50 Gb/s Ethernet Over a Single Lane and Next Generation 100 Gb/s & 200 Gb/s Ethernet Call For Interest Consensus Presentation

IEEE 802.3

Mark Nowell, Cisco

Dallas, Tx

Nov 10th, 2015

Introductions for today’s presentation

- Presenter and Expert Panel:

  Mark Nowell – Cisco

  John D’Ambrosia – Independent

  Adam Healey – Avago

  Rob Stone – Broadcom

  Chris Cole - Finisar

CFI Objectives

- To gauge the interest in 50 Gb/s Ethernet Over a Single Lane and Next Generation 100Gb/s & 200Gb/s Ethernet.

- 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 50 Gb/s electrical IO signaling technology to develop cost optimized single-lane solutions and higher speed multi-lane solutions.

Provide 50 Gb/s MAC rate and PHYs

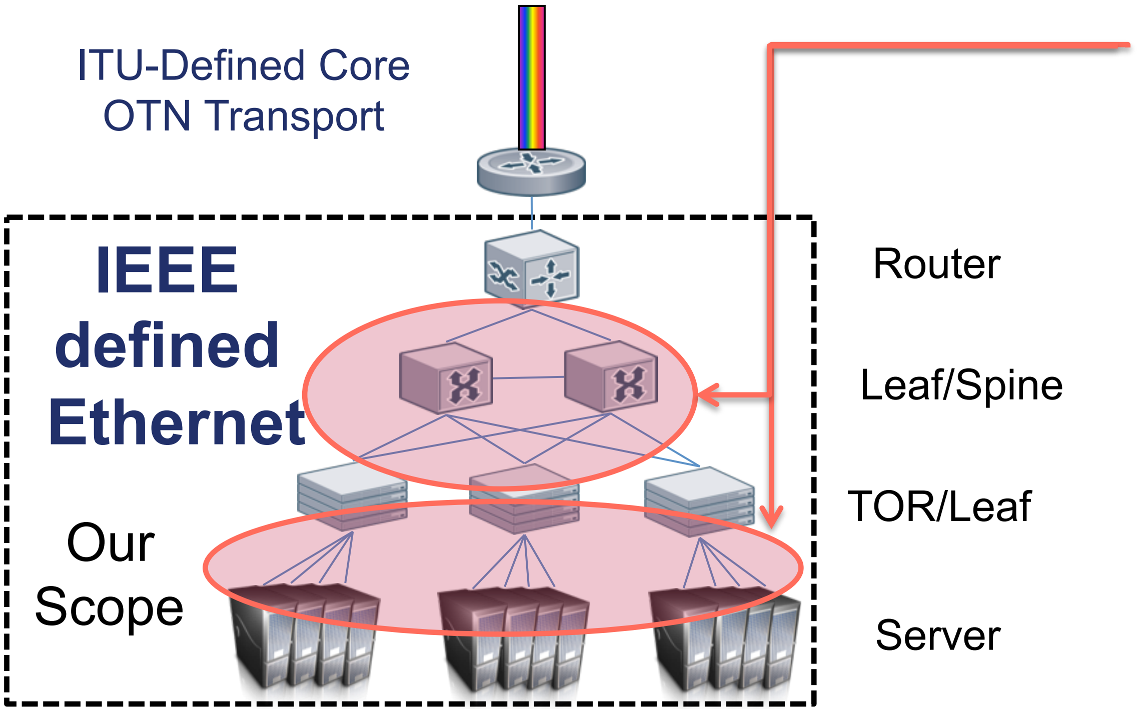
Provide 200 Gb/s MAC rate and PHYs

Provide denser 100 Gb/s PHYs

Web-scale data centers and cloud based services are presented as leading applications.

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

What Are We Talking About?



Leading Application Space for next generation Data Center Ethernet

- Optimized multi-lane interconnect for switch connectivity and intra-equipment interconnect (backplanes)
- Optimized single lane interconnect for servers and switches

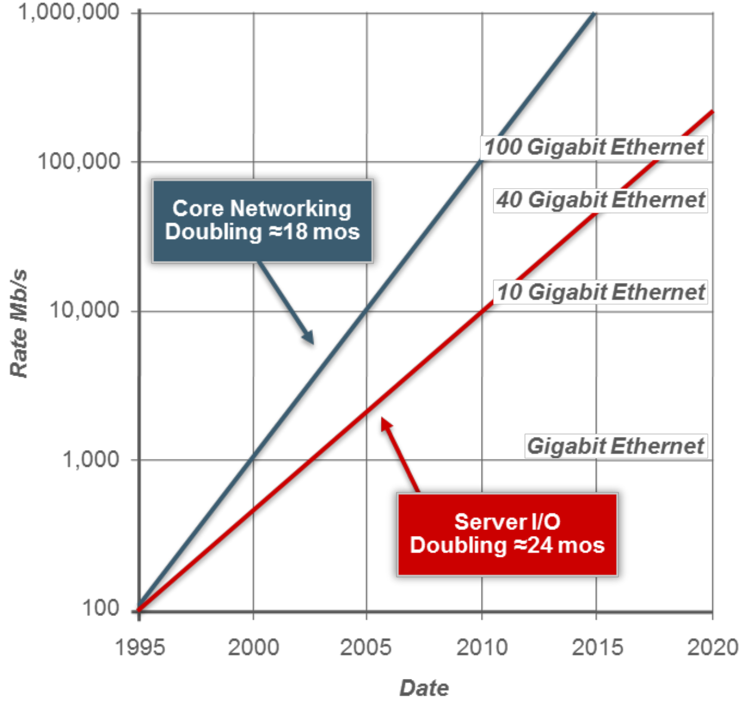
Agenda

- **Overview Discussion**
  - Next Gen Data Center Ethernet – Mark Nowell - Cisco
- **Presentations**
  - Market Drivers
    - John D’Ambrosia – Independent
  - Technical Feasibility
    - Adam Healey – Avago Technologies
  - Why Now?
    - Mark Nowell - Cisco
- **Straw Polls**

