25 Gb/s Ethernet Over a Single Lane for Server Interconnect Call For Interest Consensus

IEEE 802 July 2014 Plenary, San Diego, CA

Introductions for today's presentation

Presenter and Expert Panel:

Brad Booth - Microsoft
Dave Chalupsky - Intel
John D'Ambrosia – Dell
Howard Frazier - Broadcom
Joel Goergen - Cisco
Mark Nowell - Cisco

Objectives

- To gauge the interest in starting a study group to investigate a “25 Gb/s Ethernet over a single lane for server interconnect” project

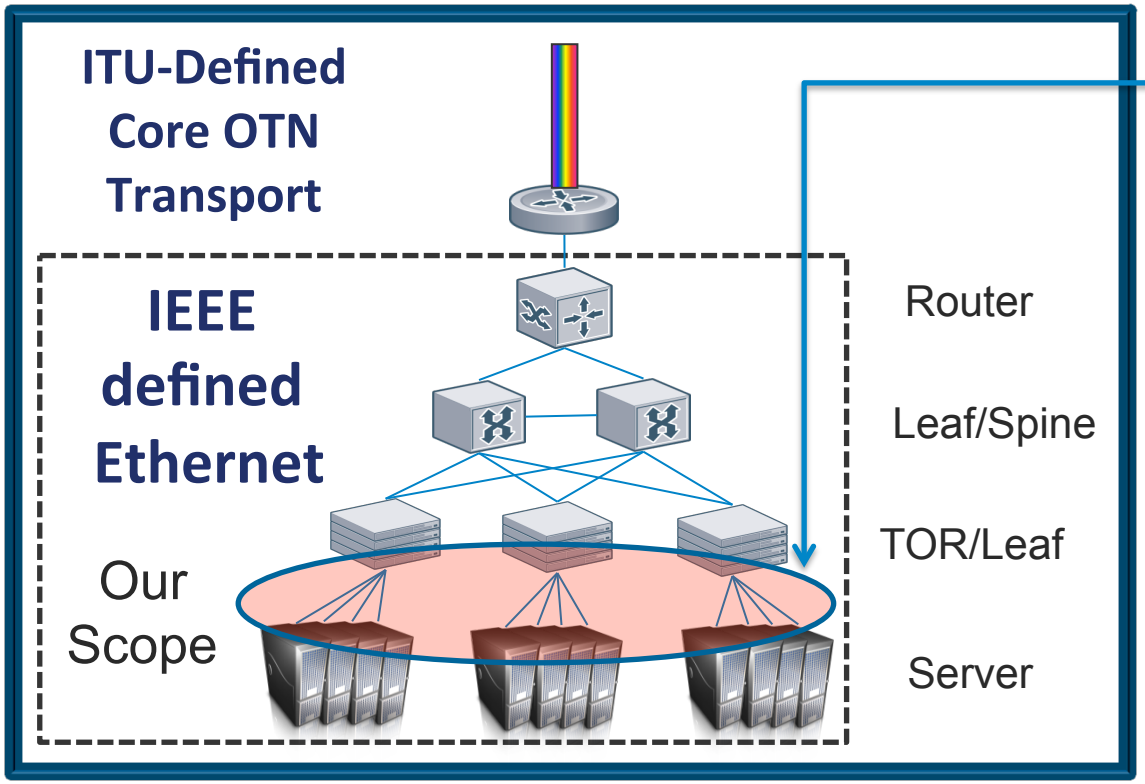
- We do not need to:
  - Fully explore the problem
  - Debate strengths and weaknesses of solutions
  - Choose a solution
  - Create a PAR or 5 Criteria
  - Create a standard
- Anyone in the room may vote or speak

Overview: 25Gb/s Ethernet Motivation

- Provide cost optimized server capability beyond 10G
- Provide a 25Gb/s MAC rate that:
  - Leverages single-lane 25Gb/s physical layer technology developed to support 100GbE
  - Maximize efficiency of server to access switch interconnect

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

What Are We Talking About?



Leading Application Space for 25Gb/s Ethernet

- Optimized interconnect from servers to first-level networking equipment (i.e. ToR, access layer, leaf…)
- A single-lane 25Gb/s Ethernet interface provides the opportunity for optimum cost-performance server interconnect

