Next-gen 400 and 200 Gb/s PHYs over Fewer MMF Pairs Call For Interest Consensus Presentation

IEEE 802.3

Draft 0.2

Agenda

- **Overview Discussion**
  - Presenter 1
- **Presentations**
  - Market Drivers
    - Presenter 2
  - Technical Feasibility
    - Presenter 3
  - Why Now?
    - Presenter 4
- **Straw Polls**

Introductions for today's presentation

- Presenter and Expert Panel:

CFI Objectives

- To gauge the interest in next-gen 400 and 200Gb/s PHYs over fewer MMF pairs .

- 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

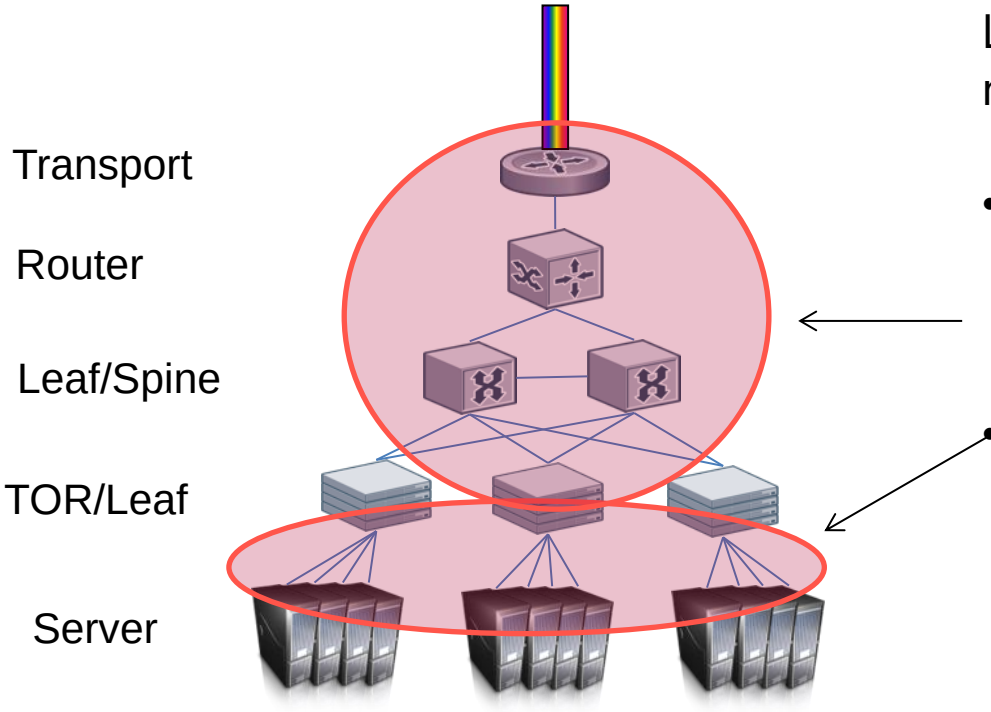
Overview: Motivation

Leverage technologies currently under development to create cost-optimized lower fiber count solutions over installed base, as well as greenfield MMF cabling, for 200 and 400 Gb/s

Global web-scale data centers and cloud based services – as well as the largest enterprise datacenters - are presented as leading applications.

Synergy with broader enterprise networking extends the application space and potential market adoption.

What Are We Talking About?



Leading application space for next generation MMF PMDs

- Switch-to-switch & switch-to-router or router-to-transport connectivity
- Breakout of 400G to 100G may be used for high density 100G or breakout to 100G servers

Market Drivers

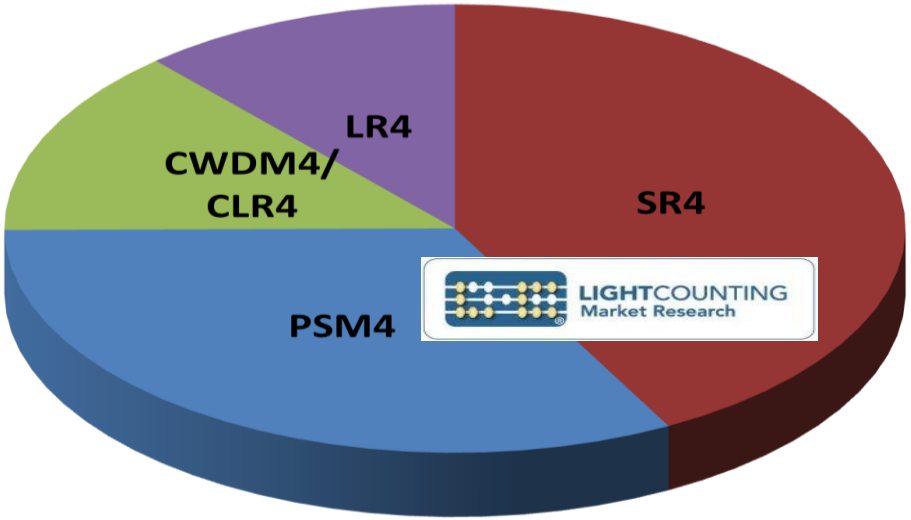

Historically VCSEL-MMF links have been seen by many as the lowest cost short-reach interconnect

- relaxed alignment tolerances
- low drive currents
- on-wafer testing
- connectors more resilient to dirt
- NEED HELP FROM EXPERTS TO CREATE A GOOD SLIDE
  Including a graphic

10, 40 and 100 Gb/s have been deployed over an installed base of MMF

- Large installed base of duplex OM3/OM4 MMF deployed for 10GBASE-SR
- Large installed base of parallel OM3/OM4 MMF deployed for 40GBASE-SR4 and 100GBASE-SR4
- Industry investment in MMF cabling continues, including OM5

100GbE QSFP28 Consumption in 2016



- Taken together, SMF modules together have majority share
- But short-reach SR4 modules had the greatest individual contribution to 2016 shipments of QSFP28 modules

Slide courtesy of Dale Murray, LightCounting

How have 40 and 100Gb/s optics been used with MMF?