Market Drivers

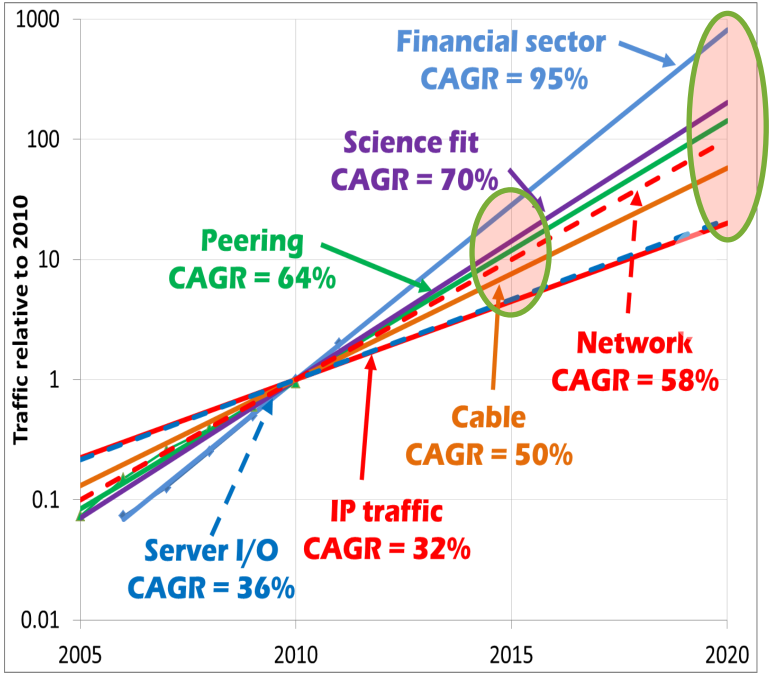
Market Diversity Examples

IEEE 802.3 HSSG - 2007



Source: http://www.ieee802.org/3/hssg/public/nov07/HSSG_Tutorial_1107.zip

IEEE 802.3 BWA - 2012



Source: http://www.ieee802.org/3/ad_hoc/bwa/BWA_Report.pdf

Ethernet Port Speed & Media Observations

Leading edge drives the higher speed technologies as soon as available

- Initial adoption: 10G ~2004; 40G ~2012; 100G ~2013; 400G ~2018

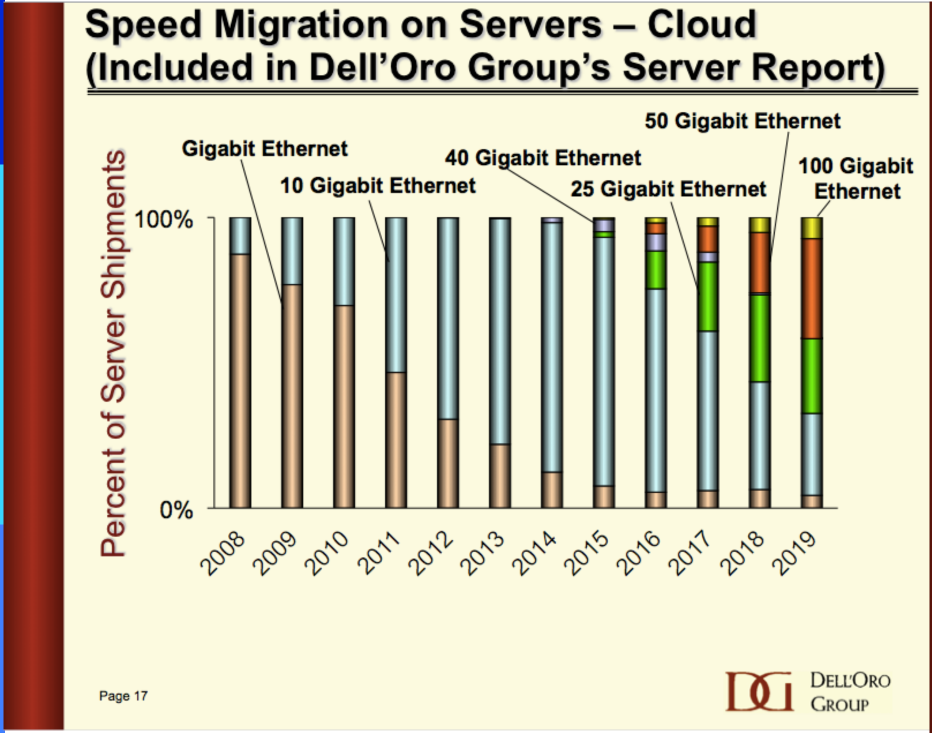
…but volume adoption is cost sensitive

- Single-lane availability can trigger volume adoption
- Multi-lane application of volume serial technology accelerates cost reduction

Server market is wide & varied – no single answer to the BW need question!

- Variety of CPU architectures, clock speeds, CPU core counts, CPUs/system
- Mix of software applications with varied needs of I/O BW vs. CPU compute power

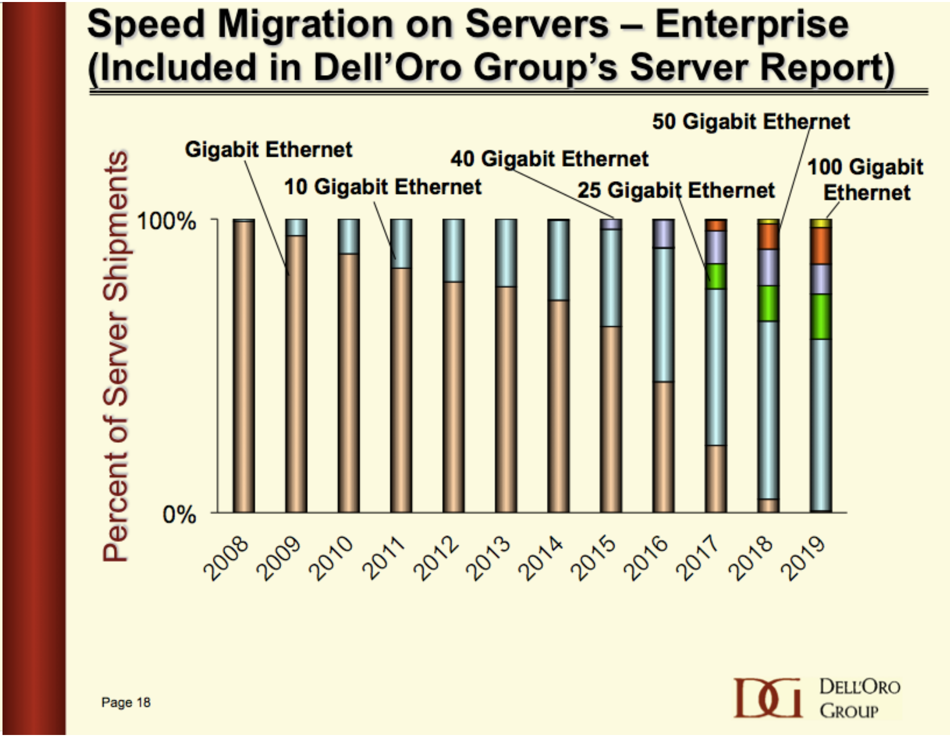
50G Server forecasts



Speed Migration on Servers – Cloud
(Included in Dell'Oro Group's Server Report)

Page 17

DELL'ORO GROUP



Speed Migration on Servers – Enterprise
(Included in Dell'Oro Group's Server Report)

