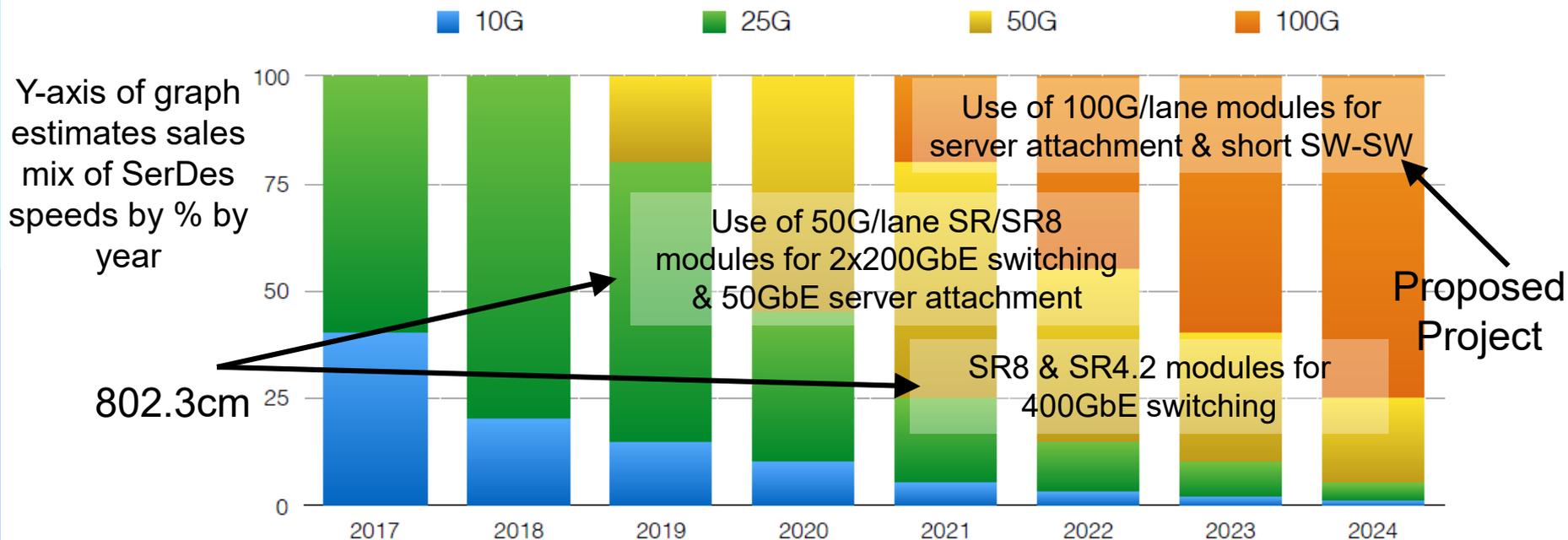


["Scaling the Cloud Network," Andreas Bechtolsheim, OCP Summit 2018](#)

(Annotations on graph by Lingle)

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

SERDES Speed Transition Over the Years [% Mix]

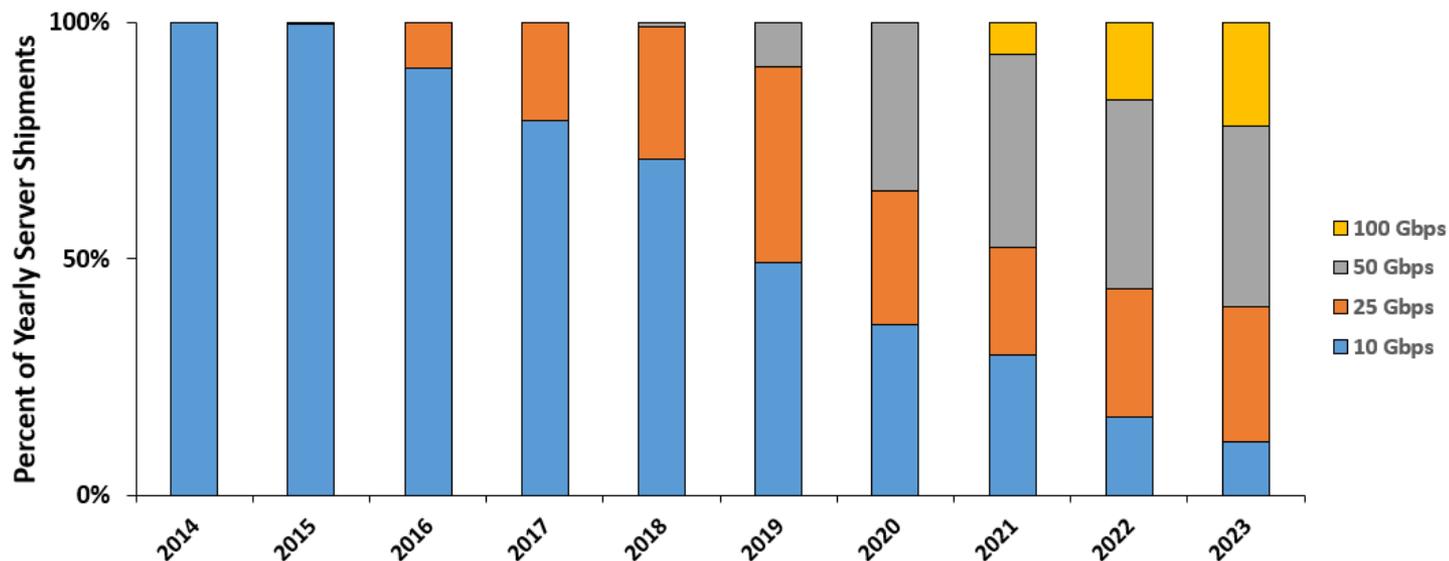


Servers also move to higher speeds over time

Server and Smart NICs: Connectivity Metrics



Server High-Speed Migration (Total Market – by SERDES speed)



