Broad market potential, economic feasibility, and distinct identity for objectives based on 100 Gb/s lanes over MMF

Robert Lingle Jr. (OFS), David Piehler (Dell EMC), Chongjin Xie (Alibaba),
James Young (CommScope)

IEEE 802.3 100 Gb/s Wavelength Short Reach PHYs Study Group

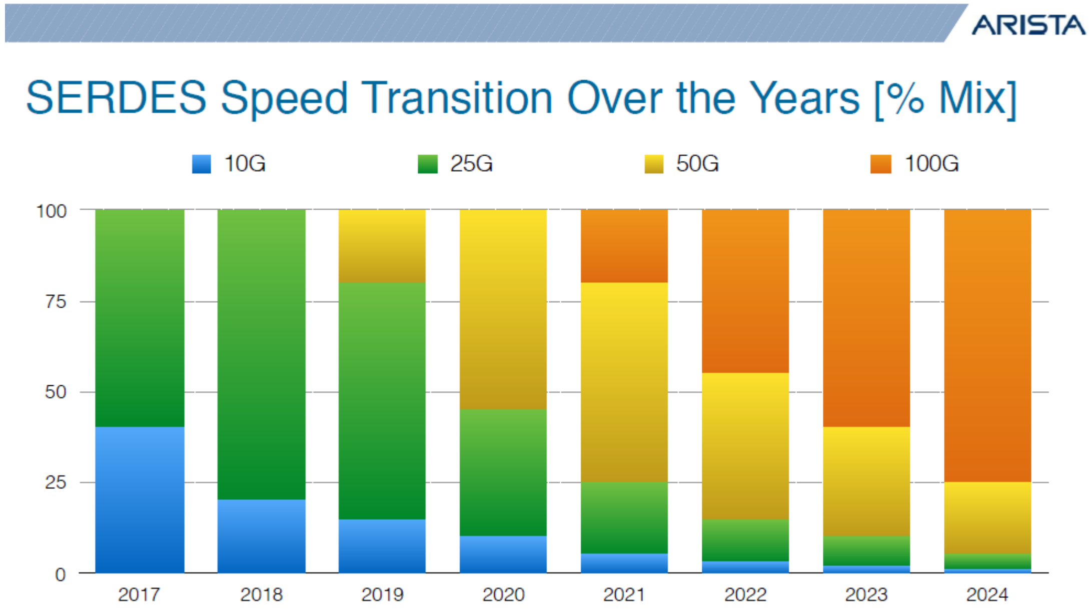
January 2020 Geneva Interim

Summary

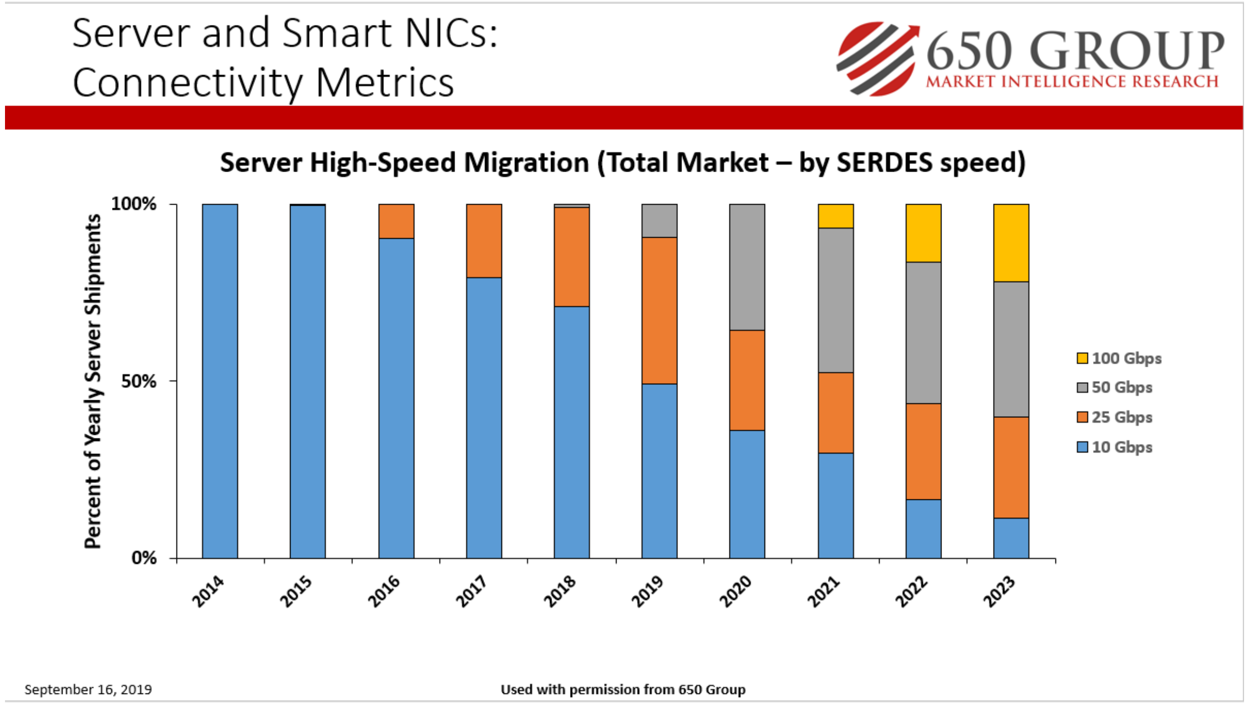
- A need is emerging for short-reach optics based on 100Gb/s wavelengths, as 100G SerDes enter the market over the next couple years
  - Reach over passive copper cable may be less than 2m for 100Gb/s lanes
  - Historically, VCSELs & MMF have advantages for lower cost interconnects, but Ethernet currently does not yet specify MMF PMDs using 100Gb/s wavelengths
  - Cost and complexity will be reduced by matching the optical speed of VCSEL-MMF links to emerging electrical speeds.
- Several trends combine to favor longer reach for server-attachment, up to as long as perhaps 30m
  - Other applications, e.g. interconnection of machine-learning clusters, may also benefit from lower cost 100G per wavelength optical links
- A meaningful fraction of switch-to-switch links, especially in the China data center market, can be covered by links up to 50m
- Technically feasible objectives for 100GbE & 400GbE MMF PMDs based on 100Gb/s wavelengths do meet the criteria for broad market potential, economic feasibility, and distinct identity
  - The SG should assess whether a similar 200 GbE objective has broad market potential