- 40G SR4
  - ~ 50% is breakout, to servers as well as a means to creating larger 10G switch fabrics
  - ~ 50% is switch-to-switch links
- 40G BiDi and SWDM4
  - Proprietary solutions used in switch-to-switch connections
- 100G SR4
  - The 100G SR4 modules deployed in 2016 represent switch-to-switch and switch-to-router connections in Cloud and Largest enterprise DCs
  - Less likely to be used for breakout in cloud or largest enterprise DCs (the early adopters), where breakout to servers often done with DAC cables from TOR, and 25G is not expected to be a popular switch fabric speed there
  - 100G Breakout may be popular in smaller enterprise DCs and campus networks (later adopters), where “25G may be the new 10G”
- 100G duplex over MMF
  - Proprietary solutions; not yet in market; two sources expected in 2017; could have been sold in 2016 if available

Market applications of 400G short reach

- Earliest use for low-cost router-transport and laboratory development applications in telecom and the cloud
- Initial volume applications in switch-router & switch-switch connections
  - in the cloud
  - largest enterprise DCs
- Breakout of 400G to 100G may be popular as for 40G-SR4, since “100G is the new 10G” in this space

Alibaba supports standardization of 400GBASE-SR4 in IEEE802.3

- Alibaba expects to deploy 100G switching for approx. three years, perhaps moving to 400G in 2019
- Alibaba uses 100GBASE-SR4 heavily for 100m switch-switch connections now
- 300m reach supports about 80% of Alibaba's data center links, and eSR4 extended reach MMF optics will be deployed when available
- Alibaba deploys CWDM4 over SMF between buildings
- 100GBASE-SR4 links over MMF cabling are lower cost for Alibaba today than PSM4 or CWDM4 links over SMF cabling
- Alibaba hopes to have 400GBASE-SR4.n in the future and supports its standardization in IEEE

Does existing 400GBASE-SR16 fulfill the needs of the datacenter market?

- 400GBASE-SR16 was envisioned as a lower-cost, fast time-to-market solution for router-transport & development needs
- 400GBASE-SR16 may not be a high-volume datacenter module
  - CFP8 will not be a common front panel port in datacenter switches
  - 32-fiber link with atypical connector will offset the low-cost nature of the transceiver.
  - Restricted to 16x25G interface (400GAUI-16)
    - No path to 400GAUI-8 without reverse gearbox
- 400GBASE-SR4 expected to be lower cost than FR8 or DR4

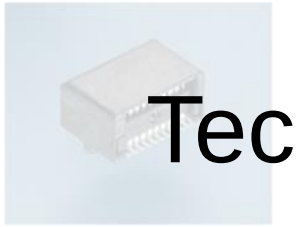
Benefits of 400GBASE-SR4 over 400GBASE-SR16 for the datacenter market

- Operate on same cabling as previous SR4 modules
  - No special connector
- Suitable for all 400G form factors
  - CFP8, QSFP-DD, OSFP
- No reverse gearbox with 400GAUI-8 interface

Market Need for 200G module for Duplex MMF

- 200G switching is expected to find acceptance in parts of the cloud and enterprise DC networking space on same time frame as 400G
- 200GBASE-SR4 is already being standardized in 802.3cd to meet expected demand over installed base of parallel MMF cabling
- Early demand for 100G duplex MMF optics is expected to be replicated for 200G optics over the installed base of duplex MMF cabling

Technology Feasibility



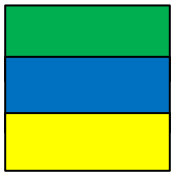
Technologies for Next-Gen MMF PMDs

- PMDs for parallel 400G and duplex 200G over MMF will require several technologies currently in advanced stages of development
  - VCSELs supporting 50Gb/s PAM4 signaling
  - Multiple wavelengths over MMF
- OM5 provides longer reach when using multiple wavelengths over MMF, but is not required

Technical options for Next-Gen MMF PMDs

Technology (per fiber)	1 fiber pair	2 fiber pairs	4 fiber pairs	8 fiber pairs	16 fiber pairs
25G-λ NRZ	25G-SR		100G-SR4		400G-SR16
50G-λ PAM4	50G-SR	100G-SR2	200G-SR4	400G-SR8	
2x50G-λ PAM4	100G-SR1.2	200G-SR2.2	400G-SR4.2		
4x25G-λ NRZ	100G-SR1.4	200G-SR2.4	400G-SR4.4		
4x50G-λ PAM4	200G-SR1.4	400G-SR2.4	800G-SR4.4		

Technology options for 100-800 Gb/s links over fewer MMF fiber pairs



Existing IEEE standard

In progress in 802.3bs

In progress in 802.3cd

Multi-Wavelength Nomenclature

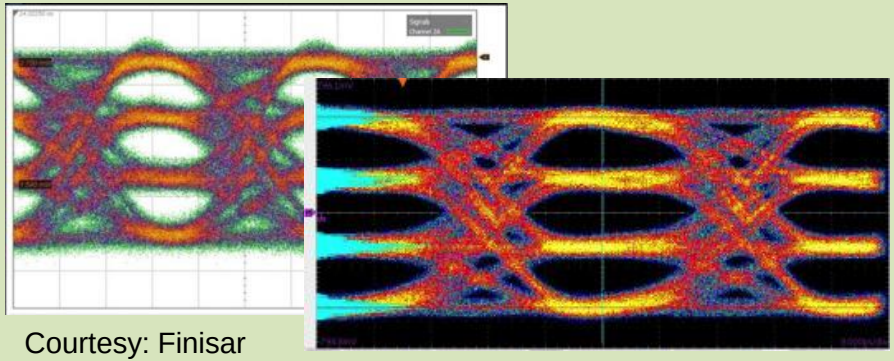
SRm.n m = # fiber pairs

n = # wavelengths

- Placeholder for slide on 50G PAM4 feasibility with VCSELs
- Include comments on coding gain assumptions
- A slide from 802.3cd CFI is reproduced on next page for reference.
- Several have been asked to supply updates; Material welcome from any supplier