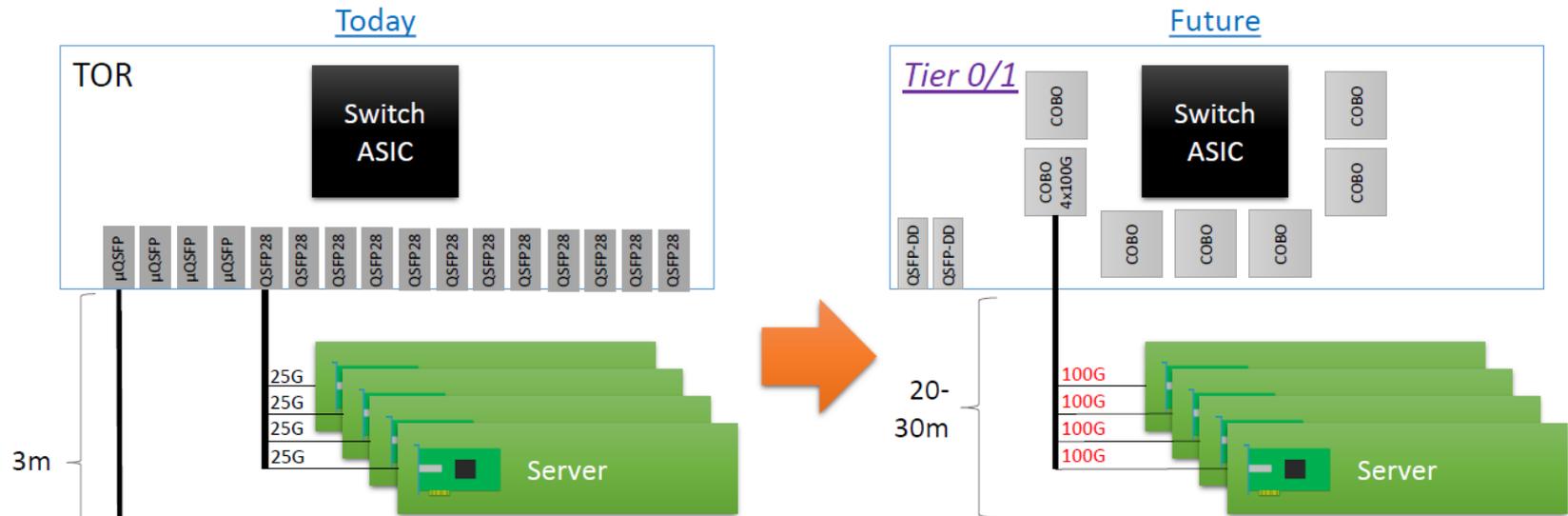
Optimized architectures evolve with server & switch technology

- Servers are moving from 25 to 50 to 100 GbE links
- Passive copper reach decreases with increasing lane speed
- As each server becomes more power hungry:
 - 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 and/or eliminate a tier of switching

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

- State-of-the-art switch technologies can support many 50G or 100G downlinks
- A 12.8T ASIC with 256x50G SerDes pairs can support 32x400GbE, 64x200GbE, or 128x100GbE links, or some combination thereof.
- Future 25.6T ASIC with 256x100G SerDes pairs can support twice as many links
- Chassis switches can leverage multiple ASICs with 50G/100G SerDes to support multiple hundreds of 50G or 100G downlinks from a MoR/EoR
- 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, requiring longer reach.
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



Consortium/MSA

IEEE P802.3cd September 2014

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 $32 \times 800\text{G}$ -capacity ports (expected in 2020).
 - Passive copper cable limited to (1-2?) m. Active copper cable limited to $\sim 5\text{m}$.
 - A possible $16 \times 50\text{G}/\lambda$ PMD would have twice the lane count & an unusual higher fiber-count connector.
 - An $8 \times 100\text{G}/\lambda$ 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 QSFP112 ports
 - Breakout to $4 \times 100\text{GBASE-SR}$
 - Dual 400GBASE-SR4 in octal module
 - Lowest-cost, low-fiber count point-to-point connection for $2 \times 400\text{G QSFP-DD800}$ or OSFP112 ports
 - Breakout to $8 \times 100\text{GBASE-SR}$