Agenda

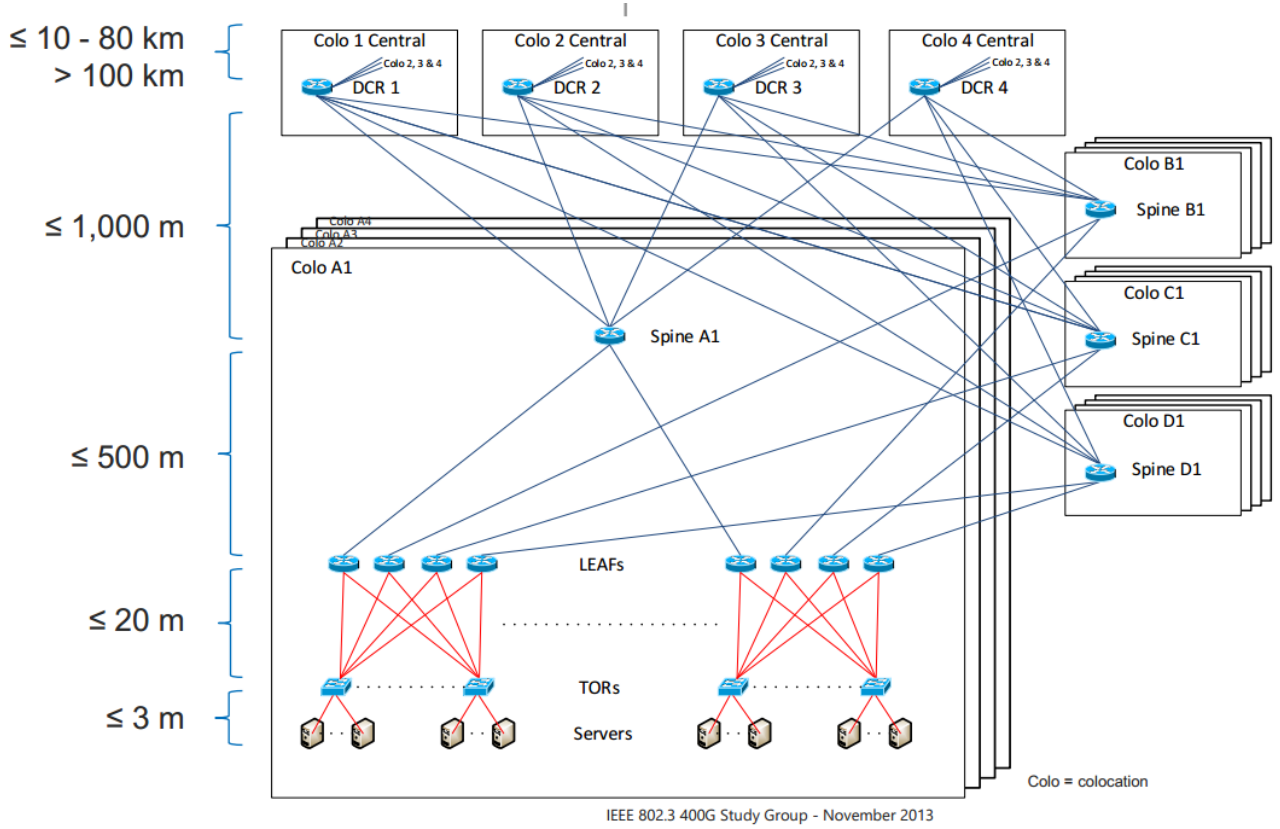
- **Overview Discussion**
  - 25 Gb/s Ethernet – Mark Nowell - Cisco
- **Presentations**
  - 25 Gb/s Ethernet Market Drivers – David Chalupsky - Intel
  - 25 Gb/s Ethernet Technical Feasibility – Howard Frazier - Broadcom
  - 25 Gb/s Ethernet: Why Now? – John D'Ambrosia - Dell
- **Straw Polls**

Market Drivers

25 Gb/s Ethernet Market Drivers – David Chalupsky - Intel

Where are the Server Links in the Cloud Data Center?

- The term “TOR” has become synonymous with server access switch, even if it is not located “top of rack”



From http://www.ieee802.org/3/400GSG/public/13_11/booth_400_01a_1113.pdf

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 ǂ	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

ǂ Server-TOR links may be served by breakout cables

IEEE 802.3 400G Study Group - November 2013

From http://www.ieee802.org/3/400GSG/public/13_11/booth_400_01a_1113.pdf

- Server interconnect drives the highest volume, has shortest reach need
- Cloud data center can have several 100k links

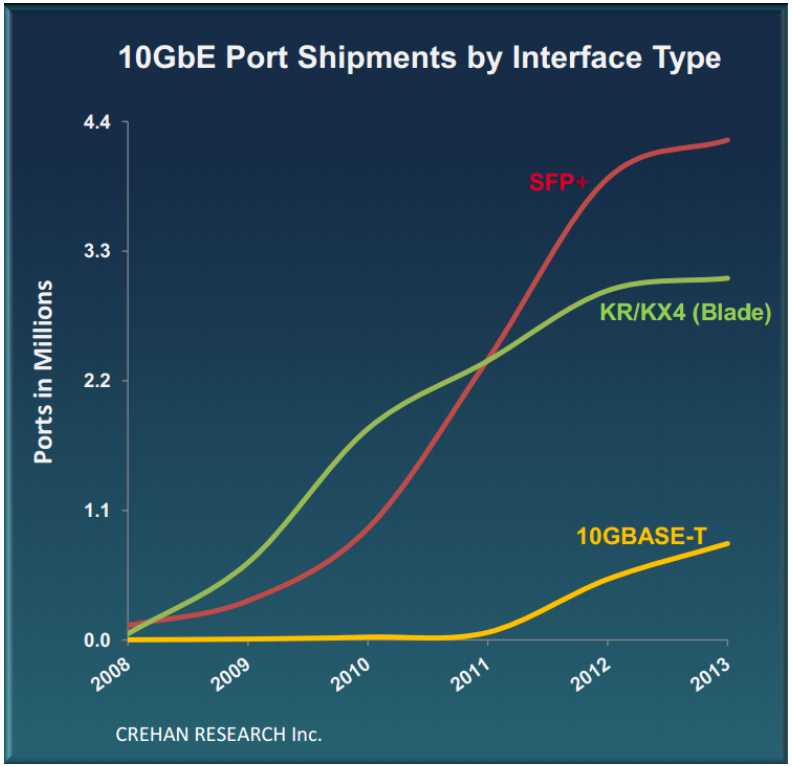
Single Lane interfaces in 10GbE Server

- 10GbE volume ramp in servers coincided with the availability of single-lane interfaces
- Early adopters (2004-2008) used
  - XAUI-based optics
  - 10GBASE-CX4
  - 10GBASE-KX4
- Single-lane backplane and twinax solutions eclipsed the early-adopter volume starting in 2009