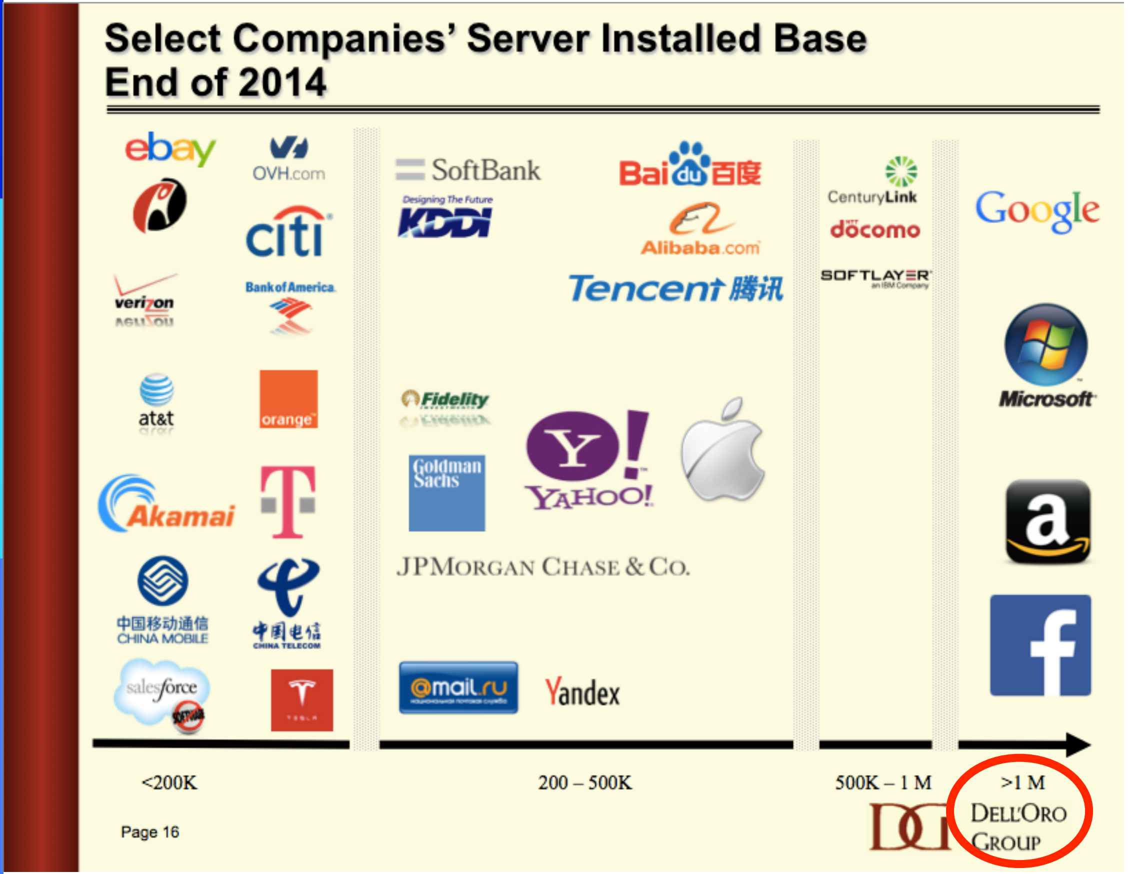
Page 18

DELL'ORO GROUP

Estimated 50G server port forecast for 2019 > 7M

Courtesy: Dell'Oro

Server Installed Base



Select Companies' Server Installed Base
End of 2014

Courtesy: Dell'Oro

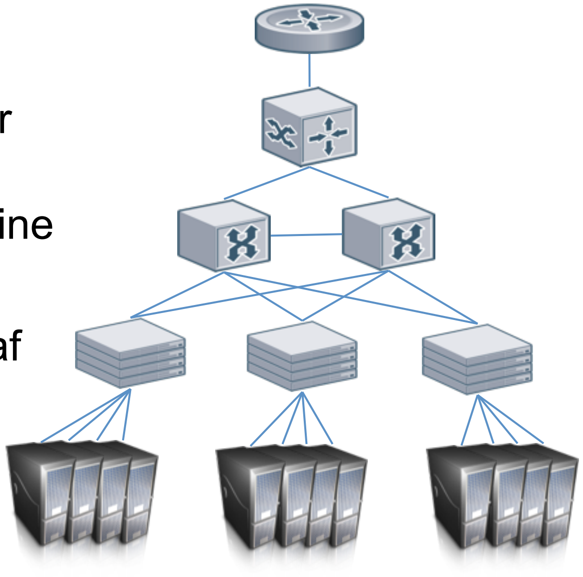
Industry shift towards large scale data centers results in large server deployments

- Cloud
- Enterprise

Motivation for large scale data centers is to drive for performance and cost optimizations

Potential Progression of Link Speeds

Interface speeds vary by location, time and Data Center scale



Router

Leaf/Spine

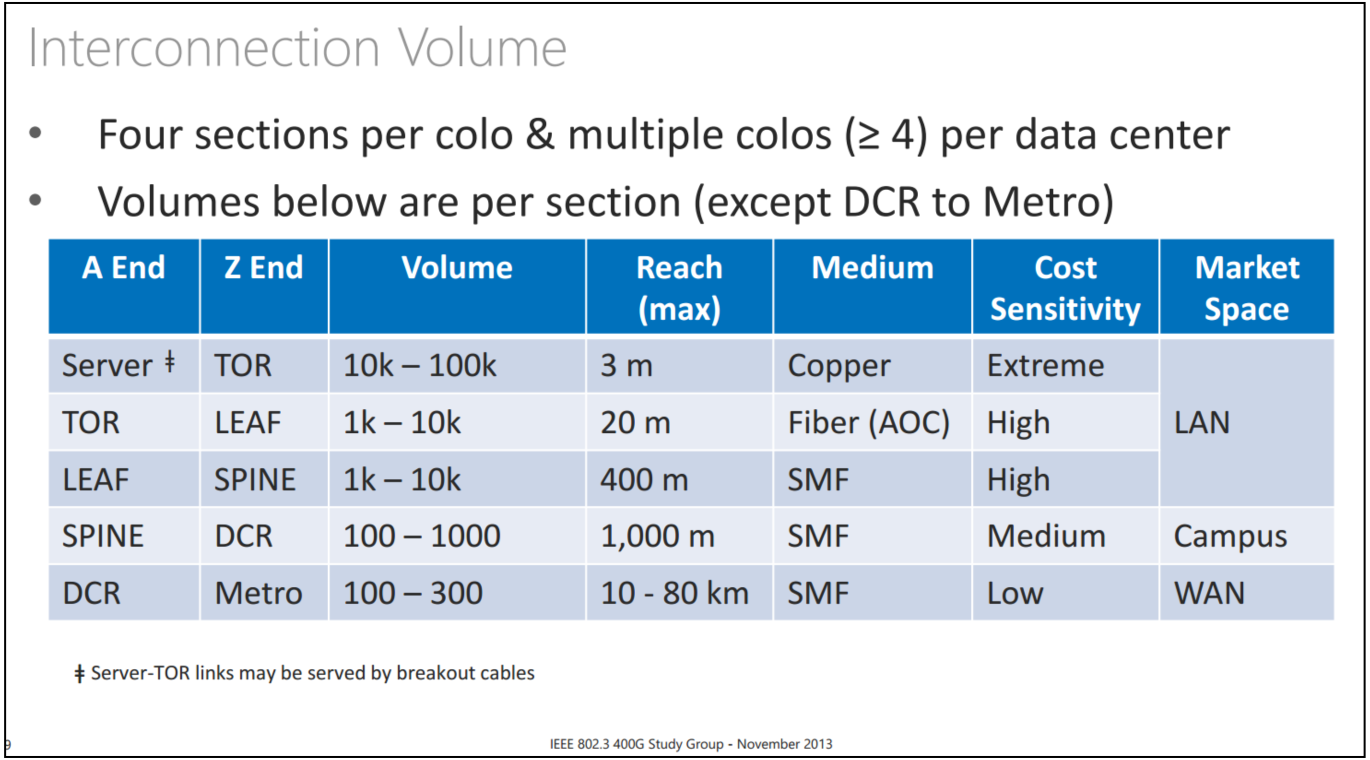
TOR/Leaf

Server

All speeds in GbE	Breadth of rates deployed		
	2012	**2016**	**2020**
Router	10, 40, 100	10, 40, 100	100, 200, 400
Leaf/ Spine	10, 40	10, 25, 40, 100	25, 40, 50, 100, 200, 400
ToR/ Leaf	1,10, 40	1,10, 25, 40,50 100	10, 25, 40, 50, 100, 200, 400
Server	1, 10	1, 10, 25, 40, 50	10, 25, 40, 50, 100, 200

50 GbE and 200 GbE have potential to be widely adopted

Data Center Interconnect Volume by Type



Interconnection Volume

- Four sections per colo & multiple colos (≥ 4) per data center
- Volumes below are per section (except DCR to Metro)

A End	Z End	Volume	Reach (max)	Medium	Cost Sensitivity	Market Space
Server ‡	TOR	10k – 100k	3 m	Copper	Extreme	LAN
TOR	LEAF	1k – 10k	20 m	Fiber (AOC)	High	
LEAF	SPINE	1k – 10k	400 m	SMF	High	
SPINE	DCR	100 – 1000	1,000 m	SMF	Medium	Campus
DCR	Metro	100 – 300	10 - 80 km	SMF	Low	WAN

‡ Server-TOR links may be served by breakout cables

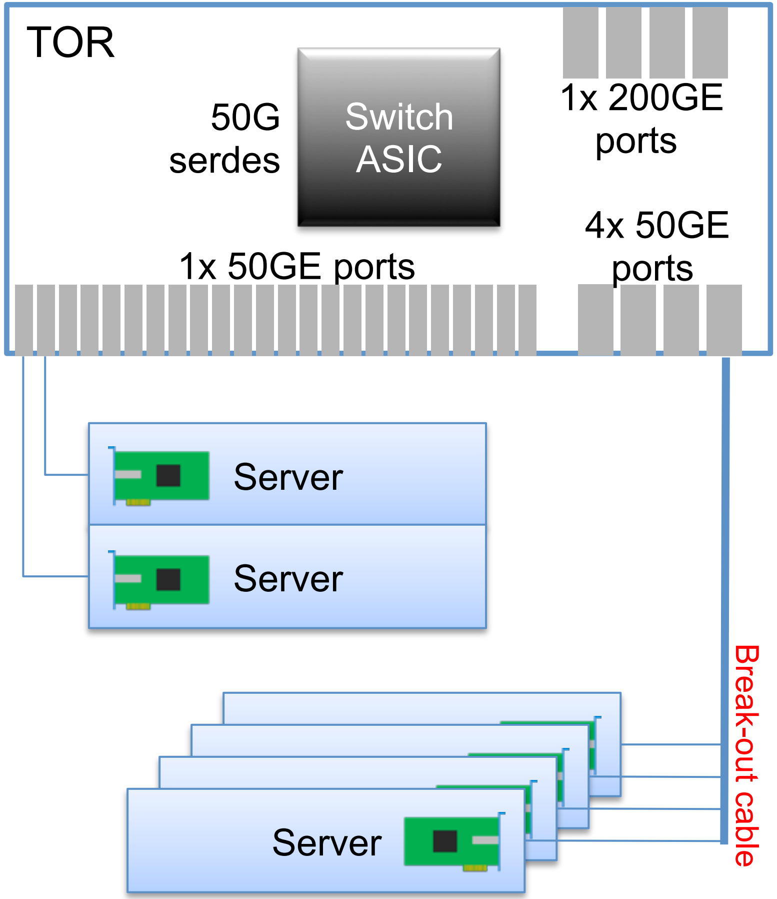
9 IEEE 802.3 400G Study Group - November 2013