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

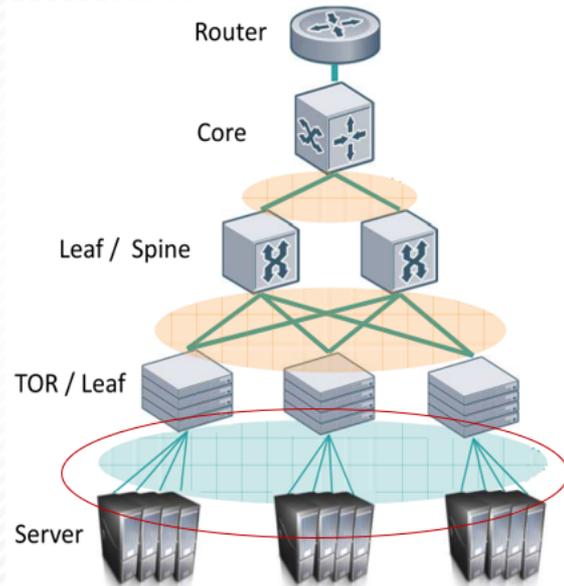
- Applications:
 - AOC (today) for server to ToR connections
 - Transceivers for ToR to leaf switch connections
- Distances:
 - 100 meters desired
 - 50 meters required for transceivers; covers 80% of ToR-LEAF switch links
 - 30 meters is currently a space for AOCs at Alibaba
 - Server connections will 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



1. Bandwidth

- Dataroom 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 Liang Guo & Jie Li from CAICT

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

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

Conclusion



- 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 interconnects.

Courtesy of Liang Guo & Jie Li from CAICT

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

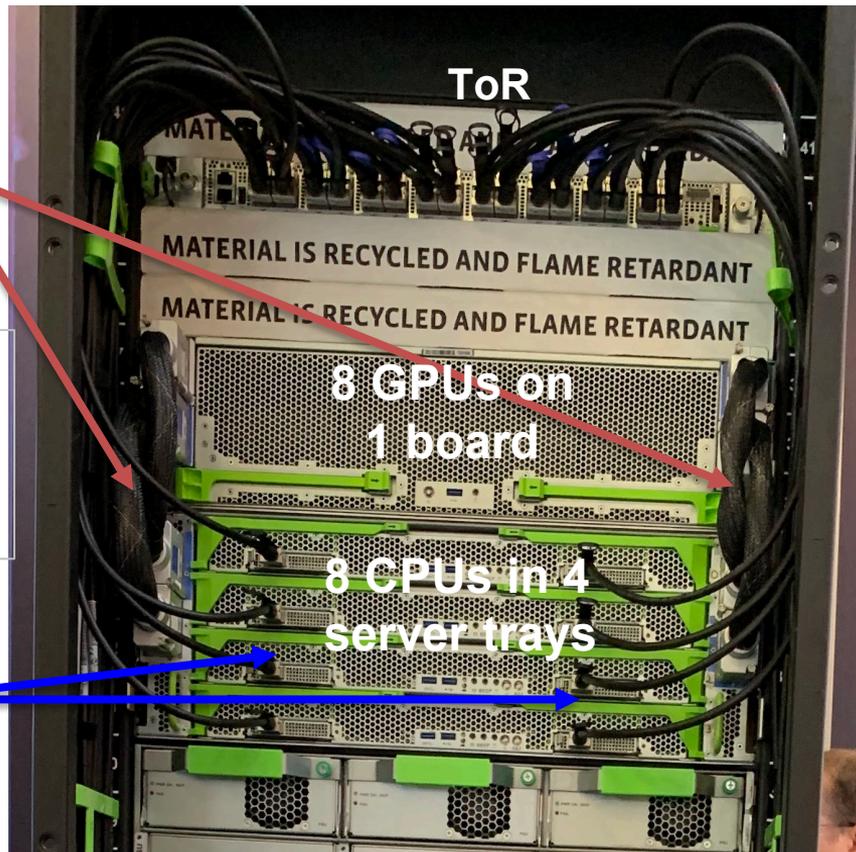
18

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



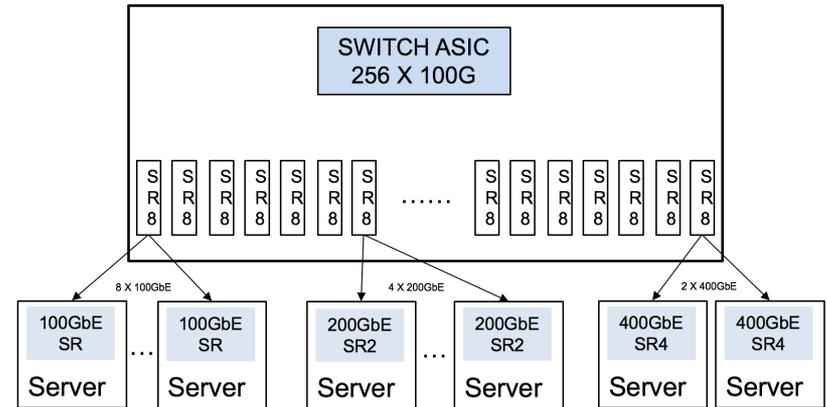
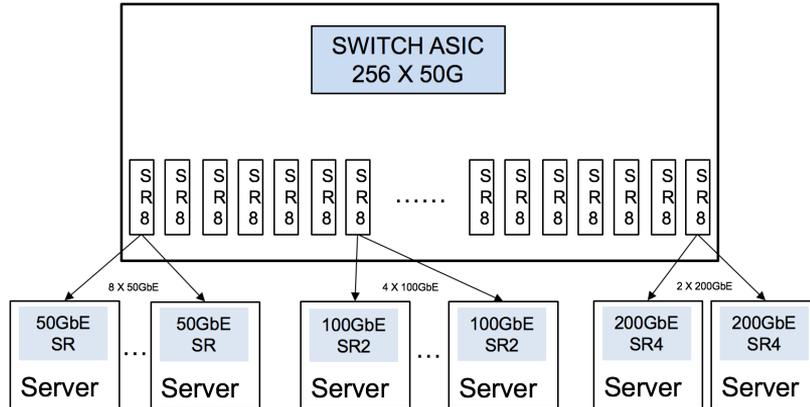
- Eight 100G connections from each cluster to ToR
- Four Zion clusters per rack, with 1.5 Tbyte memory each
- If > 1.5 Tbyte needed, problem can be spread across clusters
- Connections between racks occur through next layer of switch
- Ex. Baidu runs 10 Tbyte models now