Optical Technical Feasibility

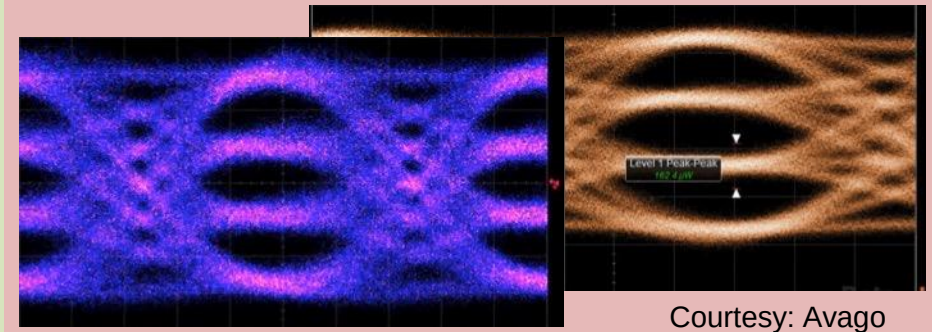
1300nm 56G PAM4



Courtesy: Finisar

Courtesy: Cisco

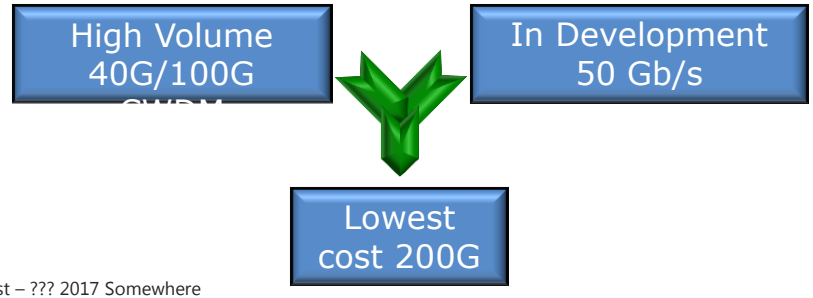
850nm 50G PAM4



Courtesy: Finisar

Courtesy: Avago

Today's high volume 40GbE and 100GbE SMF/MMF technology is directly extendable to 50GbE and 200GbE SMF/MMF applications



- Statement on status & applications of OM5
- John Kamino & Paul Kolesar requested to provide a starting point
- Material from others welcomed as well – please send me for review.

- Placeholder for 2-wavelength over MMF feasibility
- Jonathan Ingham committed to provide week of 2/27

- Placeholder for 4-wavelength over MMF feasibility
- Jonathan King will supply material

Why now?

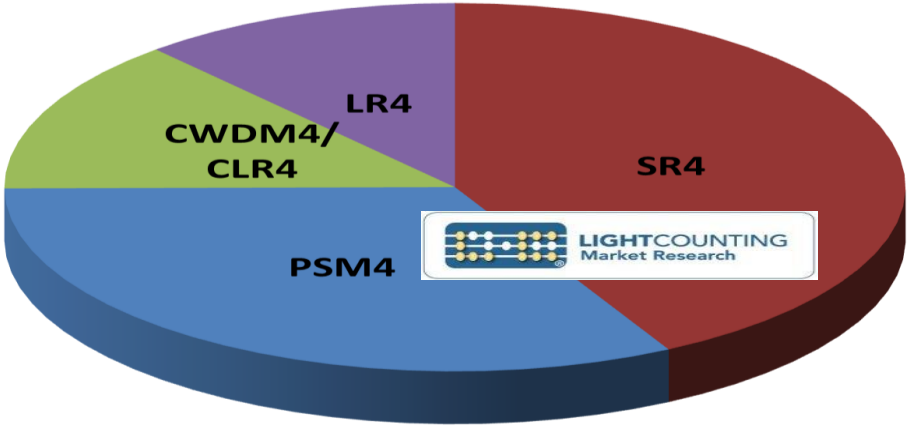
40 inches
35 dB at 12.9 GHz
Clean eye
Closed eye
Backplane

XILINX

Rationale for beginning now

- Recent history shows that higher speeds over MMF are needed in the first year that new switch speeds are commercially available
- The existing 400GBASE-SR16 solution will not meet that need
- There is no 200Gb/s duplex MMF PMD in existing IEEE standards
- These PMDs are needed in the market commercially in 2019

100GbE QSFP28 Consumption in 2016



- Taken together, SMF modules together have majority share
- But short-reach SR4 modules had the greatest individual contribution to 2016 shipments of QSFP28 modules

Slide courtesy of Dale Murray, LightCounting

Request in to a Switch Vendor to Attest to Following

- Sold 100GBASE-SR4 into the cloud DC space in 2016
- Sold 100GBASE-SR4 into large enterprise DC space in 2016
- Could have sold 100G duplex MMF transceivers in 2016 had they been available

Why Now? – rewrite this as appropriate

50 Gb/s SERDES investment and development underway

- Ethernet rates becoming defined by the optimal implementation of these SERDES rates
- Necessary to enable data center architectures (radix) and practical implementations (chip packaging)

Web-scale data centers and cloud based services need Highly Cost optimized servers with >25GbE capability

Industry has recognized the value of leveraging common technology developments across multiple applications by implementing in multiple configurations of lanes.

- Rapid standardization avoids interoperability challenges

Ethernet is immediately able to leverage technology for broader adoption and enable greater economy of scale

- There is no 50 Gb/s Ethernet single lane standardization effort under way
- There is no 200 Gb/s Ethernet standardization effort under way

Continuing Ethernet's success

- Open and common specifications; Ensured Interoperability; Security of development investment