Source: Brad Booth, Microsoft http://www.ieee802.org/3/400GSG/public/13_11/booth_400_01a_1113.pdf

- Cloud data center can have several 100k links
- Server interconnect, which drives the highest volume, requires cost effective solutions

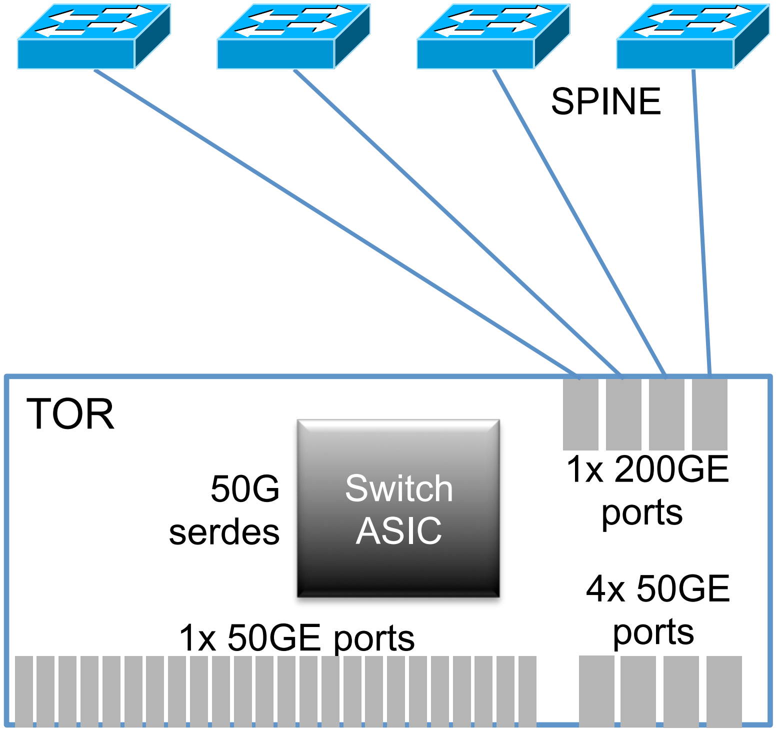
Single and Quad 50 Gb/s Ethernet Connectivity

- Enables similar topology as 10&40 Gb/s and 25 &100 Gb/s Ethernet
  - Single 50 Gb/s SFP56 port implementation or Quad 50 Gb/s QSFP56 breakout implementation possible
  - Maximizes ports and bandwidth in ToR switch faceplate
  - Dense rack server
  - Within rack, less than 3m typical length



200 Gb/s Ethernet Connectivity

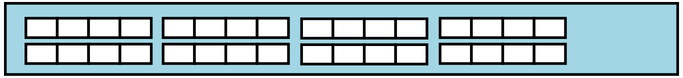
- Enables DC fabric topology similar to 40 Gb/s and 100 Gb/s Ethernet
  - Switch to switch fabric interconnect
  - Single-mode and multi-mode fiber or AOC
  - Switch-to-Switch typical reaches 100m (MMF) to ~2km (SMF)



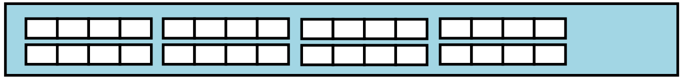
Network Demand for 200GbE

- 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.

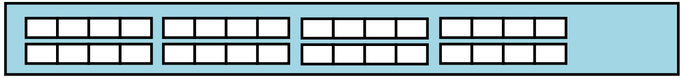
32 QSFP+ port Switch in 2012
4x10GbE Down, 40GbE up

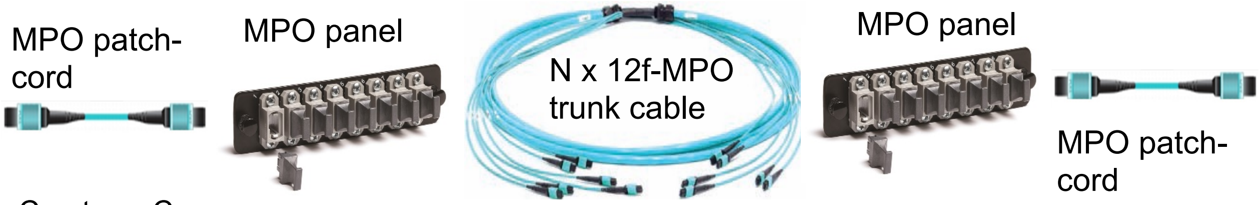


32 QSFP28 port Switch in 2016
4x25GbE Down, 100GbE up



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

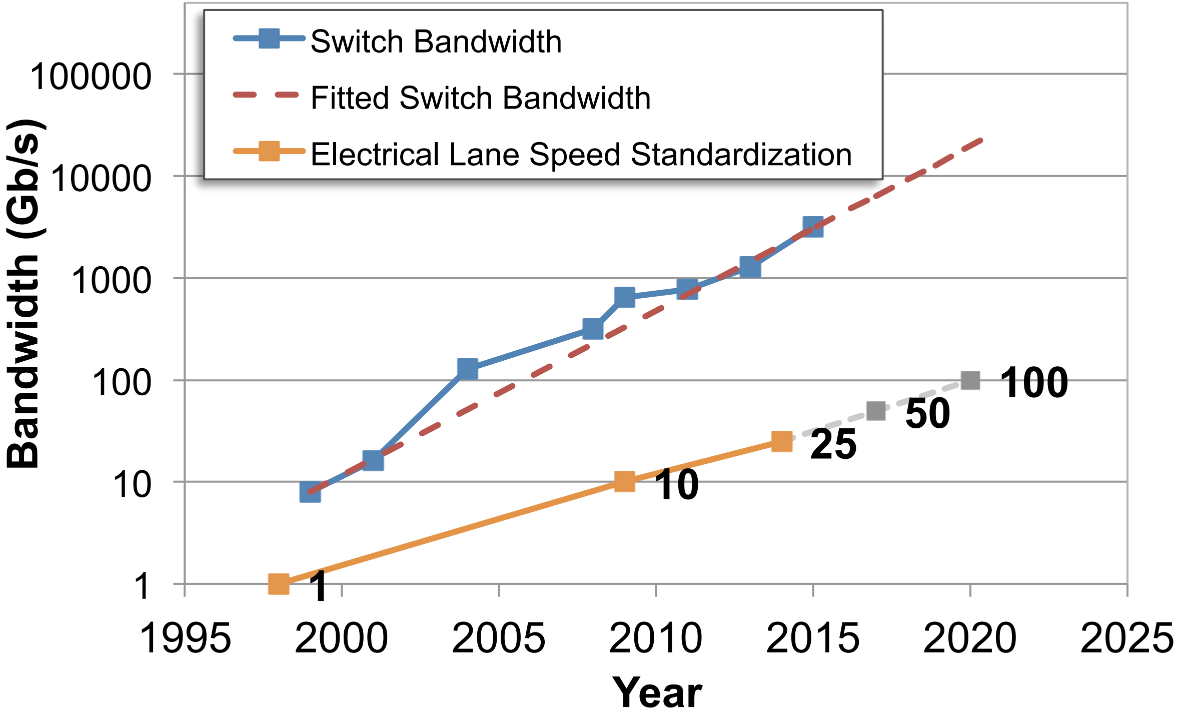




Courtesy: Commscope

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

Ethernet switch chip history and projections



Switch chip bandwidths continue to grow in order to support network capacity requirements

Switch IO rates have to increase in conjunction to maintain feasible implementations

Final Thoughts

- “Diversity” – the new “norm” for Ethernet
  - Server – applications, rates, number, and refresh cycles
  - Server to switch infrastructure - PHY type, density, and refresh cycles
  - Switches – bandwidth to support, PHY type, and refresh cycles
- Switching capacity requirements continue to grow
  - Industry investment in 50 Gb/s I/O
  - Re-use and familiarity with x1 / x4 architecture
- Leverage 50 Gb/s I/O technology for a family of Ethernet rates that leverages high volume server opportunities to lower cost everywhere!