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? (1 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

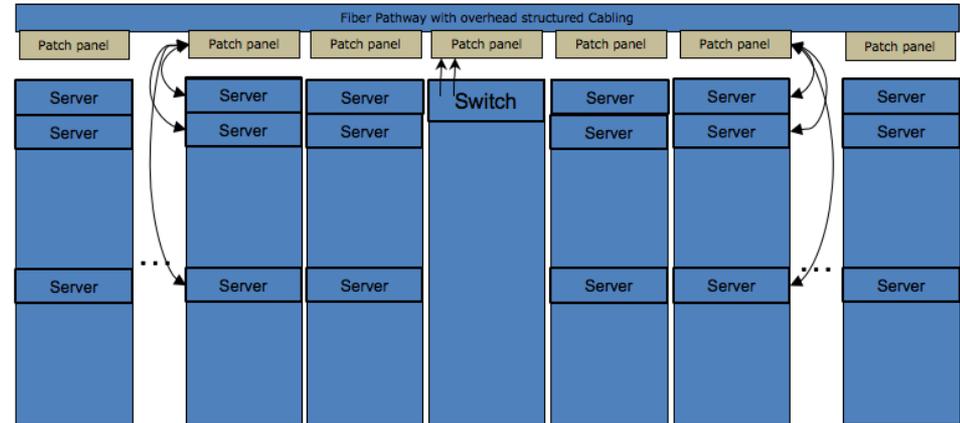


http://grouper.ieee.org/groups/802/3/NGMMF/public/Jan18/shen_NGMMF_01_jan18.pdf

Why use transceivers? (2 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

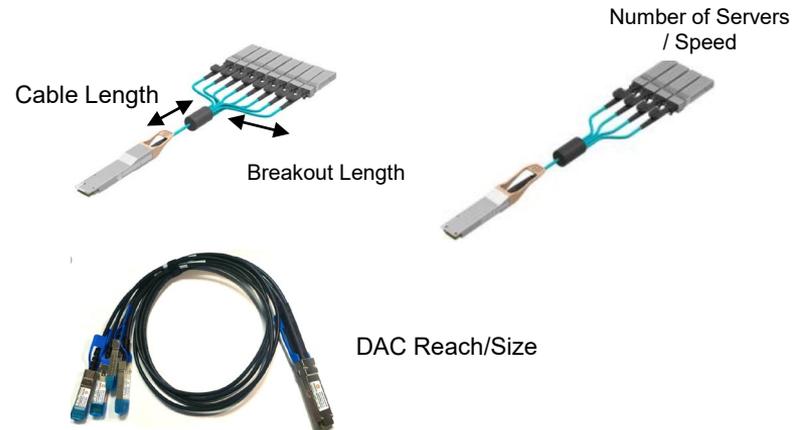
Why not use DACs or AOCs?

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

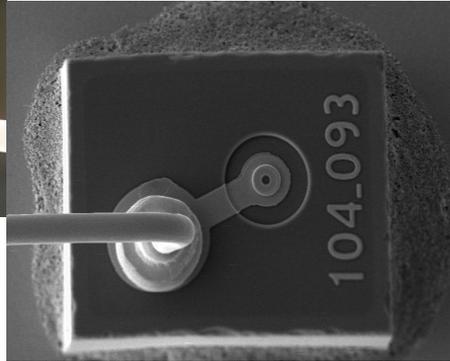
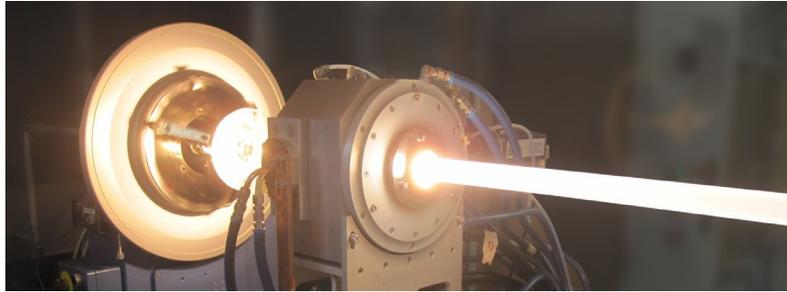


Cost & density benefits accrue from higher lane speeds

- 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.