Chart notes

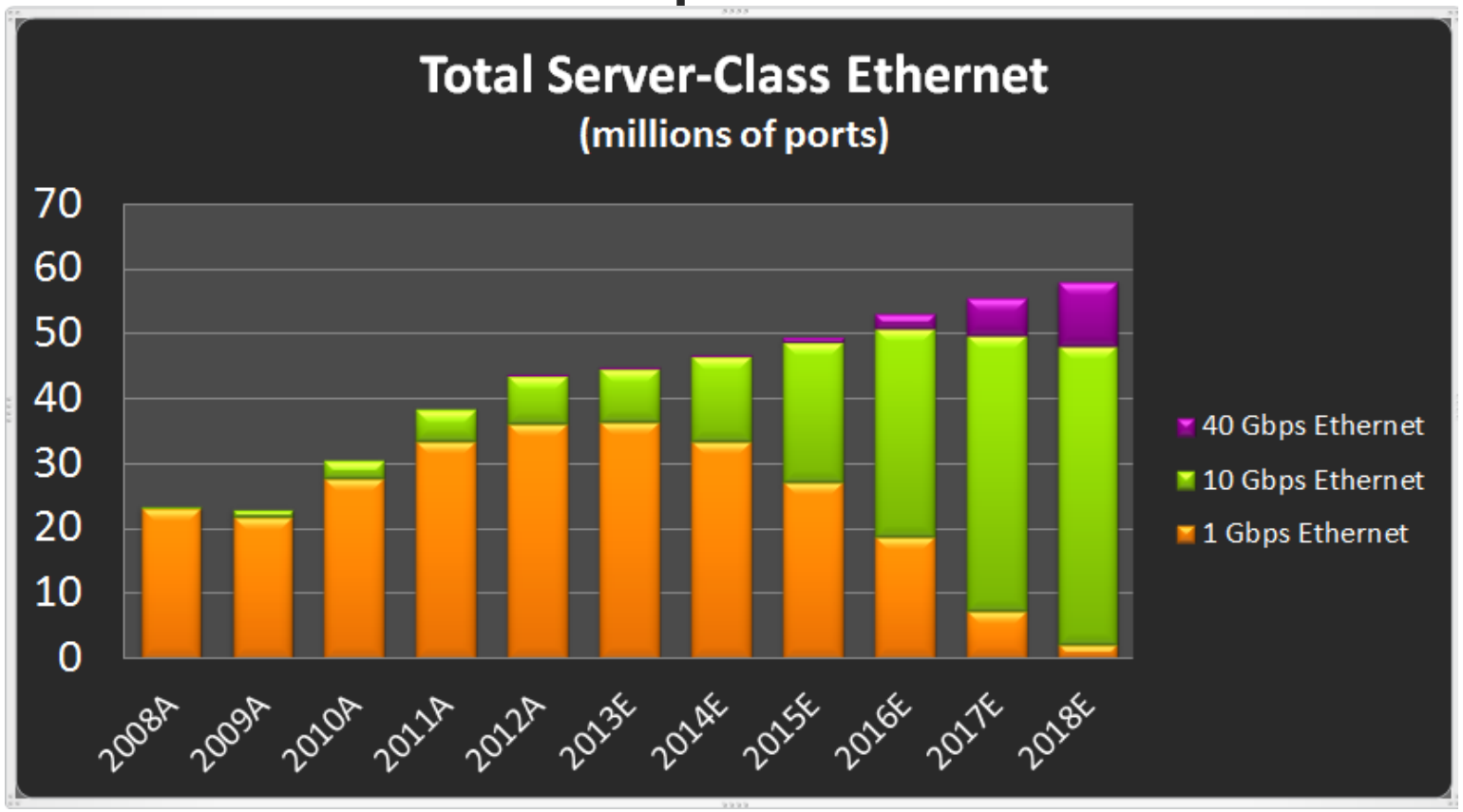
- "Other" category, not shown, went from ~12% in 2008 to <1% in 2013
- SFP+ majority use is twinax, then SR; accurate share data unavailable
- Blade server is mostly KR based upon system configuration. KX4 vs. KR split data unavailable.