Technology Feasibility

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

MAC/PCS Technical Feasibility

- 50G and 100G MACs have been implemented in the industry already, 200G MACs are also feasible in existing technology
- A PCS for each possible speed (50G, 100G and 200G) is feasible and can leverage existing technology, some possible PCS choices are:
  - Leverage 802.3bj and/or 802.3bs logic architectures
  - Can include either KR4 FEC (RS[528,514,10]) and KP4 FEC (RS[544,514,10])
- With FinFET process technology, it is possible to develop faster, lower power, compact IP that occupies a small fraction of an ASIC/FPGA.
- Time-sliced MAC/PCS designs are feasible and can handle multi-rate implementations

50 Gb/s SERDES Technical Feasibility

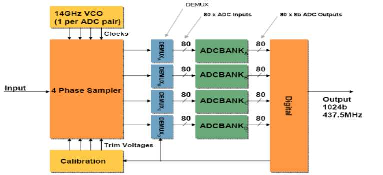
- P802.3bs specifying 8x50G PAM4 CDAUI-8 C2C/C2M interfaces, as well as 8x50G and 4x100G optical interfaces.
- Several OIF projects targeting 56-60G data rate including LR, MR, VSR, XSR, and USR with PAM4 and NRZ modulation.
- PHYs using 50G PAM4 with FEC can reuse electrical channels similar to those specified and designed for 25G NRZ such as 100GBASE-CR4/KR4, 25GBASE-CR/KR, and CAUI-4.
- Similarly, previously defined channel models for circuit boards, direct attach cables, and connectors can be used for 50G PAM4 with minor modification.
- Reduced lane counts of CDAUI-8 for C2C and C2M.

50 Gb/s SERDES Technical Feasibility

From 802.3bs

Existing 63GS/s 8 bit ADC

- 63GS/s ADC
- 320X time interleaving
- 8 bit ADC ➔ ENOB ≈ 6bit
- 40nm CMOS process
- Power = 1250mW
- OFC 2010

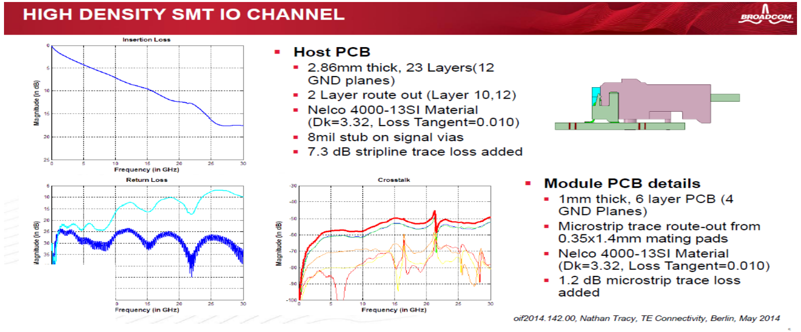


http://www.fujitsu.com/downloads/MICRO/fme/dataconverters/OFC-2010-56Gss-ADC-En

- 28nm provides ~ 30% power saving
- 13.5G ADC requires ~ 76% less power

➔ 28nm 8bit 13.5G ADC power ~ 190mW

PAM4 digital receiver perf.. & feasibility
http://www.ieee802.org/3/bj/public/jan12/parthasarathy_01_0112.pdf



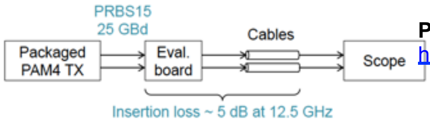
HIGH DENSITY SMT IO CHANNEL

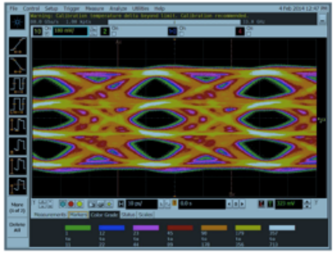
- Host PCB
  - 2.86mm thick, 23 Layers(12 GND planes)
  - 2 Layer route out (Layer 10,12)
  - Nelco 4000-13SI Material (Dk=3.32, Loss Tangent=0.010)
  - 8mil stub on signal vias
  - 7.3 dB stripline trace loss added
- Module PCB details
  - 1mm thick, 6 layer PCB (4 GND Planes)
  - Microstrip trace route-out from 0.35x1.4mm mating pads
  - Nelco 4000-13SI Material (Dk=3.32, Loss Tangent=0.010)
  - 1.2 dB microstrip trace loss added

oif2014.142.00, Nathan Tracy, TE Connectivity, Berlin, May 2014

PAM4 MODULATION FOR THE 400G ELECTRICAL INTERFACE
http://www.ieee802.org/3/bs/public/14_07/domb_3bs_01a_0714.pdf

Measured results: Experiment #1





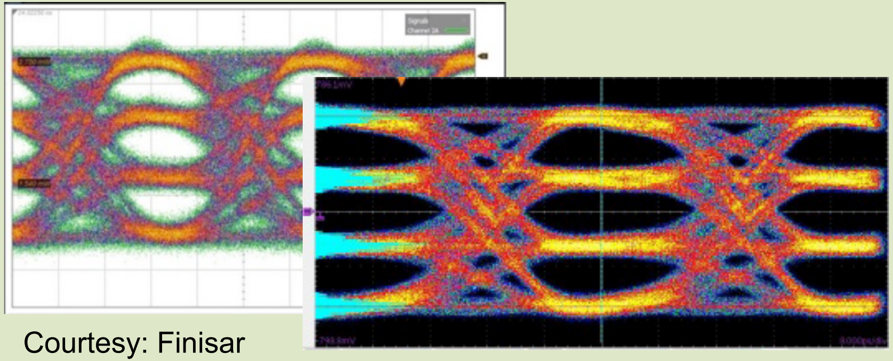
IEEE P802.3bs 400 GbE Task Force - November 2014

CDAUI-8 chip-to-module and chip-to-chip interfaces using PAM4
http://www.ieee802.org/3/bs/public/14_11/healey_3bs_01_1114.pdf