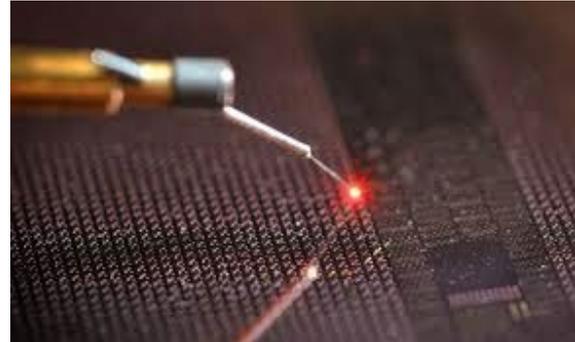
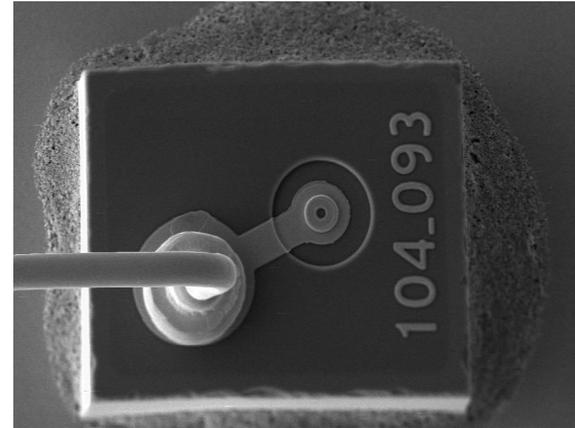


Technical feasibility



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

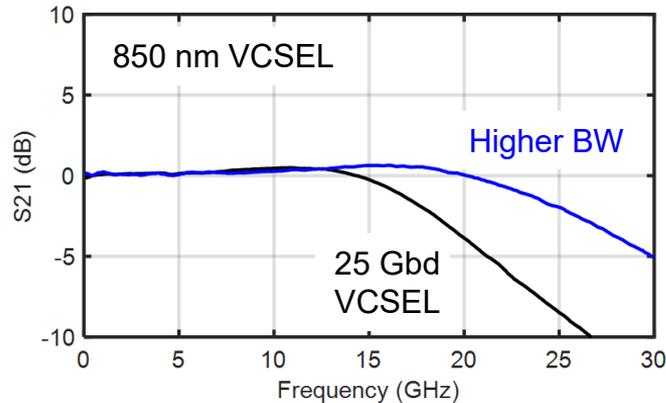


PCS & FEC Re-use

- Goal is to re-use existing PCS/FEC/PMAs for this project
 - C2M support from 802.3ck, which reuses PCS/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 2.4×10^{-4} on the PMD

Technical Feasibility: 100G Multimode Fiber Link

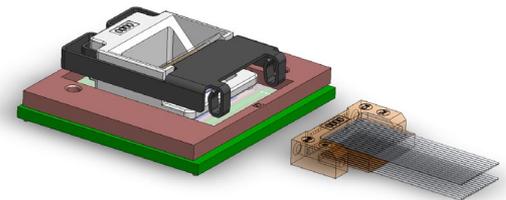
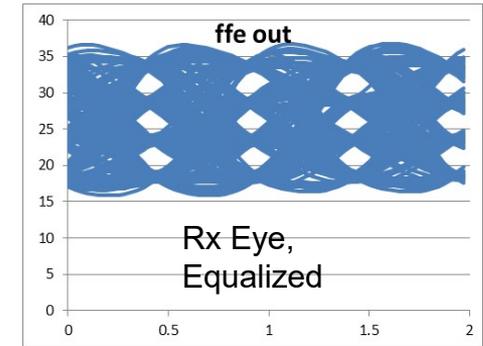
- Development of 50 Gbd VCSEL is in progress.
 - Substantial increase in bandwidth over 25Gbd VCSELs for 400G-SR8 & 400G-SR4.2 achieved.
 - Targets low RIN and k-factor to reduce modal noise & MPN penalties in the link.
- 50 Gbd VCSEL can lead to lower cost solution for next-gen short reach optical links.
 - 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

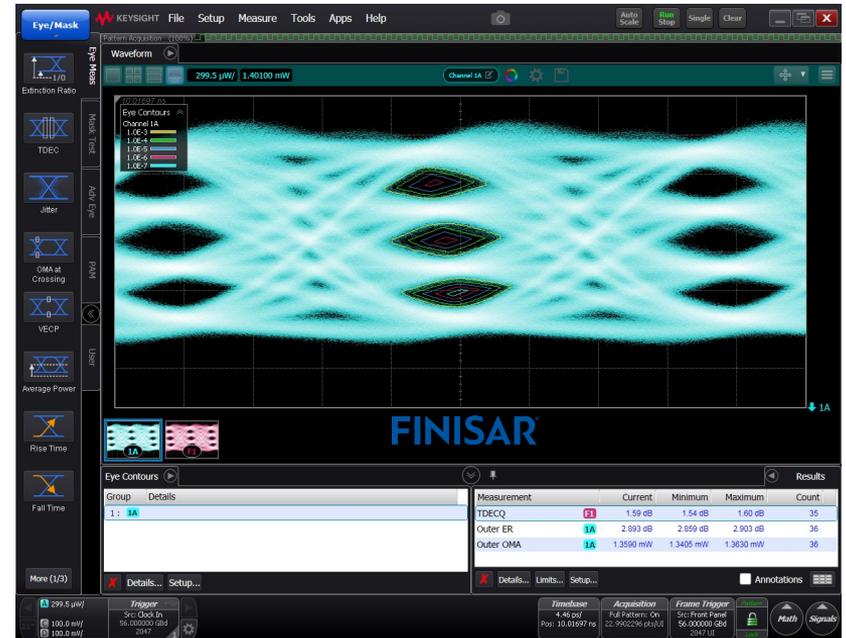
Technical Feasibility: 100G Multimode PMD (1 of 3)