100 Gb/s fundamental SerDes rates will be available for switch ASICs and server NICs in near future

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



"Scaling the Cloud Network," Andreas Bechtolsheim, OCP Summit 2018


Server and Smart NICs:
Connectivity Metrics

September 16, 2019

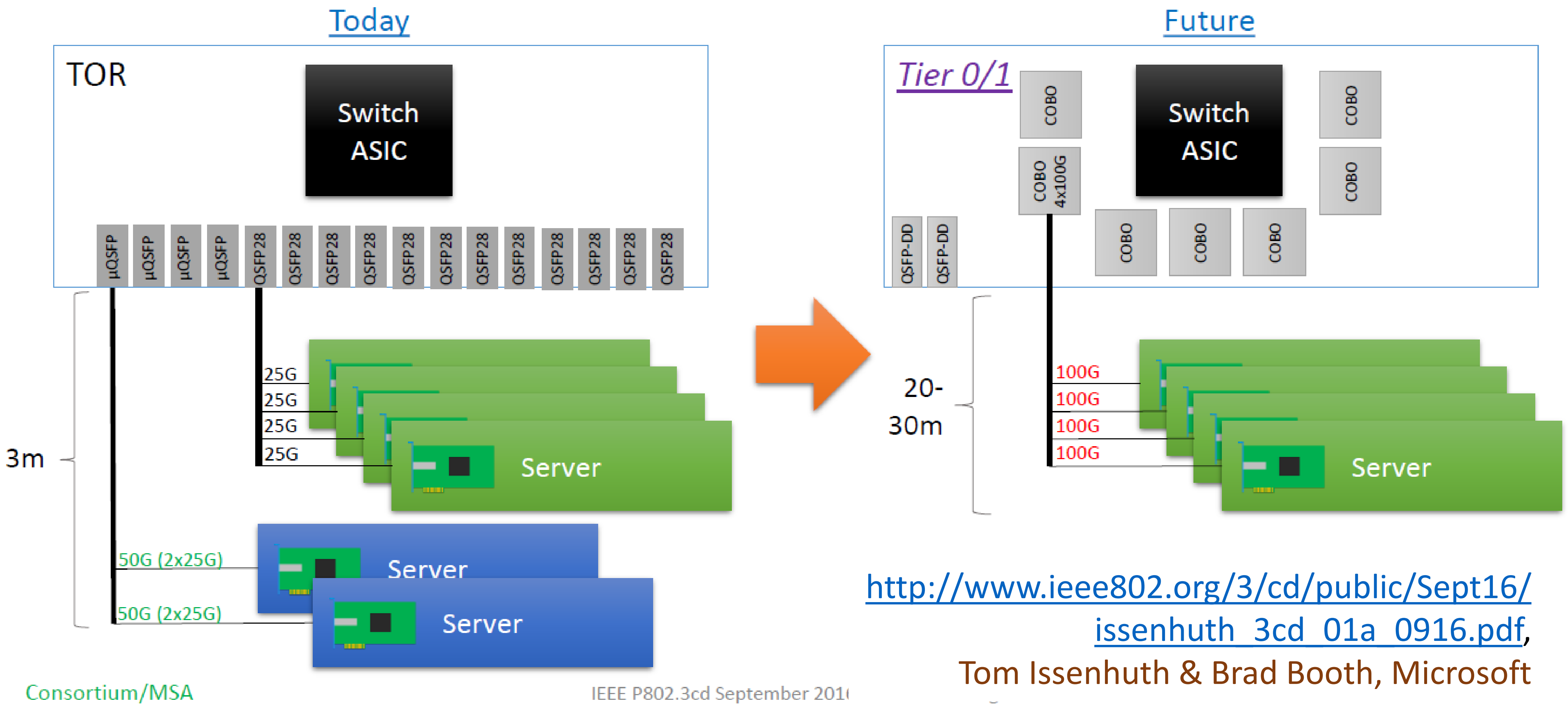
Used with permission from 650 Group

Used with permission from 650 Group

Optimized architectures evolve with server & switch technology

- Servers are moving from 25 to 50 to 100 G SerDes-based 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
- SerDes pair counts around switch ASICs have increased from 128 to 256 to 512, while fundamental speeds are also increasing.
- 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

Emerging MoR/EoR architectures will require compact optical cable and ease of breakout over 20-30m, predicted in 2016 (link below)