10GbE Adapter/LOM Port Interface Mix



Data source: Crehan Research, Inc., Q1'2014

Server Ethernet Port Speed Forecast



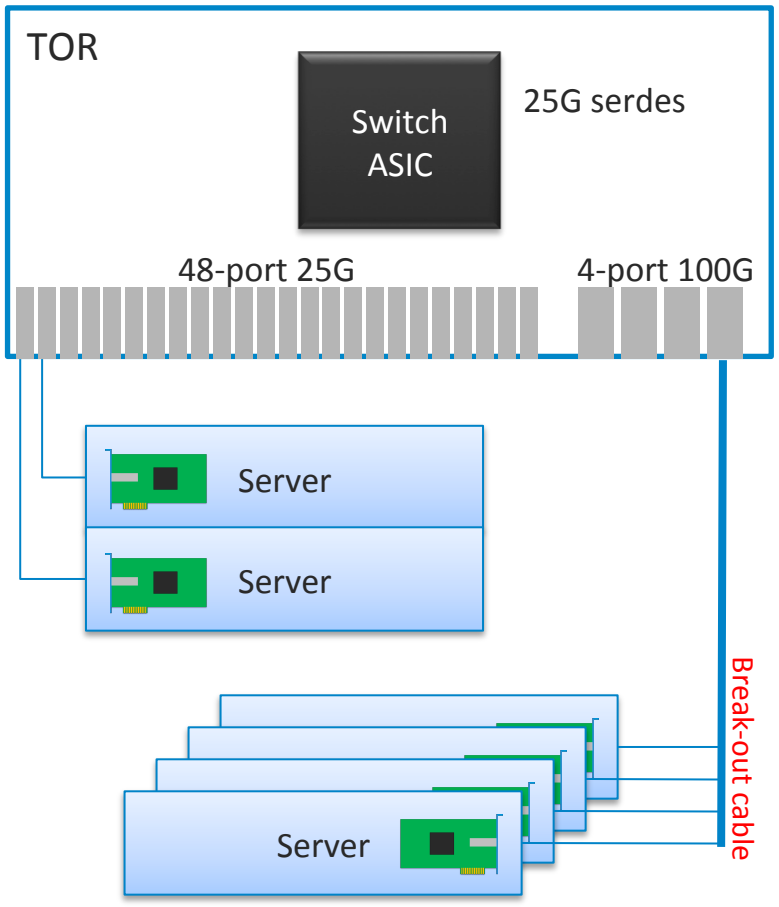
Data source: Crehan Research, Inc., Q1'2014

Server Ethernet Port Speed & Media Observations

- Market is wide & varied – no single answer to the BW need question!
  - Port speeds from 1Gb/s to 100Gb/s will co-exist
  - Variety of CPU architectures, clock speeds, core counts, CPUs/system
  - Mix of software applications with varied needs of I/O BW vs. CPU compute power
- Leading edge drives the higher speed as soon as available
  - Initial adoption: 10G ~2004; 40G ~2012; 100G ~2015
- …but volume adoption is cost sensitive
- 1G→10G crossover forecast has repeatedly shifted right
    - 2012→2014→2016
    - In turn, transition to 40G slower than prior forecasts
    - Creates a window where new technology can provide the higher port speed at lower cost
    - Some portion of today's 10G KR & SFP+ users are likely to adopt 25Gb/s on the way to a higher speed

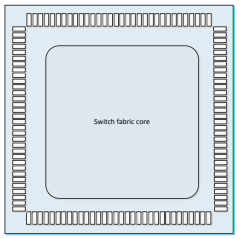
25Gb/s Ethernet Connectivity

- Enables similar topology as 40Gb/s & 10Gb/s
  - Single 25Gb/s SFP28 port implementation or Quad 25Gb/s QSFP28 breakout implementation possible
  - Maximizes ports and bandwidth in ToR switch faceplate
  - Dense rack server
  - Within rack, less than 3m typical length



25Gb/s I/O Efficiency

- Switch ASIC Connectivity limited by serdes I/O
- 25Gb/s lane maximizes bandwidth/pin and switch fabric capability vs. older generation
- Single Lane port maximizes server connectivity available in single ASIC
- 25Gb/s port optimizes both port count and total bandwidth for server interconnect



For a 128 lane switch:

Port Speed (Gbps)	Lane Speed	Lanes / port	Usable ports	Total BW (Gbps)
10	10	1	128	1280
25	25	1	128	3200
40	10	4	32	1280
40	20	2	64	2560
100	25	4	32	3200

Using 25Gb/s ports maximizes connectivity and bandwidth.

25 Gb/s Technology Feasibility

25 Gb/s Ethernet Technical Feasibility – Howard Frazier – Broadcom

Wealth of Prior Experience

Technology	Nomenclature	Description	Status
Backplanes	100GBASE-KR4 100GBASE-KP4	4 x 25 Gb/s (NRZ) 4 x 25 Gb/s (PAM-4)	IEEE Std 802.3bj™-2014 Ratified
Cu Twin-Axial	100GBASE-CR4	4 x 25 Gb/s	
Chip-to-Chip	CAUI-4	4 x 25 Gb/s	IEEE P802.3bm in Sponsor Ballot
Chip-to-Module	CAUI-4	4 x 25 Gb/s	
Module Form Factor	SFP28	1 x 25 Gb/s	Summary Document SFF-8402
	QSFP28	4 x 25 Gb/s	Style 1 - MDI for 100GBASE-CR4 Summary Document SFF-8665
	CFP2	4 x 25 Gb/s	
	CFP4	4 x 25 Gb/s	Style 2 MDI for 100GBASE-CR4