- 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



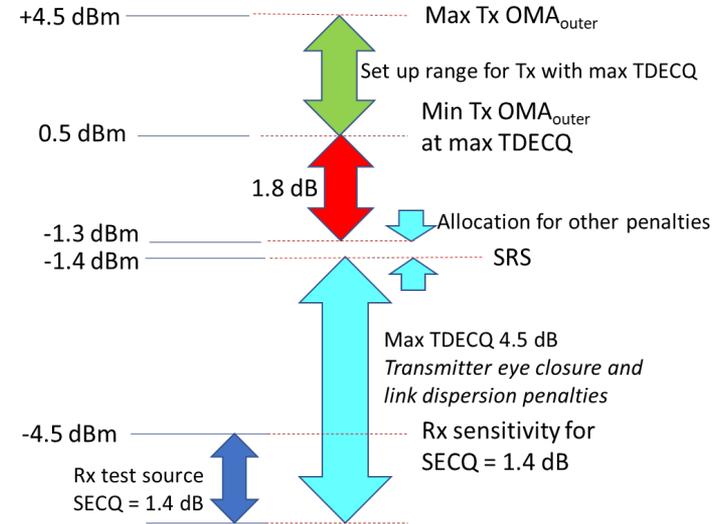
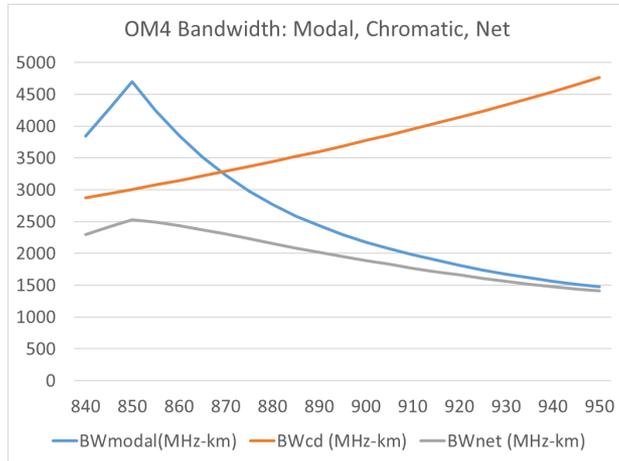
Technical Feasibility: 100G Multimode PMD (2 of 3)

- 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

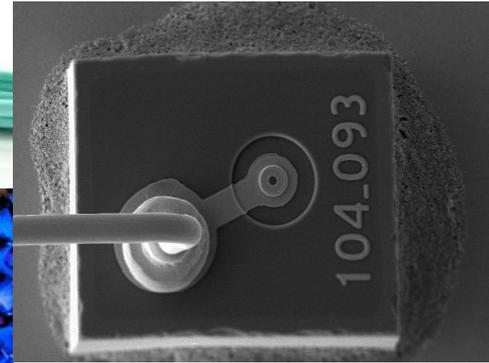
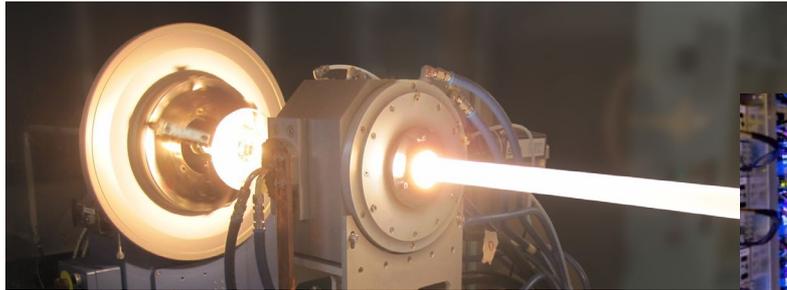
Summary: We can demonstrate that cost-effective 100G/lane multimode PMD for short reach is technically feasible.

Re-use of Receiver Components

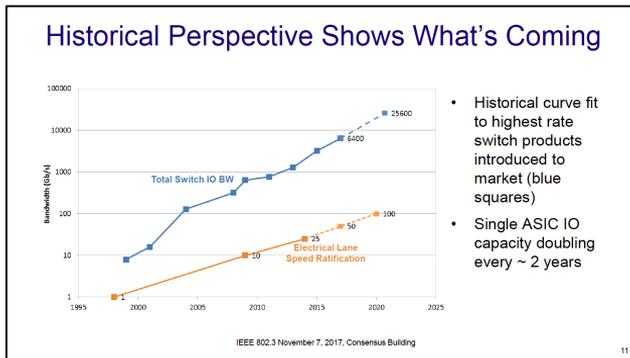
- Except for the low-cost photodiode, receiver components for 100G optics are not unique to multimode receivers
- Transimpedance amplifiers and downstream clock recovery and signal processing circuits will be the same functions as used in longwave receivers (802.3cu implementations, for example)
- In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



Why now?