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

Market need for 100G per wavelength, VCSEL-MMF interconnects from a North American perspective

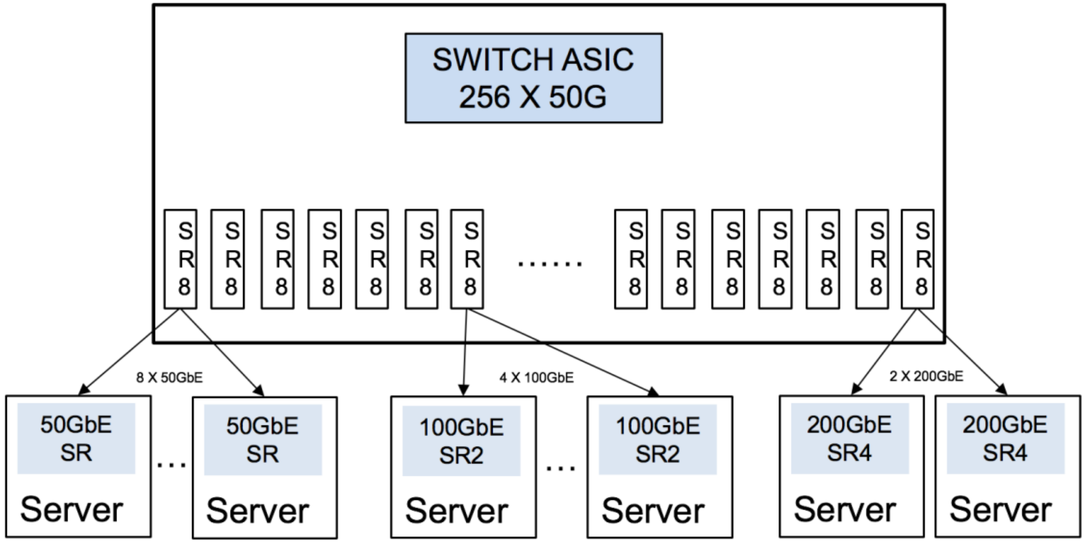
- Market need
  - Low-cost interconnect between switches with 32×800G-capacity ports (expected in 2020).
    - Passive copper cable limited to (1-2?) m. Active copper cable limited to ~ 5m.
    - A possible 16x50G/λ PMD would have twice the lane count & an unusual higher fiber-count connector.
    - 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
    - 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 4×100GBASE-SR
  - Dual 400GBASE-SR4 in octal module
    - Lowest-cost, low-fiber count point-to-point connection for 2×400G QSFP-DD800 or OSFP112 ports
    - Breakout to 8×100GBASE-SR