25Gb/s MAC/PCS Technical Feasibility

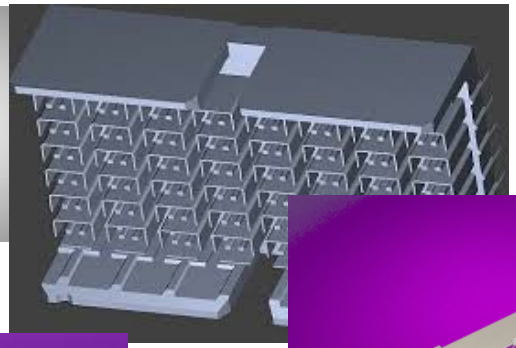
- The MAC is feasible in existing technology, and designs can leverage a 40GbE MAC and run it slower, or run a 10GbE MAC faster (with possibly a wider bus width)
- The PCS is feasible in existing technology, some possible PCS choices are:
  - Re-use the 10GbE PCS, 64B/66B, but run 2.5x faster (at possibly a wider bus width than a current 10GbE PCS). Can re-use the 10GBASE-KR FEC if desired and if it provides enough gain for possible PMDs
  - Re-use the 10GbE PCS and re-use the 802.3bj RS-FEC sublayer (both run at 25G), use transcoding to keep the same lane rate after adding the RS-FEC. Note the latency will be longer than it is for 100GbE.
- Possible data path widths in FPGAs: 64b @400MHz
  - Compact IP is possible, taking a small fraction of an FPGA
- Possible data path widths in ASICs: 32b @800MHz
  - Compact IP is possible
- Time-sliced MAC/PCS designs are feasible and can handle multi-rate implementations

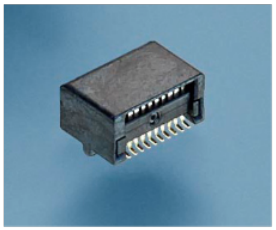
25Gb/s Single Lane Technical Feasibility

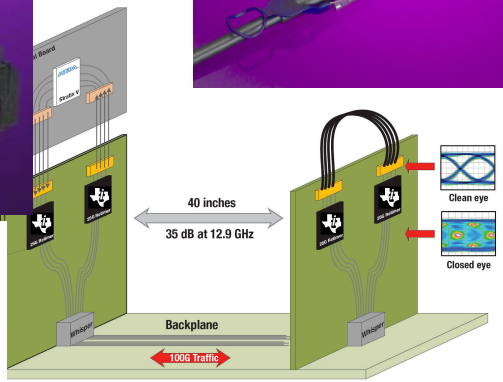
- SERDES Technology widely available
  - Under discussion among SERDES vendors since ~2002
  - OIF Project in July 2005
  - Several OIF CEI-25 and CEI-28 flavors in 2010/2011 time frame
  - Defined in IEEE P802.3bj as a 25Gb/s 4 lane electrical interface
  - Shipping ASIC cores for ~3 to 4 years
  - Defined channel models for circuit boards, direct attach cables, and connectors
- Technology re-use
  - Single-lane of 100GbE 4-lane PMD and CAUI-4 specifications
  - SFP28 being developed for 32G FC