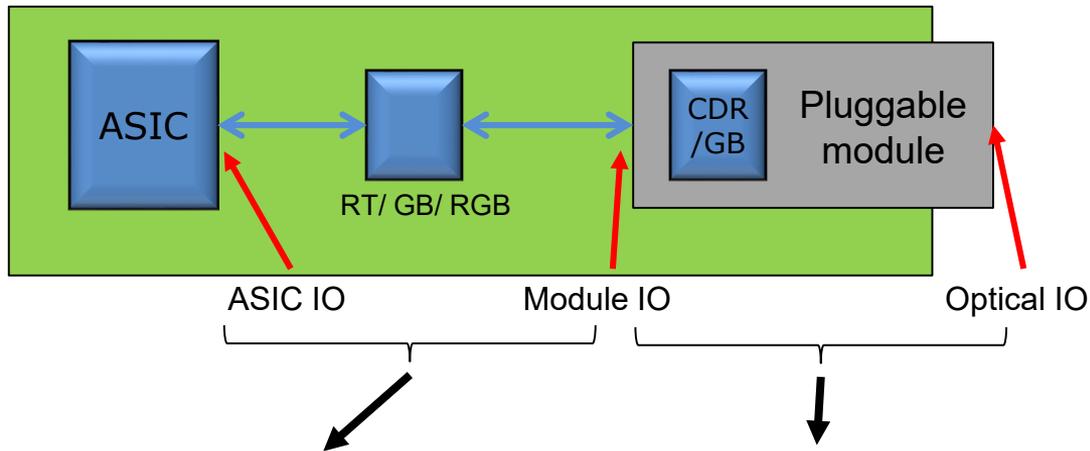
There are strong reasons to match ASIC IO to module IO at 100G



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



		Module IO			Optical IO			
		25 Gb/s	50 Gb/s	100 Gb/s	25 Gb/s	50 Gb/s	100 Gb/s	100 Gb/s
ASIC IO	25 Gb/s	RT	GB	GB	RT	GB	GB	GB
	50 Gb/s	RGB	RT	GB	RGB	RT	GB	GB
	100 Gb/s	RGB	RGB	RT	RGB	RGB	RT	RT
	Module IO							

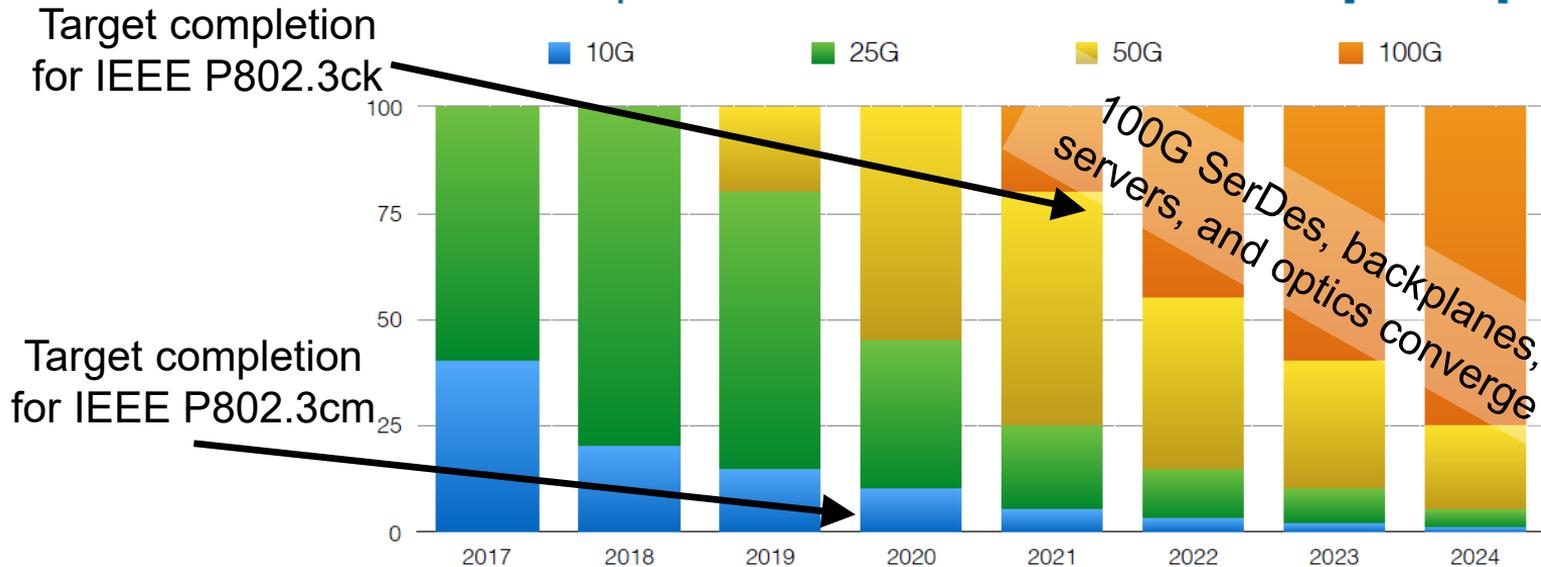
Optional (vs. Mandatory)