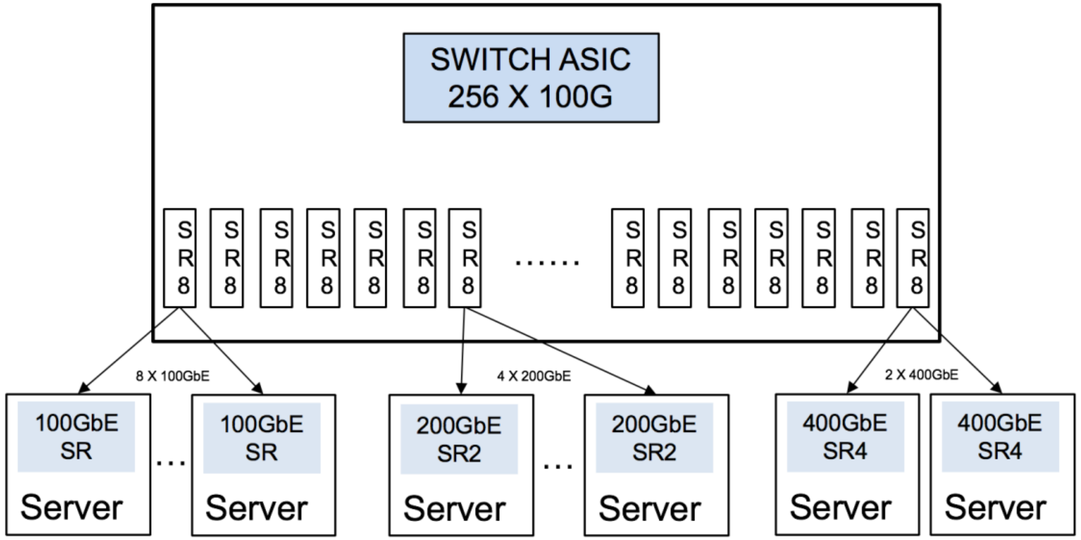
One perspective on potential applications for 100G per wavelength, short-reach, VCSEL-MMF links in big cloud datacenters in China

- Applications:
  - AOC used today for server-to-ToR connections
  - Transceivers used for ToR-to-leaf switch connections
- Distances targets:
  - 100m reach desired
  - 50m reach required to use transceivers
  - 50m reach covers 80% of ToR-LEAF switch links at Alibaba
  - ≤ 30m is currently a space for AOCs at Alibaba
  - Server connections will be longer than 2-3m in the future
- Configurations:
  - Need for breakout depends on network architecture
  - Use of breakout in the future could favor transceivers over AOC in some cases

8 fp cabling with octal transceiver modules form a very flexible paradigm for breakout applications to servers