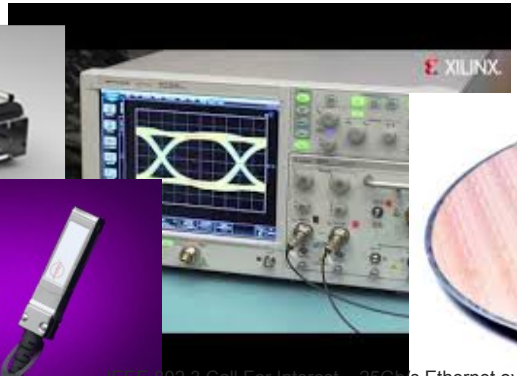
25Gb/s Technologies Readily Available

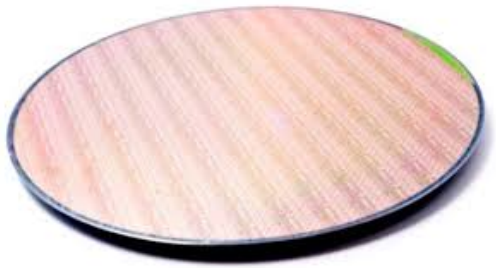




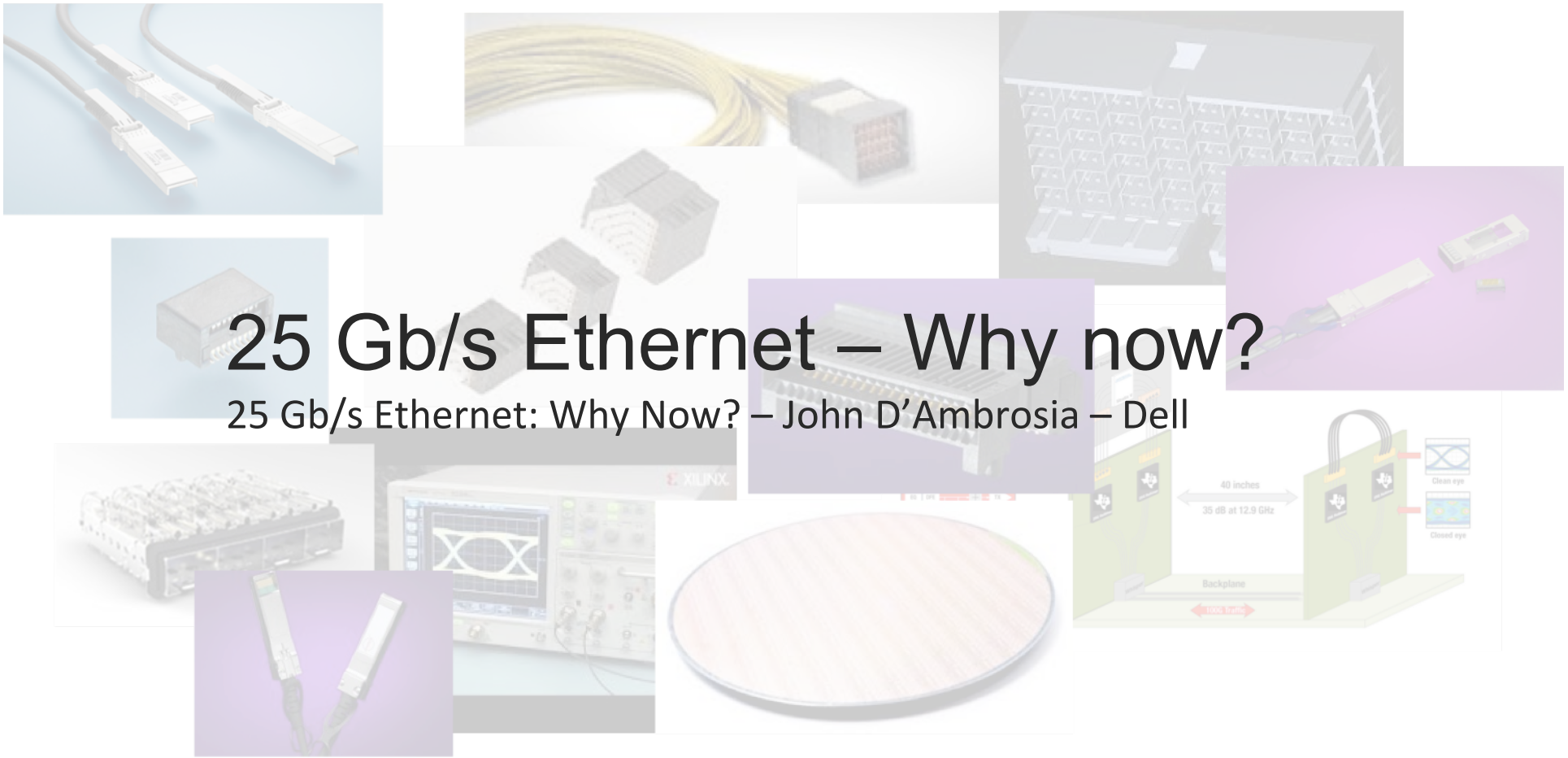




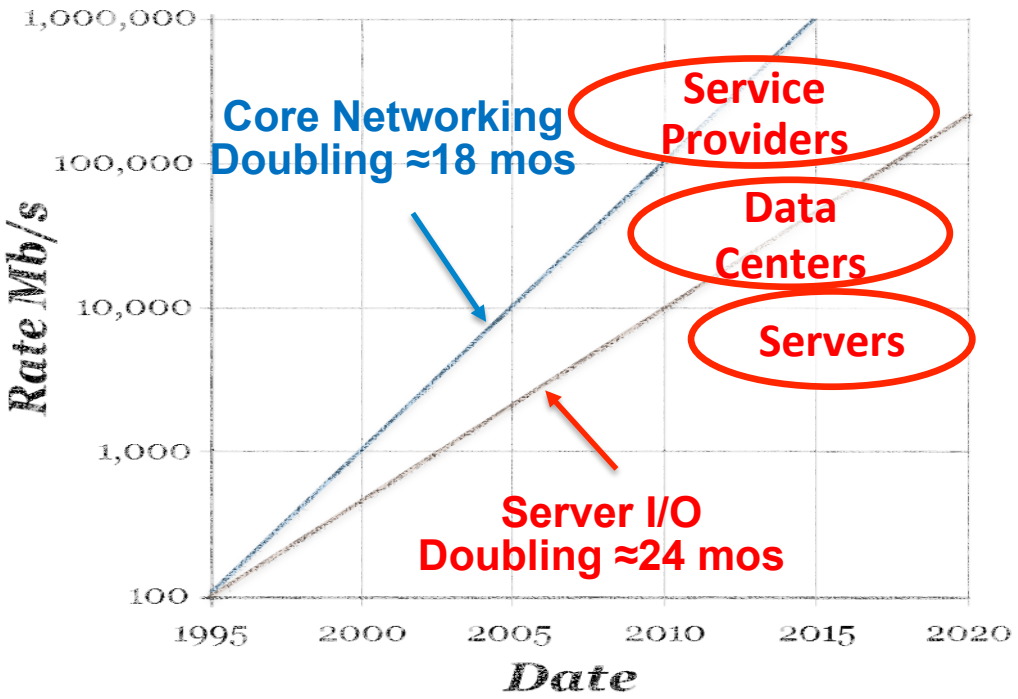

Provided by Amphenol, Molex, TE, Xilinx



25 Gb/s Ethernet – Why now?

25 Gb/s Ethernet: Why Now? – John D'Ambrosia – Dell

Before…..