Simplest

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

ARISTA

SERDES Speed Transition Over the Years [% Mix]



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

Liang Guo, CAICT

Jie Li, CAICT

Peng Dong, Huawei

Supporters (55 supporters with 38 affiliations)

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

Carl Rumbolo – Wells Fargo

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

Ronald Nordin – Panduit

Chris Cole – II-VI

John Johnson – Broadcom

Masaru Terada – Furukawa Electric

Leon Bruckman – Huawei

David Malicoat – Senko

Jan Filip – Maxim Integrated

Jonathan Ingham – Broadcom

Rich Mellitz – Samtech

Rich Baca – Microsoft

Christophe Metivier - Arista

Jeff Maki – Juniper Networks

Piers Dawe – Mellanox

Mark Nowell – Cisco

Ray Nering – Cisco

Kent Lusted – Intel

Greg McSorley – Amphenol

David Chen – AOI

Tom Palkert – MACOM

Adrian Butter – Marvell

Dean Wallace – Marvell

Peter Stassar – Huawei

John D’Ambrosia – Futurewei, U.S.
subsidiary of Huawei

Nathan Tracy – TE Connectivity

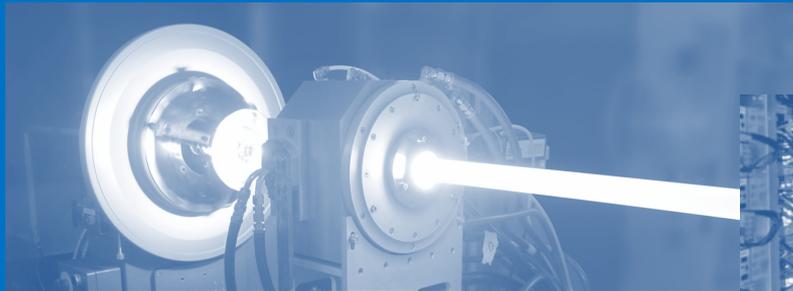
Scott Sommers – Molex

Dave Lewis – Lumentum

Hai-Feng Liu – HG Genuine



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: 55 N: 0 A: 10

Room Count: 65

Participation

I would participate in the “100 Gb/s per Wavelength Short Reach PHYs Study Group” in IEEE 802.3.

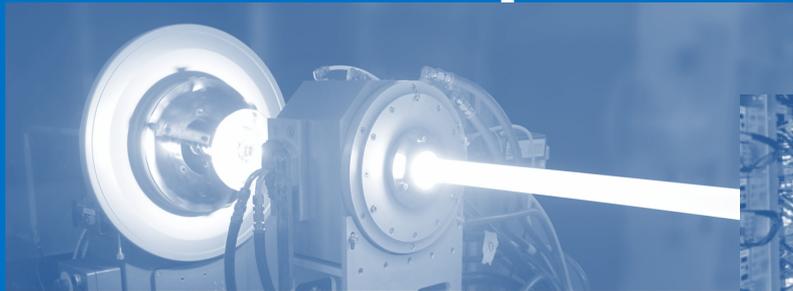
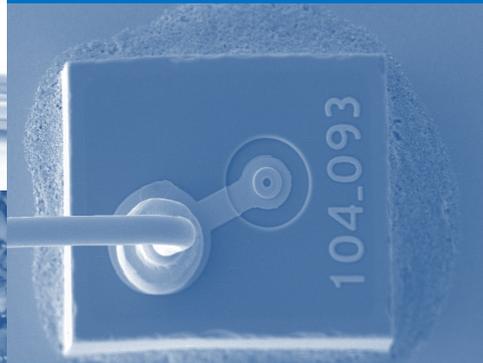
Tally: 22

My company would support participation in the “100 Gb/s per Wavelength Short Reach PHYs Study Group” in IEEE 802.3.

Tally: 18



Back Up



Call-for-Interest: 100 Gb/s Wavelength Short Reach PHYs

Matching the module I/O rate to the ASIC I/O rate reduces component count and complexity, leading to lower cost & power optical interfaces. SerDes rates on both switch ASICs and server network interface cards are expected to move to 100 Gb/s per lane in the next few years, driving current work in the IEEE P802.3ck and P802.3cu Task Forces. However, the highest speed, shortest reach PMDs over multimode fiber under development in Ethernet are built on 50 Gb/s optical lanes.

As the electrical power & interconnect bandwidth of servers & accelerators are increasing, high-radix switches are also able to connect more servers, creating a trend towards fewer servers-per-rack and thus longer server attachment connections at 100 Gb/s and higher. At the same time the achievable passive copper cable reach with 100 Gb/s lanes is expected by some experts to be shorter than the reach with 50 Gb/s lanes.

In order to develop the lowest cost and complexity optical modules for next-generation server attachment, as well as shorter switch-to-switch connections, with emerging SerDes rates, we request a Call-for-Interest to assess support for a Study Group to develop the PAR and CSD for lower cost, short reach, optical PHYs using 100 Gb/s wavelengths.