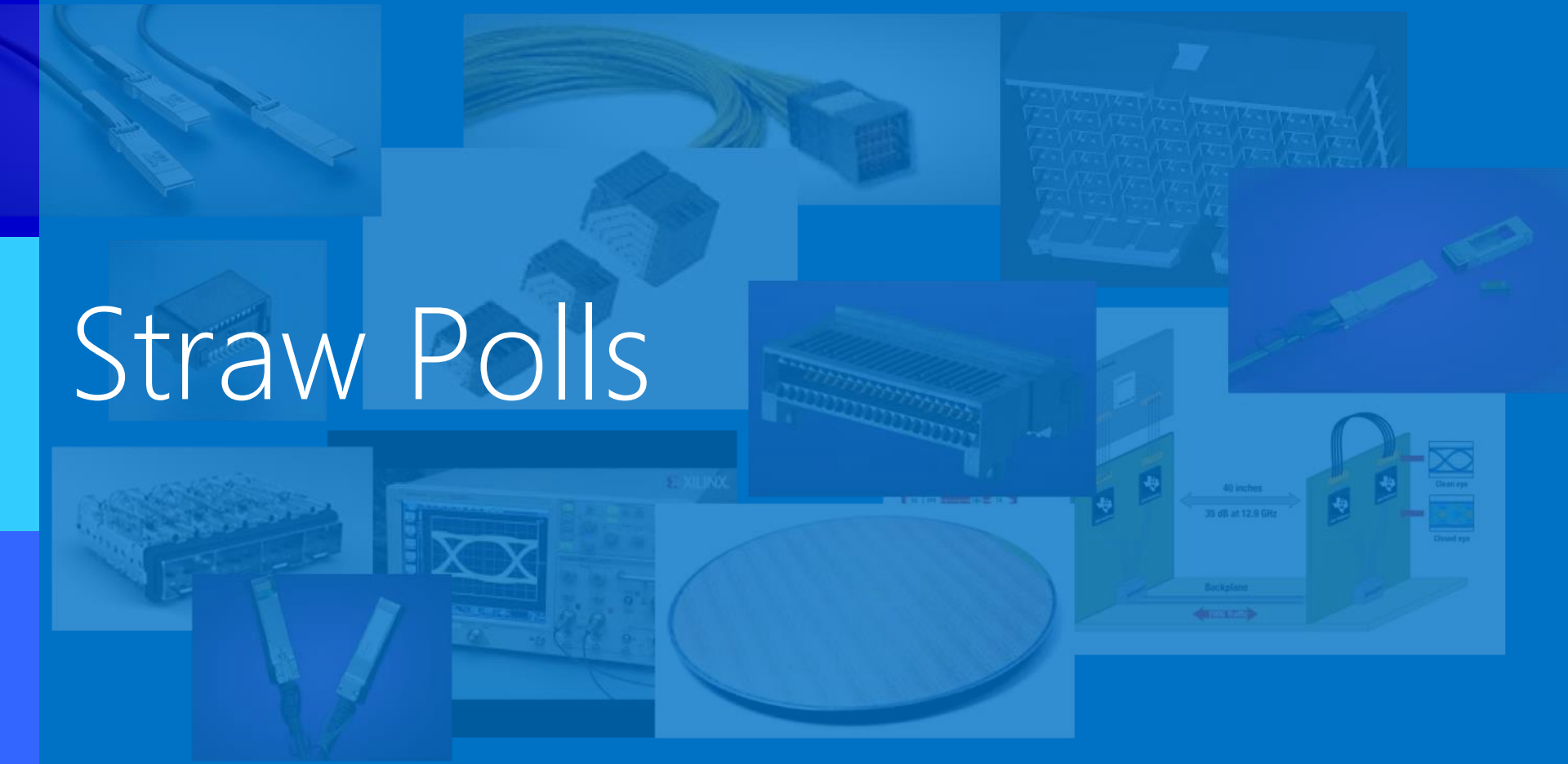
Contributor Page

Dale Murray, LightCounting
Chongjin Xie, Alibaba
Steve Swanson, Corning

Supporters (p Individuals from q companies)

Supporters (2)

Straw Polls



Existing and in-process Ethernet standards have prepared the ecosystem for new MMF PMDs

- 25 Gb/s electrical lanes
- 50 Gb/s electrical lanes
- 100 Gb/s MAC, PHYs, and PMDs
- 200 Gb/s MAC, PHYs, and PMDs
- 400 Gb/s MAC, PHYs, and PMDs

{Make this into a table with relevant information, 802.3 references}

Call-for-Interest Consensus

- Should a study group be formed for “Next-gen MMF PMDs”?
- Y: N: 0 A:
- Room count:

Participation

- I would participate in a “Next-gen MMF PMDs” study group in IEEE 802.3
  - Tally:
- My company would support participation in a “Next-gen MMF PMDs” study group
  - Tally:

Future Work

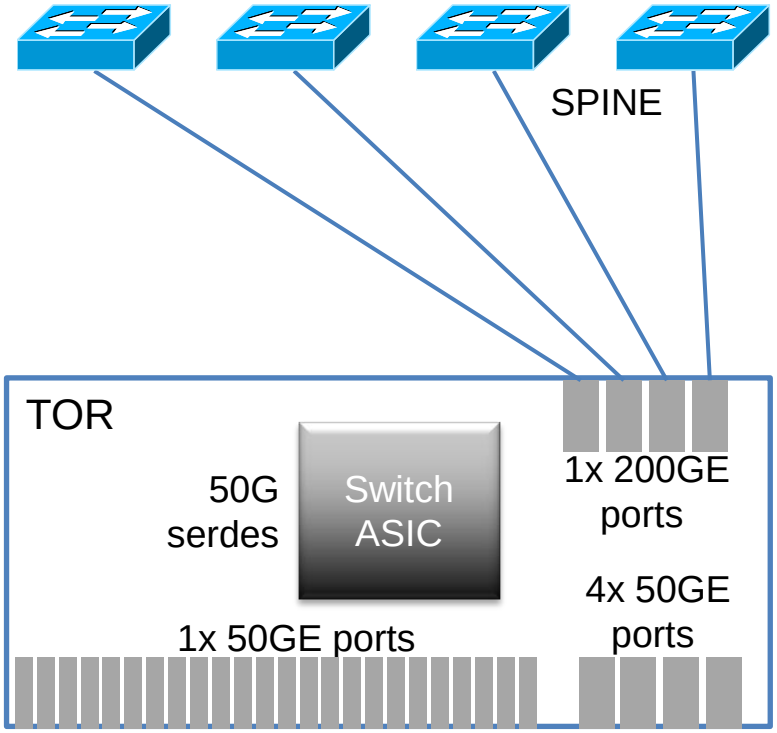
- Ask 802.3 at Thursday's closing meeting to form study groups
- If approved:
  - Request 802 EC to approve creation of the study groups on Friday
  - First joint study group meeting would be during Jan 2016 IEEE 802.3 interim meeting

Slides from November CFI as reference

I doubt many are useful?