50 Gb/s Single lane technology direction is established and in development

Optical Technical Feasibility

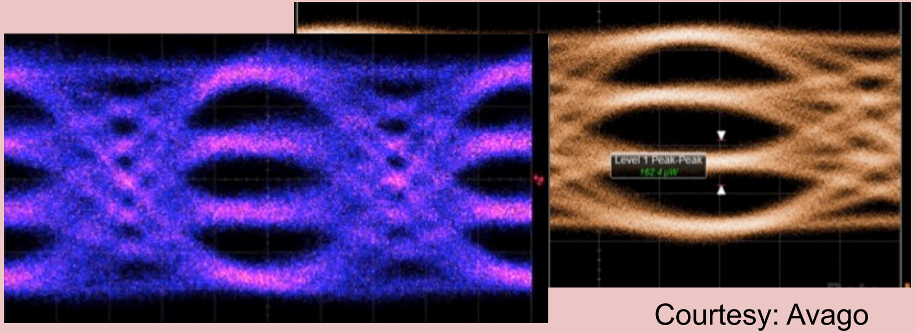
1300nm 56G PAM4



Courtesy: Finisar

Courtesy: Cisco

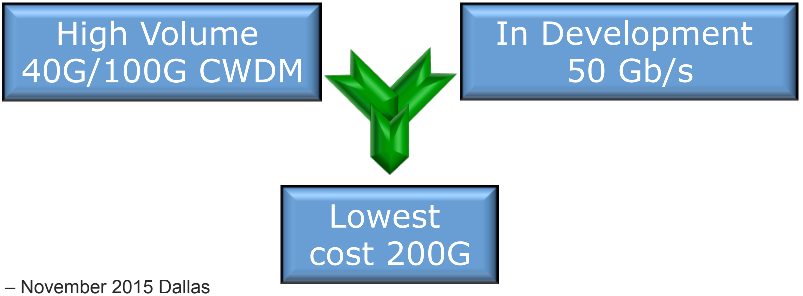
850nm 50G PAM4



Courtesy: Avago

Courtesy: Finisar

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



High Volume 40G/100G CWDM

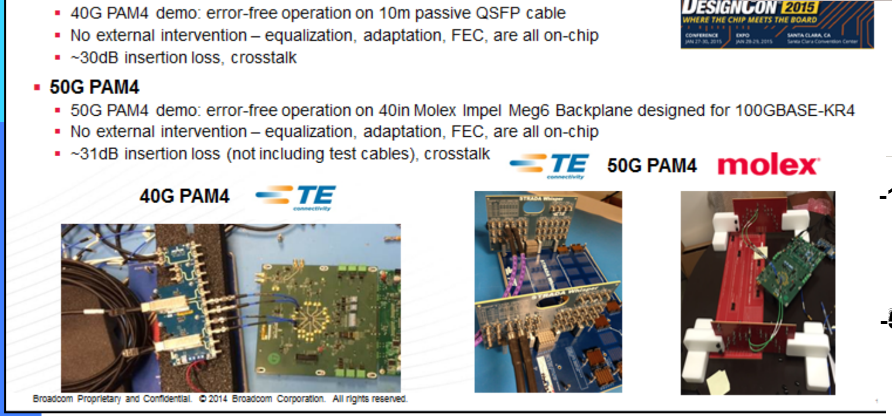
In Development 50 Gb/s

Lowest cost 200G

Connector/Cable/Backplane Technical Feasibility

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

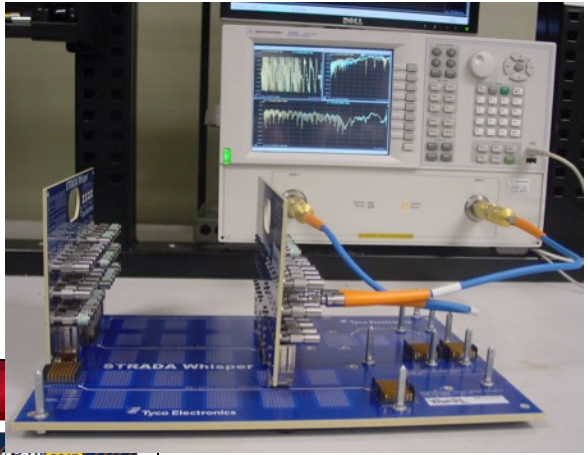
BROADCOM DESIGNCON 2015 DEMOS



- **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

Broadcom Proprietary and Confidential. © 2014 Broadcom Corporation. All rights reserved.

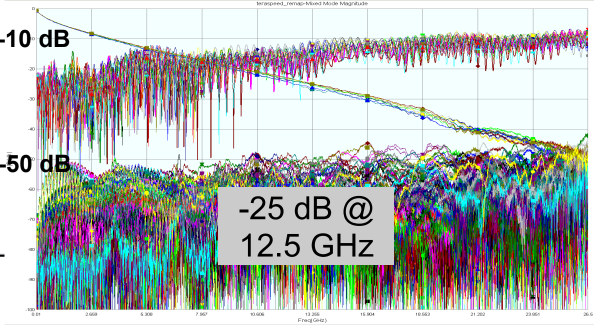
http://www.ieee802.org/3/bs/public/adhoc/elect/15_0212/parthasarathy_01_0215_elect.pdf



1m BP @ DesignCon 2015

- Demonstrated with both 50G PAM4 & NRZ

Courtesy: TE

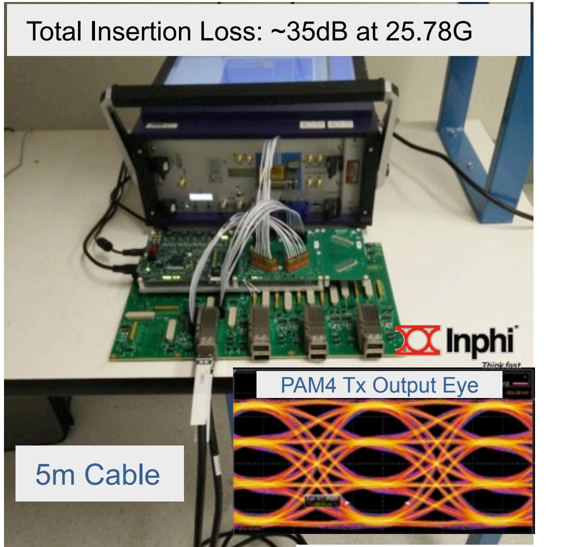


29” Backplane design

Courtesy: Teraspeed Consulting - A Division of Samtec

50G PAM4

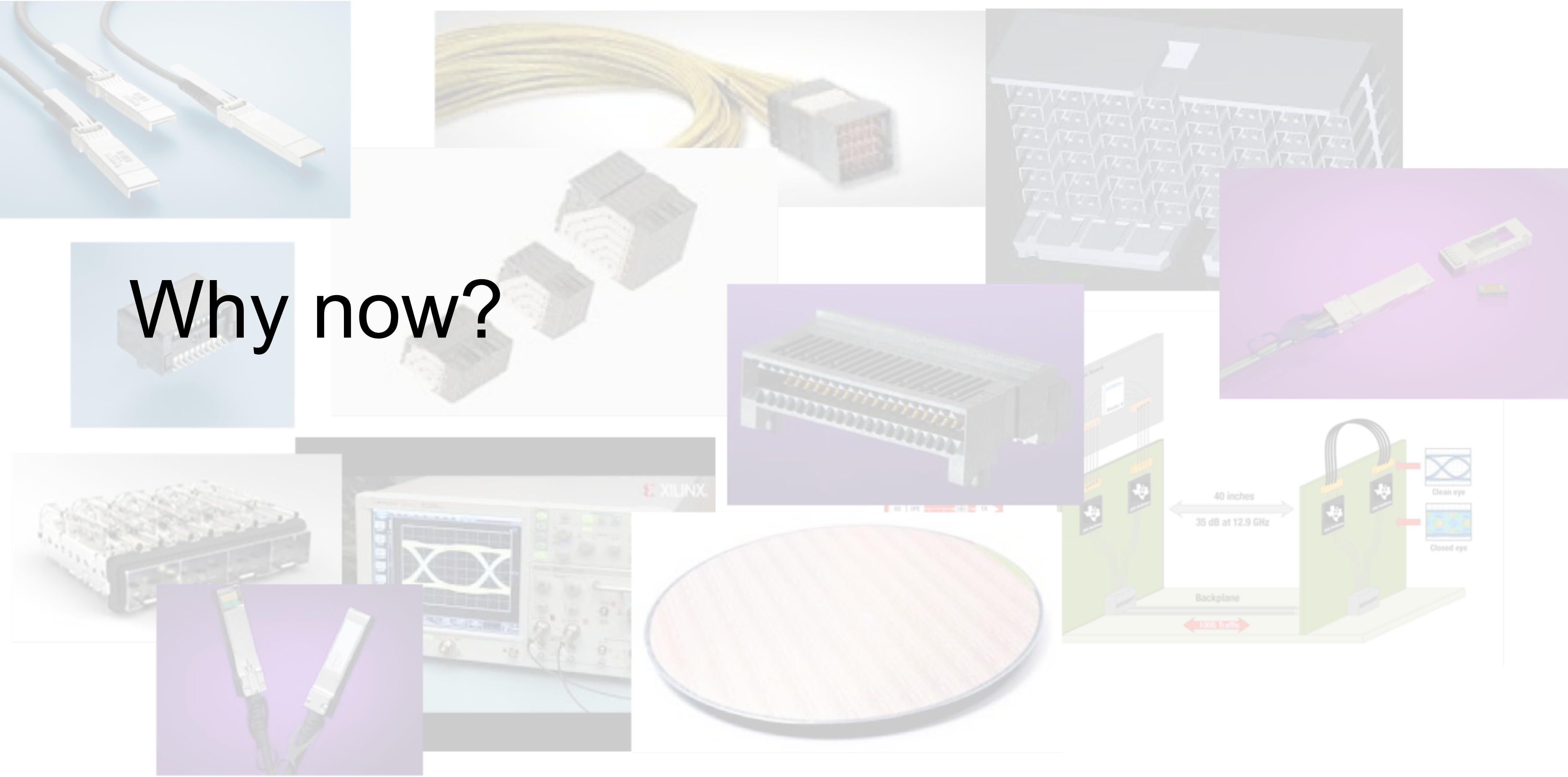
Total Insertion Loss: ~35dB at 25.78G



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

Courtesy: Inphi

Why now?