Graphic Source: IEEE P802.3ba Tutorial, Nov 07

Crystal Balls aren't always clear

- 100GbE took off in service provider networks
- 40GbE took off in data centers
- Servers – slow transition to 10 GbE for some, but not for others

Consider Today's Cloud Scale Data Centers

	Top of Rack Box, Based on Single 128 I/O (3.2Tb) Silicon Switch Device					# TORs for a 100K Server Data Center
Server I/O	Oversubscription	Servers	100G Uplinks	Throughput (Tb/s) per ToR Switch	Utilization (%)	
40GbE (4x10G)	2.8:1	28	4	1.52	47.5	**3572**
40GbE (2x20G)	2.4:1	48	8	2.72	85	**2084**
25GbE Single Lane	3:1	96	8	3.2	100	**1042**

- Total Cost of Ownership – **Optimize cost per bit per second!**
  - CAPEX – Top of Rack Switches, Interconnect Structure
  - OPEX – Power / Cooling

Why Now?

- Web-scale data centers and cloud based services need
  - Servers with >10GbE capability
  - Cost sensitive for nearer-term deployment
- Industry has recognized the need & solution
  - Switching & PHY silicon under development
  - Formation of 25GbE Consortium targeting cloud-scale networks
- 25Gb/s technology standardized, developed, productized for 100GbE can be leveraged now!
  - There are no 40Gb/s single lane standardization efforts under way
- The Ethernet Ecosystem has been very successful
  - Open and common specifications
  - Ensured Interoperability
  - Security of development investment

Contributor Page

Hugh Barrass - Cisco
Brad Booth - Microsoft
Dave Chalupsky - Intel
John D'Ambrosia - Dell
Howard Frazier - Broadcom
Joel Goergen - Cisco
Mark Gustlin - Xilinx
Greg McSorley - Amphenol
Richard Mellitz - Intel
Mark Nowell - Cisco
Tom Palkert - Molex
Megha Shanbhag - TE
Scott Sommers - Molex
Nathan Tracy - TE

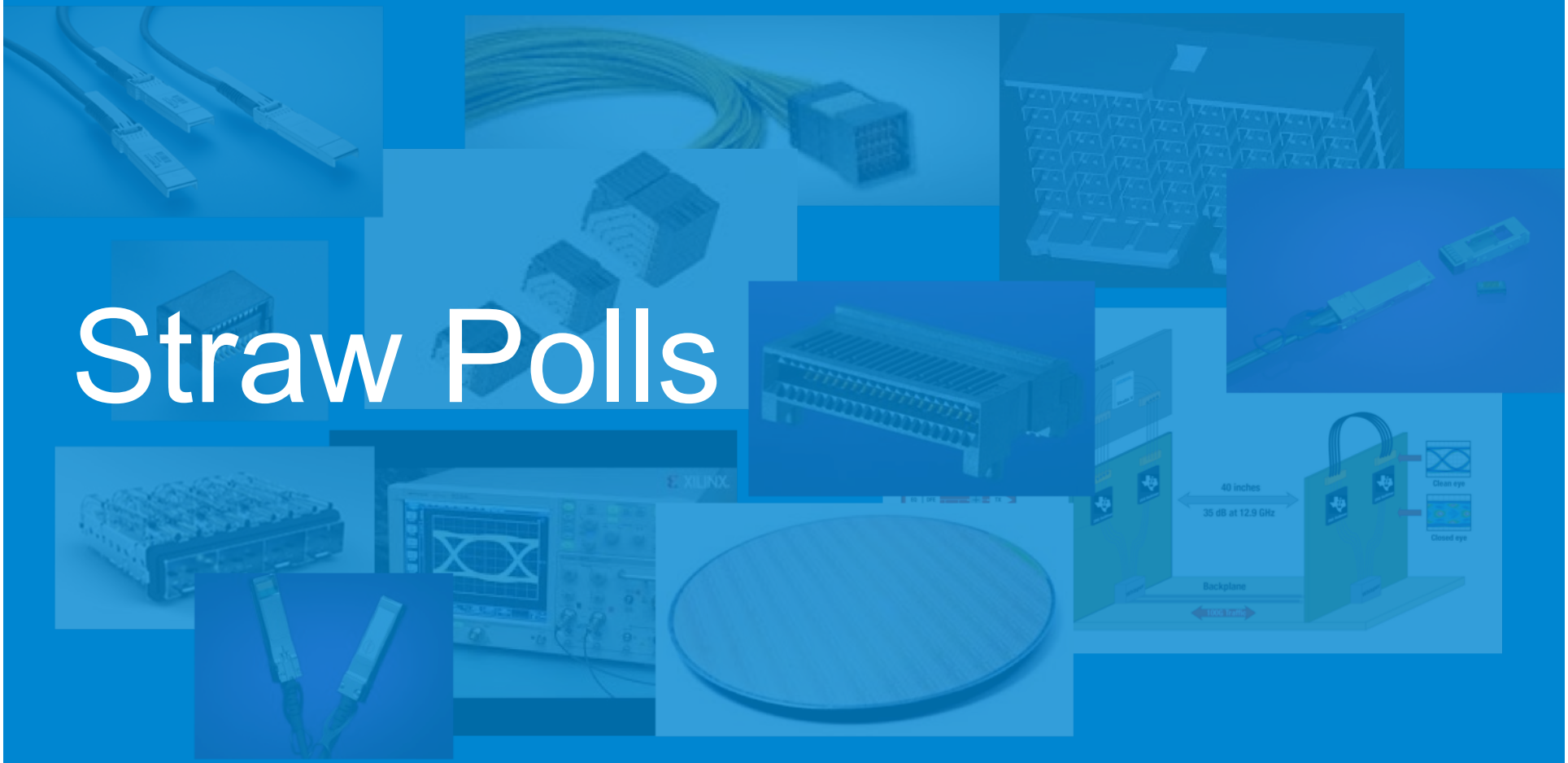
Supporters (Page 1 of 2) (87 individuals from 48 companies)