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



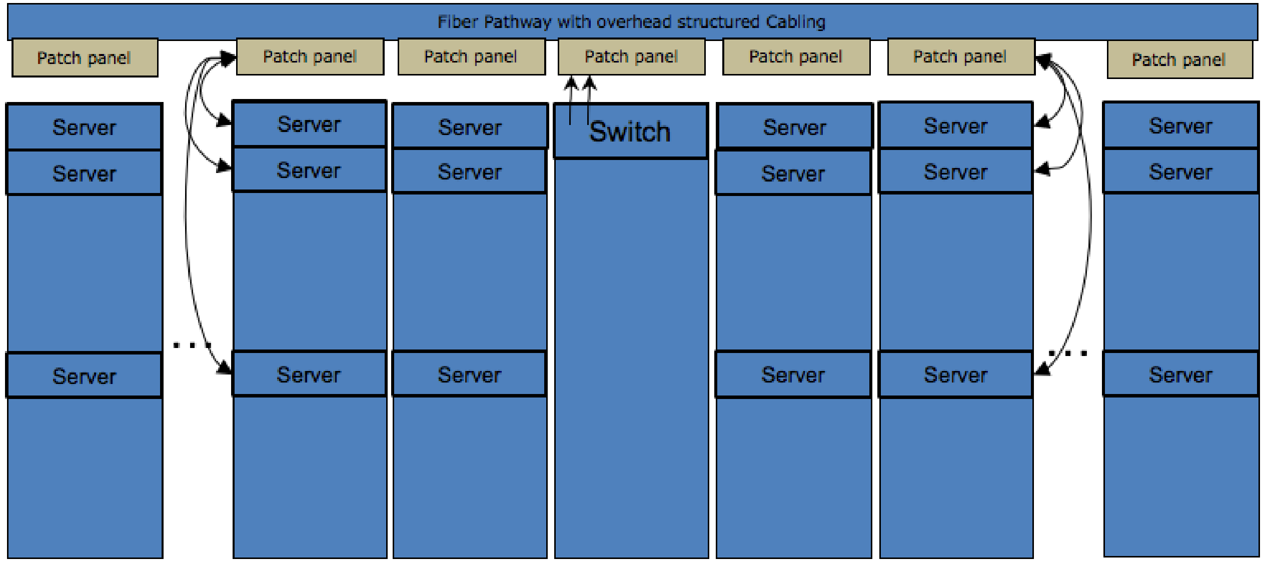


http://grouper.ieee.org/groups/802/3/NGMMF/public/Jan18/shen_NGMMF_01_jan18.pdf

One example of the use of 100G short reach optical links with transceivers to facilitate TOR elimination

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)



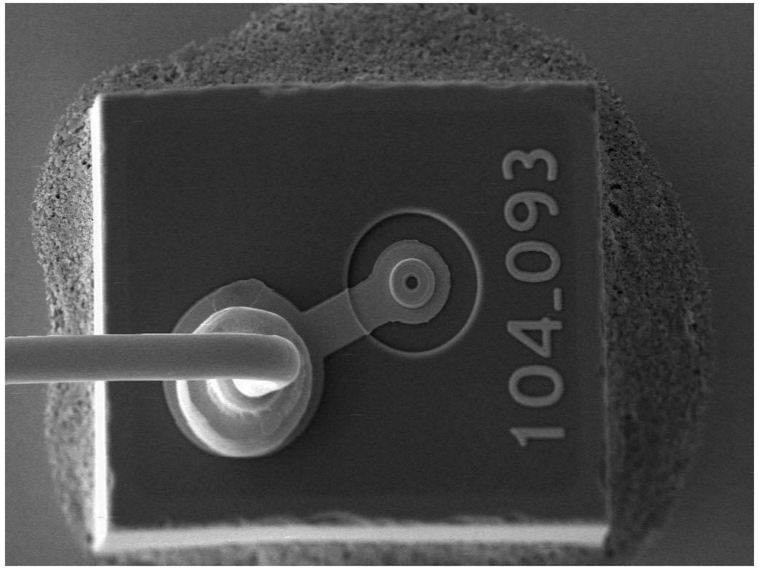
- Typical server row 16 – 20 cabinets
- Cabinets arrive on site with servers installed
- Overhead cable is pre-installed with pathway
- Simple patching from server to overhead patch panel