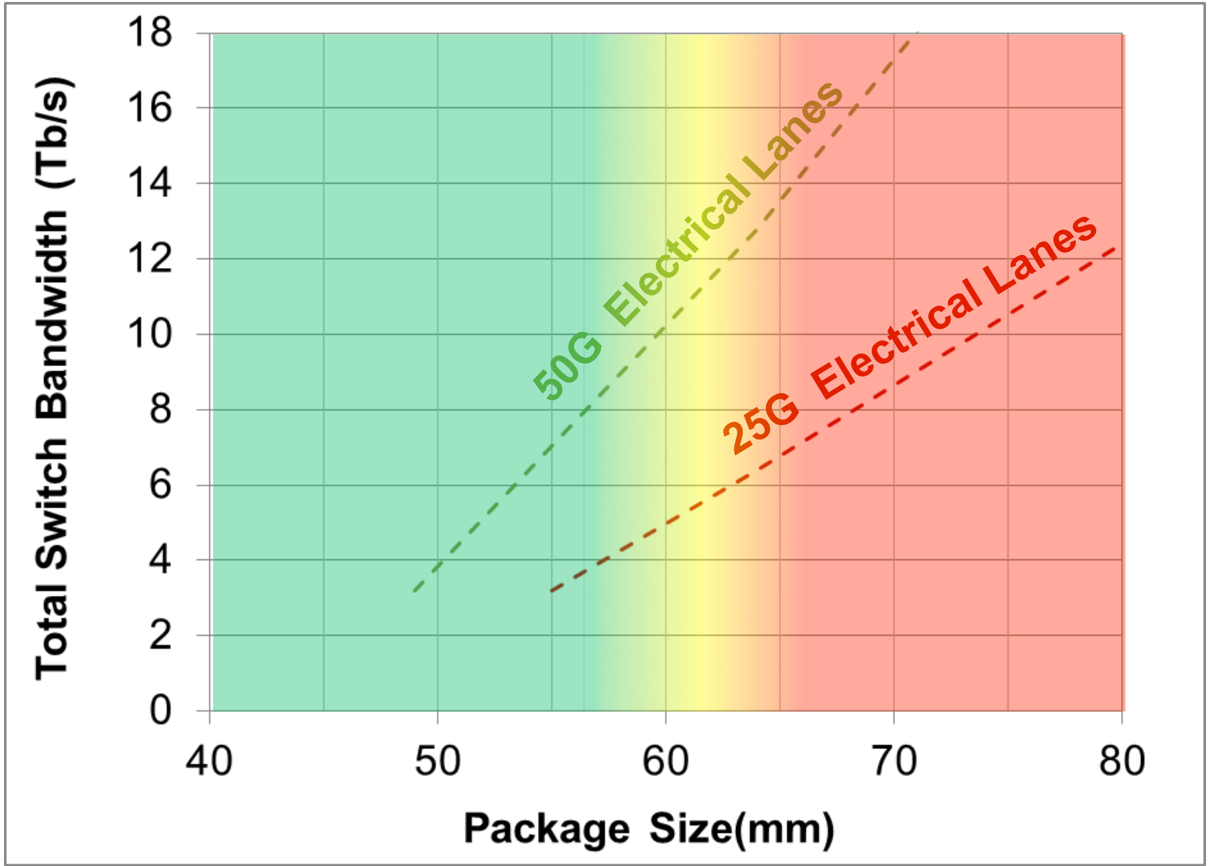

XILINX
40 inches
35 dB at 12.9 GHz
Clean eye
Closed eye
Backplane

The key questions

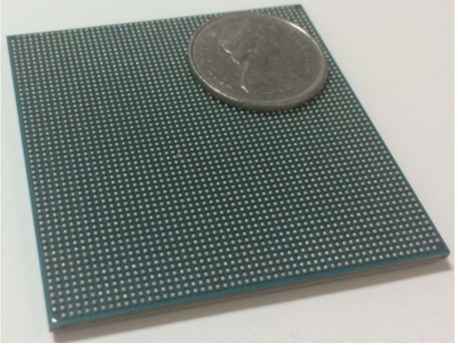
- Can solutions be built to keep up with the requirements driven by the bandwidth growth?
- Can those solutions be cost effective?

Switch Chip Package Limitations

(Avoiding Technical Infeasibility)



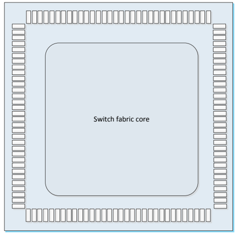
BGA package limitations require eventual migration to higher electrical lane speeds to support higher bandwidth switches



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.