200 Gb/s Ethernet Connectivity

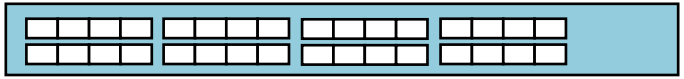
- Enables DC fabric topology similar to 40 Gb/s and 100 Gb/s Ethernet
  - Switch to switch fabric interconnect
  - Single-mode and multi-mode fiber or AOC
  - Switch-to-Switch typical reaches 100m (MMF) to ~2km (SMF)



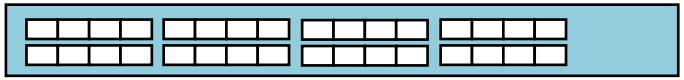
Network Demand for 200GbE

- As servers virtualize more applications, they drive more bandwidth into the network
- The network uplinks need to progress to higher speeds to match the server speeds
- 200GbE can provide a similar network infrastructure and oversubscription as servers migrate from 25GbE to 50GbE.

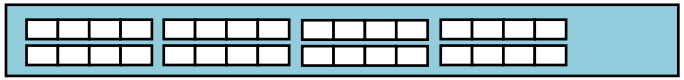
32 QSFP+ port Switch in 2012
4x10GbE Down, 40GbE up



32 QSFP28 port Switch in 2016
4x25GbE Down, 100GbE up



32 QSFP56 port Switch in 2020
4x50GbE Down, 200GbE up



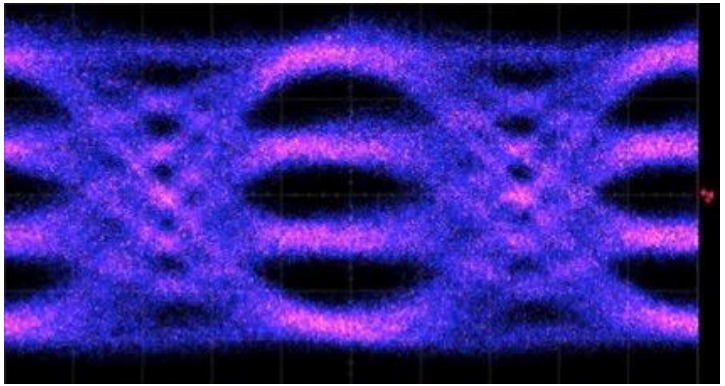
Courtesy: Commscope

E.g. 1:12f trunk cables
1x40GE→1x100GE→1x200GE
using SR4 technology

Leverage of Industry investment

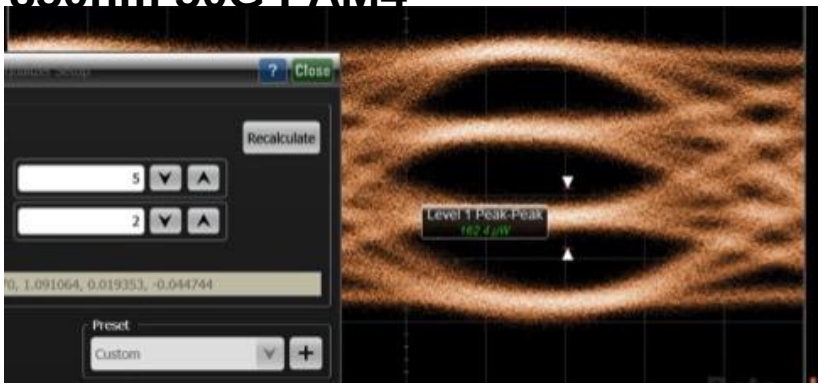
Technology	Nomenclature	Description	Status
Backplanes	100GBASE-KP4 & KR4 CEI-56G-LR-PAM4	4 x 25 Gb/s backplane 56 Gb/s PAM4	IEEE 802.3bj Published Straw Ballot
Chip-to-Module	CDAUI-8 CEI-56G-VSR-PAM4	8 x 50 Gb/s PAM4 60 Gb/s PAM4	IEEE P802.3bs in Task Force Rev Straw Ballot
Chip-to-Chip	CDAUI-8 CEI-56G-MR-PAM4	8 x 50 Gb/s PAM4 60 Gb/s PAM4	IEEE P802.3bs in Task Force Rev Straw Ballot
SMF Optical	400GBASE-FR8 & LR8 400GBASE-DR4	8 x 50 Gb/s PAM4 4 x 100 Gb/s PAM4	IEEE P802.3bs in Task Force Review
Module Form Factor	SFP56	1 x 50 Gb/s	Extension to Summary Document SFF-8402
	QSFP56	4 x 50 Gb/s	Extension to Summary Document SFF-8665

Optical Technical Feasibility



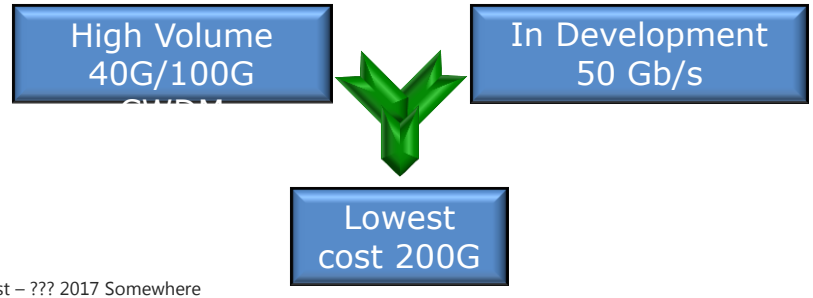
Courtesy: Finisar

850nm 50G PAM4



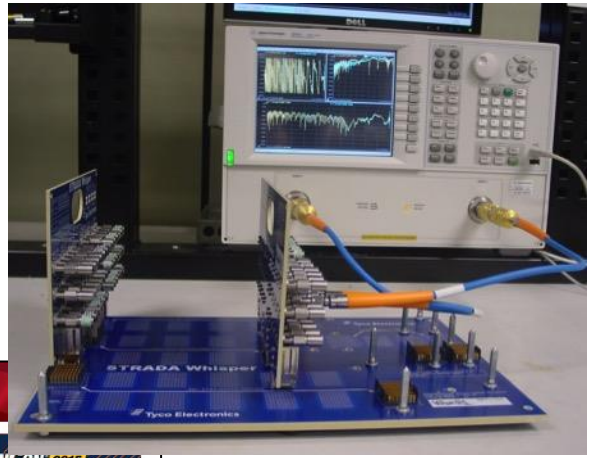
Courtesy: Avago

Today's high volume 40GbE and 100GbE SMF/MMF technology is directly extendable to 50GbE and 200GbE SMF/MMF applications