John Abbott - Corning
Venu Balasubramonian - Marvell
Thananya Baldwin - Ixia
Mike Bennet - 3MG Consulting
Vipul Bhatt - Inphi
Sudeep Bhoja - Inphi
Brad Booth - Microsoft
Bill Brennan - Credo Semicondictor
Matt Brown - Applied Micro
Dave Brown - Semtech
Mark Bugg - Molex, Inc
Carlos Calderon - Cortina-Systems
Dave Chalupsky - Intel
Chris Cole - Finisar
Chris Collins - Applied Micro
John D'Ambrosia - Dell
Mike Dudek - Qlogic
David Estes - Spirent Communications
Nathan Farrington - Packetcounter, Inc.
Bob Felderman - Google
Scott Feller - Cortina-Systems
Howard Frazier - Broadcom
Mike Furlong - ClariPhy
Mike Gardner - Molex, Inc
Ali Ghiasi - Ghiasi Quantum LLC
Joel Goergen - Cisco
Mark Gustlin - Xilinx
Steffen Hagene - TE
Dave Helster - TE
Yasuo Hidaka - Fujitsu Labs of America, Inc.
Kiyo Hiramoto - Oclaro Japan, Inc
Tom Issenhuth - Microsoft
Peter Jones - Cisco
Myles Kimmitt - Emulex
Scott Kipp - Brocade
Elizabeth Kochuparambil - Cisco
Paul Kolesar - CommScope
Subi Krishnamurthy - Dell
Ryan Latchman - Macom
Arthur Lee - MediaTek Inc.
David Lewis - JDSU
Mike Li - Altera
Kent Lusted - Intel
Jeffery Maki - Juniper
Arthur Marris - Cadence

Supporters (Page 2 of 2)

Beck Mason - JDSU
Erdem Matoglu - Amphenol
Greg McSorley - Amphenol
Richard Mellitz - Intel
Paul Mooney - Spirent
Andy Moorwood - Infinera
Ed Nakamoto - Spirent
Gary Nicholl - Cisco
Takeshi Nishimura - Yamaichi Electronics
Mark Nowell - Cisco
David Ofelt - Juniper
Tom Palkert - Luxtera
Vasu Parthasarathy - Broadcom
Neel Patel - ClariPhy
Pravin Patel - IBM
Jerry Pepper - Ixia
John Petrilla - Avago Technologies
Scott Powell - ClariPhy
Haoli Qian - Credo Semicondictor
Adee Ran - Intel
Ram Rao - Oclaro, Inc
Michael Ressl - Hitachi Cable America
Mike Rowlands - Molex, Inc
Anshul Sadana - Arista
Megha Shanbhag - TE
Kapil Shrikhande - Dell
Jeff Slavik - Avago Technologies
Scott Sommers - Molex, Inc
Steve Swanson - Corning
Norm Swenson - ClariPhy
Atsushi Takai - Oclaro Japan, Inc
Michael Johas Teener - Broadcom
Anthony Torza - Xilinx
Nathan Tracy - TE
Vincent Tseng - MediaTek Inc.
David Warren - HP
Brian Welch - Luxtera
Oded Wertheim - Mellanox
Chengbin Wu - ZTE
George Zimmerman - CME Consulting
Pavil Zivny – Tektronix
Adam Healey – Avago Technologies

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 “25 Gigabit/s Ethernet over a single lane for server interconnects”?
- Y: 121 N: 1 A: 14

- Room count: 148

Participation

- I would participate in a “25 Gigabit/s Ethernet over a single lane for server interconnects” study group in IEEE 802.3
  - Tally: 59

- My company would support participation in a “25 Gigabit/s Ethernet over a single lane for server interconnects” study group
  - Tally: 36

Future Work

- Ask 802.3 at Thursday's closing meeting to form a "25 Gigabit/s Ethernet over a single lane for server interconnects" study group
- Prepare ITU liaison letter for WG approval if Study Group formation is approved by WG.
- If approved:
  - 802 EC informed on Friday of formation of the study group
  - First study group meeting would be during Sept 2014 IEEE 802.3 interim meeting

Thank you!

From:
Joel Goergen
Mark Nowell
Dave Chalupsky
Brad Booth
John D'Ambrosia
Howard Frazier