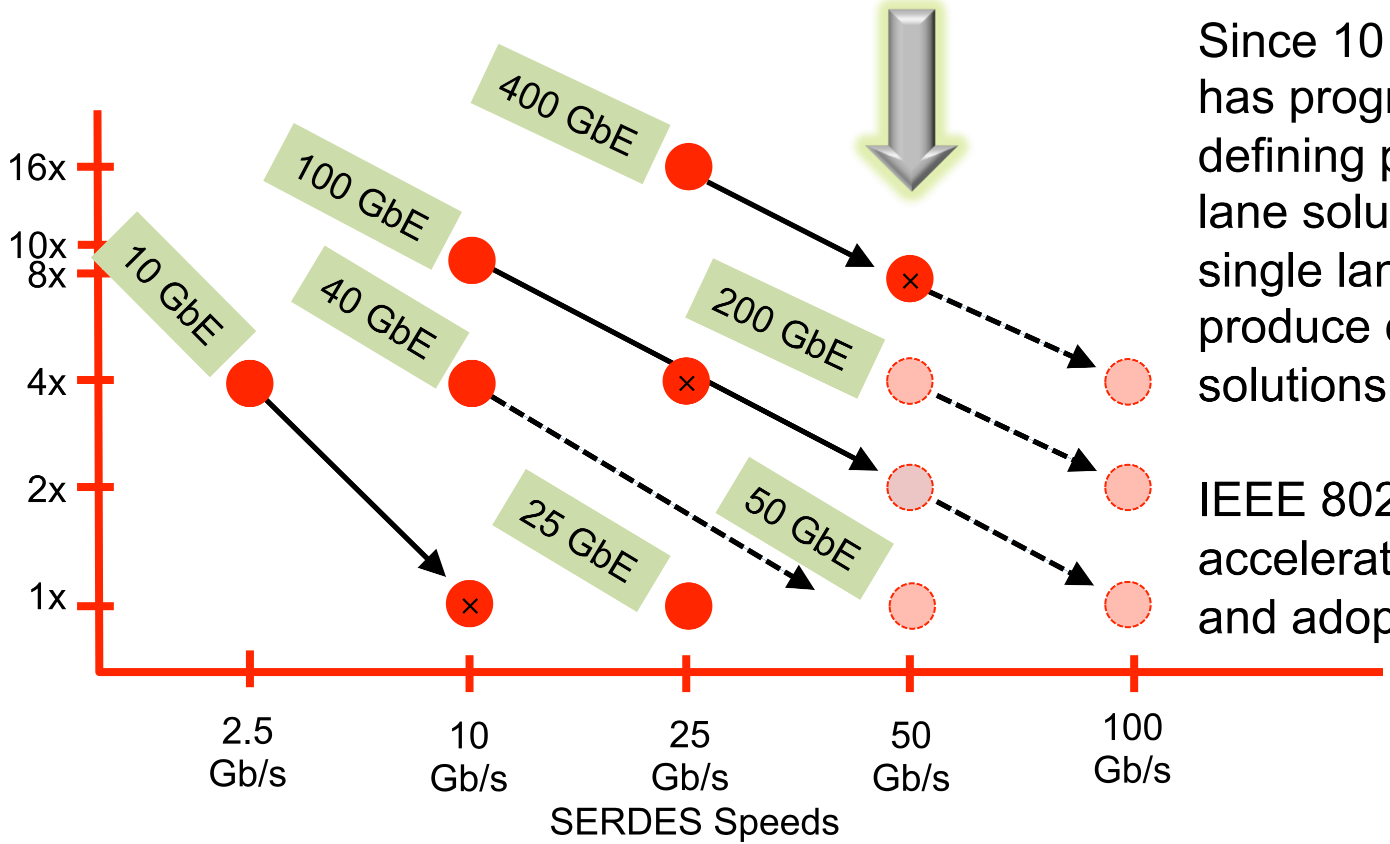
Switch Radix Capabilities

	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**

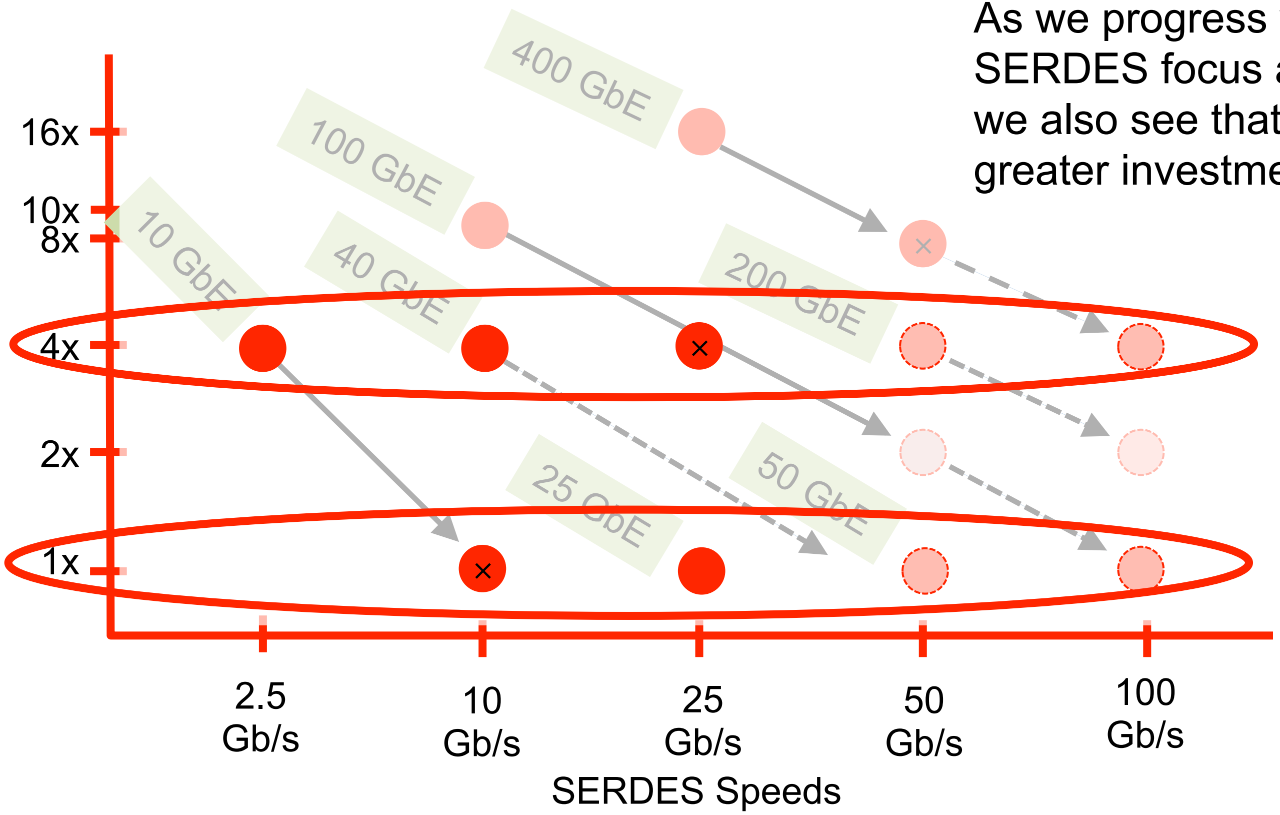
The new normal – multi-lane and re-use



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).

Why Now?

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

Mark Nowell – Cisco
Rob Stone – Broadcom
John D’Ambrosia – Independent
Brad Booth – Microsoft
Gary Nicholl – Cisco
Scott Kipp – Brocade
Yong Kim – Broadcom
Matt Brown – APM
Mark Gustlin – Xilinx
Chris Cole – Finisar
Ed Sayre - Samtec
Chris Roth – Molex
David Ofelt - Juniper

Mike Li – Altera
Joel Goergen – Cisco
Kent Lusted – Intel
Jonathan Ingham – Avago
Marco Mazzini – Cisco
Nathan Tracy – TE
Sudeep Bhoja – Inphi
Vipul Bhatt – Inphi
Adam Healey – Avago
Vasu Parthasarathy – Broadcom
Scott McMorrow – Samtec
Kapil Shrikhande – Dell

Supporters (98 Individuals from 58 companies)

Vasudevan Parthasarathy - Broadcom
Jacky Chang - HPE
Kapil Shrikhande - Dell
Kohichi Tamura - Oclaro
Scott Irwin - MoSys
Matt Brown - APM
Gary Nicholl - Cisco
Robert Lingle - OFS
Velu Pillai - Broadcom
Vivek Telang - Broadcom
Rajiv Pancholy - Broadcom
Andre Szczepanek - Inphi
Sudeep Bhoja - Inphi
Michael Johas Teener - Broadcom
Thananya Baldwin - Ixia
Martin White - Cavium Networks
Chris Roth - Molex
Dave Ofelt - Juniper
Dale Murray - LightCounting

Jonathan Ingham - Avago
Randy Rannow - APIC
Vipul Bhatt - Inphi
Dan Dove - Dove Networking
Paul Kolesar - Commscope
Jerry Pepper - Ixia
Ron Muir - JAE
Sam Sambasivan - AT&T
Andy Moorwood - Ericsson
Oded Wertheim - Mellanox
Dave Chalupsky - Intel
Jon Lewis - Dell
Brian Teipen - Adva
Bharat Tailor - Semtech
Brian Welch - Luxtera
Peter Jones - Cisco
Rakesh Sambaraju - Berk-Tek
Arthur Marris - Cadence
Vineet Salunke - Cisco

Gary Bernstein - Leviton
Mark Gustlin - Xilinx
Kent Lusted - Intel
Mike Li - Altera
Zhao Wenyu - China Academy of
Information & Comm. Tech.
Xu (Helen) Yu - Huawei
Moonsoo Park - OE Solutions
Scott Kipp - Brocade
Xinyuan Wang - Huawei
Keith Conroy - Multi-PHY
Ryan Latchman - MaCom
Scott Sommers - Molex
Adam Healey - Avago
Richard Mellitz - Intel
Yong Kim - Broadcom
Rick Rabinovich - ALE USA
Chris Cole - Finisar
John D'Ambrosia - Independent
Vittal Balasubramani - Dell

Supporters (2)

Paul Mooney - Spirent
Greg LeCheminant - Keysight Tech
Shoukei Kobayshi - NTT
Jeff Maki - Juniper
David Malicoat - HPE
Daniel Dillow - FCI
Tom Issenhuth - Microsoft
James Fife - eTopus
Ali Ghiasi - Ghiasi Quantum
Raj Hegde - Broadcom
Greg McSorley - Amphenol
Henry Chen - Broadcom
Pete Anslow - Ciena
Ghani Abbas - Ericsson
Jim Theodoras - ADVA
Doug Coleman - Corning
Peter Cibula - Intel
Jonathan King - Finisar
Steve Trowbridge - Alcatel-Lucent
Mike Andrewartha - Microsoft
Pirooz Tooyserkani - Cisco
Martin Skagen - Brocade
Stefano Valle - ST Micro
Mike Ressl - Hitachi Cable
Hai-Feng Liu - Intel
Scott Schube - Intel
Arnold Sodder - Mercury Systems
Ed Ulrichs - Source Photonics
Mike Dudek - Qlogic
Mike Bennett - 3MG Consulting
Hesham ElBakoury - Huawei
Brad Booth - Microsoft
Nathan Tracy - TE
Peter Stassar - Huawei
Tom Palkert - Molex
Adee Ran - Intel
Paul Brooks - Viavi
David Estes - Spirent
Gary Bernstein - Leviton
Matt Traverso - Cisco
Andy Zambell - FCI
Petar Pepaljugoski - IBM

Straw Polls

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

Call-for-Interest Consensus

- Should a study group be formed for “50 Gigabit/s Ethernet over a single lane”?
- Y: 127 N: 0 A:5
- Room count: 134

Call-for-Interest Consensus

- Should a study group be formed for “Next Generation 100 & 200 Gigabit/s Ethernet”?
- Y: 124 N: 0 A:4
- Room count: 134

Participation

- I would participate in a “50 Gigabit/s Ethernet over a single lane” study group in IEEE 802.3
  - Tally: 102
- I would participate in a “Next Generation 100 & 200 Gigabit/s Ethernet” study group in IEEE 802.3
  - Tally: 103
- My company would support participation in a “50 Gigabit/s Ethernet over a single lane” study group
  - Tally: 66
- My company would support participation in a “Next Generation 100 & 200 Gigabit/s Ethernet” study group
  - Tally: 66

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