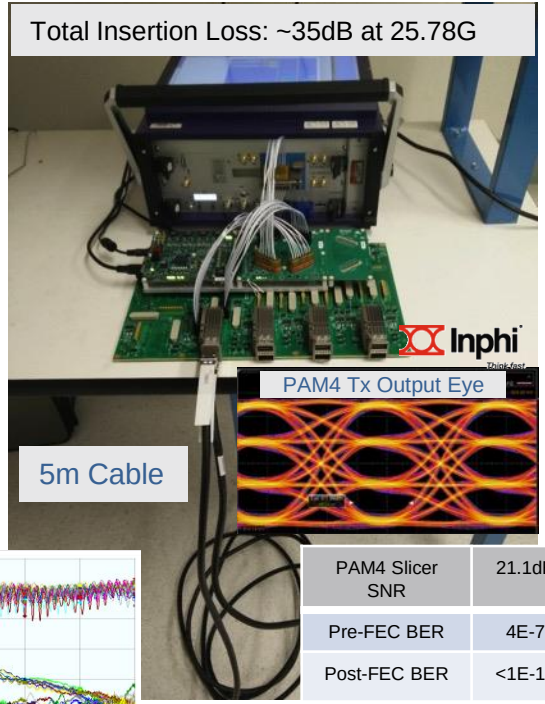
Connector/Cable/Backplane Technical Feasibility

Numerous industry demos showing 50G connector, cable and backplane capabilities



50G PAM4

Total Insertion Loss: ~35dB at 25.78G



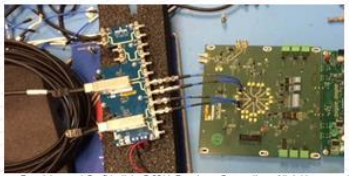
PAM4 Slicer SNR	21.1dB
Pre-FEC BER	4E-7
Post-FEC BER	<1E-15

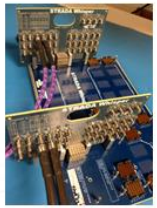
Courtesy: Inphi

BROADCOM DESIGNCON 2015 DEMOS

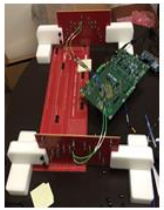
- **40G PAM4**
  - 40G PAM4 demo: error-free operation on 10m passive QSFP cable
  - No external intervention – equalization, adaptation, FEC, are all on-chip
  - ~30dB insertion loss, crosstalk
- **50G PAM4**
  - 50G PAM4 demo: error-free operation on 40in Molex Impel Meg6 Backplane designed for 100GBASE-KR4
  - No external intervention – equalization, adaptation, FEC, are all on-chip
  - ~31dB insertion loss (not including test cables), crosstalk

40G PAM4 TE connectivity





50G PAM4 molex

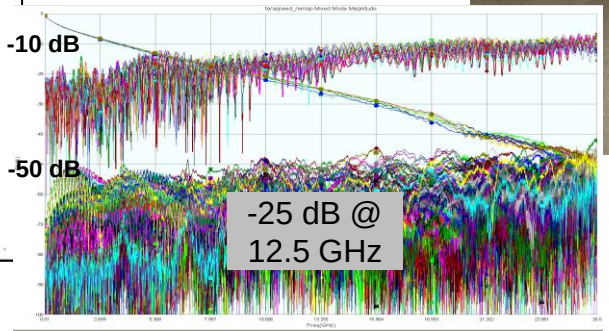


http://www.ieee802.org/3/bs/public/adhoc/elect/15_0212/parthasarathy_01_0215_elect.pdf