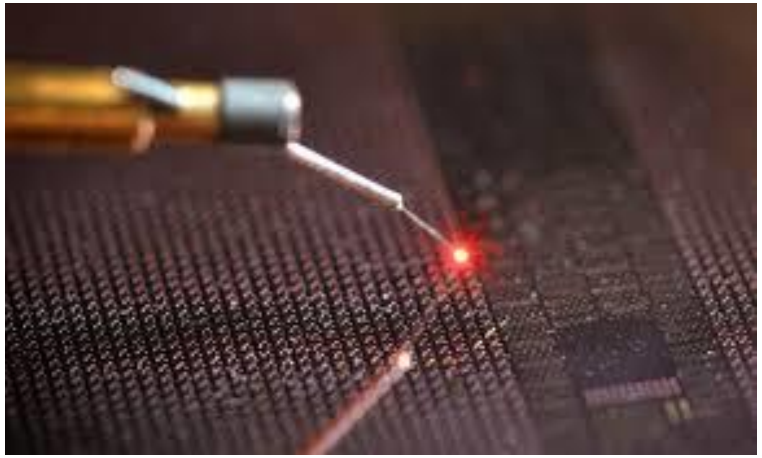
Broad market potential for 400GBASE-SR4 and lower speed breakout optics is supported by a reasonable estimate

- Roughly 15-20% of fiber interconnect needs in large data centers are less than 30m, whether transceivers or AOCs
- With a transition from ToR to EoR/MoR switches, where the server ports may be tens of meters away, there is another large potential application for 100G per wavelength fiber links
  - Assume that half the server connections convert from TOR to MoR/EoR switching architecture
  - Then an additional 15-20% of the market is available for these optics
  - 400G-SR4 in quad or octal modules could break out to 4x100G, 8x100G, 2x200G, or 4x200G server attachments, respectively
- This represents a potential 30-40% of the projected 400G volume, including both transceivers and AOCs, that will benefit from short-reach 100G per wavelength fiber links

Historically VCSEL-MMF links have advantages for lower cost 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





Cost & density benefits accrue from higher lane speeds

- A suite of MMF PMDs have been defined by IEEE 802.3 using 50G wavelengths:
  - 400GBASE-SR8 (8x50G) over 8 fp
  - 400GBASE-SR4.2 (8x50G) over 4 fp
  - 200GBASE-SR4 (4x50G) over 4 fp
  - 100GBASE-SR2 (2x50G) over 2 fp
  - 50GBASE-SR (1x50G) over 1 fp

- Higher speed 100Gb/s wavelengths lead to reduced lane counts, reduced fiber & component counts, reduced complexity, and lower cost than previously standardized PMDs, enabling potential PMDs such as:
  - 400GBASE-SR4 (4x100G) over 4 fp
  - 200GBASE-SR2 (2x100G) over 2 fp
  - 100GBASE-SR (1x100G) over 1 fp

Conclusion & Recommendation

- We have demonstrated Broad Market Potential, Economic Feasibility, and Distinct Identity for 100 Gb/s per wavelength VCSEL-MMF PMDs for applications in the cloud
- The following objectives should be adopted, assuming technical feasibility is demonstrated by other contributions:
  - Define a physical layer specification that supports 100 Gb/s operation over 1 pair of MMF with lengths up to at least 50m
  - Define a physical layer specification that supports 400 Gb/s operation over 4 pairs of MMF with lengths up to at least 50m
- The SG should investigate whether a 200GbE objective has BMP
  - For example, 200G QSFP56 ports on mezzanine NIC card exist today
  - IEEE P802.3ck has an objective for 200GBASE-CR2

Supporters

- Yan Zhuang – Huawei
- Vipul Bhatt – II-VI
- Mark Kimber – Semtech
- Leon Bruckman – Huawei
- Jose Castro – Panduit
- Flavio Marques – Furukawa LATAM
- John Abbott – Corning
- Rich Baca – Microsoft
- Ramana Murty - Broadcom
- Rick Pimpinella – Panduit
- Masaru Terada – Furukawa Electric
- Earl Parsons – CommScope
- Kobi Hasharoni – Dust Photonics
- Rob Stone – Broadcom
- Ken Jackson – Sumitomo
- Ed Sayre – North East Systems Associates
- Dean Wallace – Marvell
- Rich Mellitz – Samtec
- David Chen – AOI
- Piers Dawe – Mellanox