1m BP @ DesignCon 2015

- Demonstrated with both 50G PAM4 & NRZ

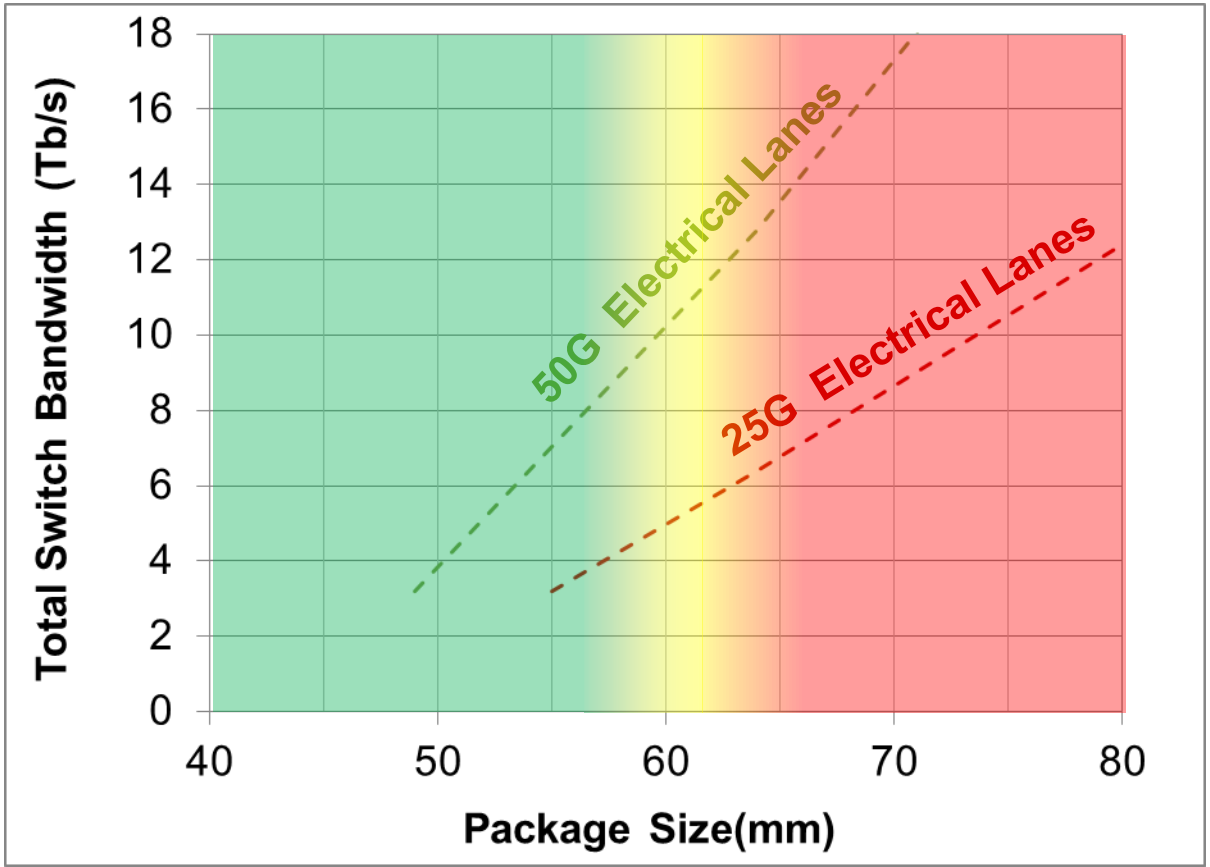
Courtesy: TE



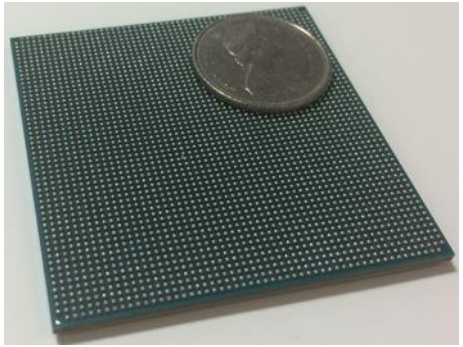
29” Backplane design

Courtesy: Teraspeed Consulting - A Division of Samtec

Switch Chip Package Limitations
(Avoiding Technical Infeasibility)



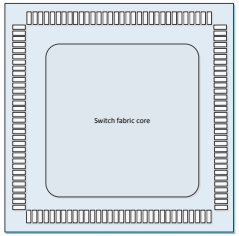
BGA package limitations require eventual migration to higher electrical lane speeds to support higher bandwidth switches



Current technology: ~ 65mm supports up to ~ 6 Tb/s of at 25 Gb/s per lane

50 Gb/s I/O Efficiency

- Switch ASIC Connectivity limited by serdes I/O
- 50 Gb/s lane maximizes bandwidth/pin and switch fabric capability vs. older generation
- Single Lane port maximizes server connectivity available in single ASIC
- 50 Gb/s port optimizes both port count and total bandwidth for server interconnect



For a 128 lane switch:

Port Speed (Gb/s)	Lane Speed (Gb/s)	Lanes /port	Usable ports	Total BW (Gb/s)
10	10	1	128	1280
25	25	1	128	3200
50	25	2	64	3200
50	50	1	128	6400
100	50	2	64	6400
200	50	4	32	6400

Using 50 Gb/s ports maximizes connectivity and bandwidth.

Switch Radix Capabilities

	Total Number of Ports Possible		
Total ASIC IO (Tb/s)	**100G Ports @ 25G IO**	**200G Ports @ 50G IO**	**400G Ports @ 50G IO**
3.2	32	16	8
4.8	48	24	12
6.4	64	32	16
9.6	-	48	24
12.8	-	64	32

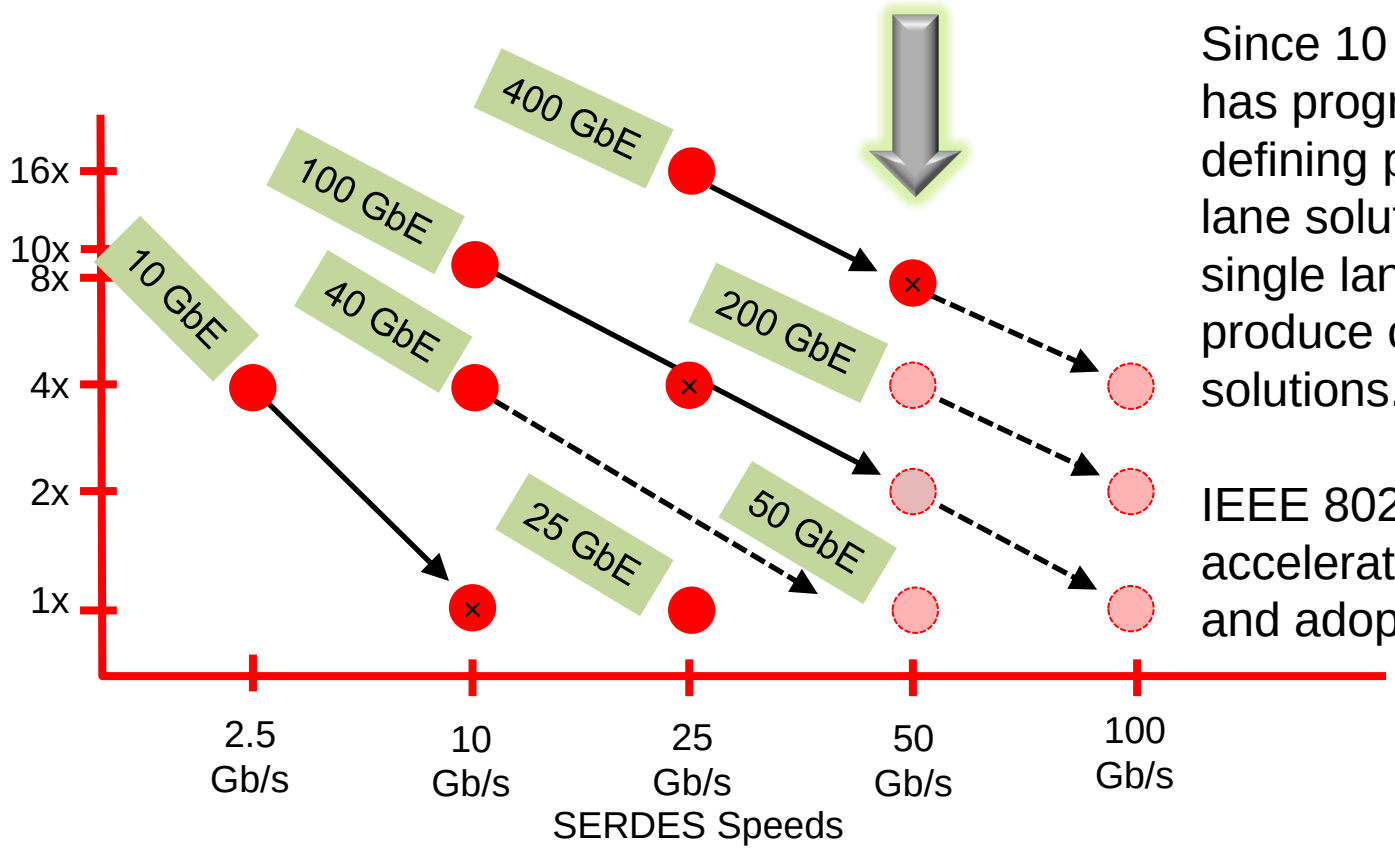
Red: < current # ports

Green: ≥ current # ports

- For Leaf / Spine Switch Applications, networks need sufficient number of ports to minimize number of switch stages in the fabric
- Current topologies utilize 32 ports @ 100G per Leaf/ToR (and per switch silicon)
- Historically the number of ports have been maintained while total ASIC BW has increased

- **Using 200G interfaces for leaf / spine applications enables a sufficient number of ports at a lower ASIC BW (i.e. earlier in time) than 400G interfaces**

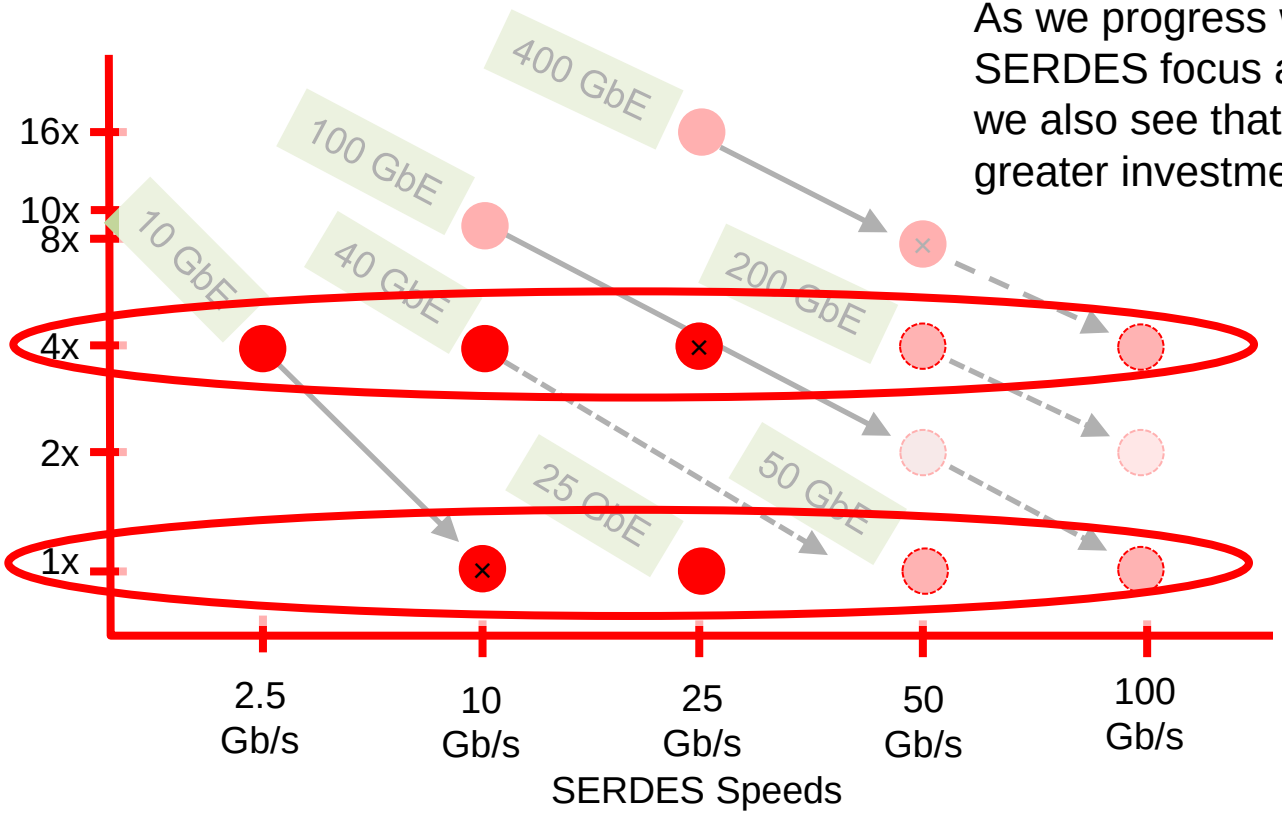
The new normal – multi-lane and re-use



Since 10 GbE, Ethernet has progressed by defining pragmatic multi-lane solutions and fastest single lane technologies to produce cost-effective solutions.

IEEE 802 definition accelerates market focus and adoption.

Single lane and multiple lanes



As we progress with the trends around SERDES focus and multi-lane variants we also see that certain multiples have greater investment and focus.

4x consistent with existing implementation experience resulting in cost effective multi-lane solutions

1x (single-lane) always represents lowest cost solution (once technical and manufacturability issues are addressed).