400Gb/s Ethernet CFI Proposal

IEEE 802.3 Industry Connections
Higher Speed Ethernet Consensus Ad Hoc

Geneva, Switzerland

23 September 2012

Contributors

- Ghani Abbas, Ericsson
- Pete Anslow, Ciena
- David Chalupsky, Intel
- Chris Cole, Finisar
- Kai Cui, Huawei
- John D’Ambrosia*, Dell
- Dan Dove, APM
- Ali Ghiasi, Broadcom
- Mark Gustlin, Xilinx
- Mike Peng Li, Altera
- Jeff Maki, Juniper
- Andy Moorwood, Infinera
- Gary Nicholl, Cisco
- Mark Nowell, Cisco
- David Ofelt, Juniper
- Peter Stassar, Huawei
- Matt Traverso, Cisco
- Steve Trowbridge, ALU
- Brad Turner, Juniper

* Contributed to the content of this presentation but has taken no position with respect to supporting the 400Gb/s Ethernet CFI Proposal

Supporters

End Users

- Ralf-Peter Braun, DT
- Martin Carroll, Verizon
- Lu Huang, China Mobile
- Tom Issenhuth, Microsoft
- Junjie Li, China Telecom
- Sam Sambasivan, ATT
- Masahito Tomizawa, NTT
- Cheng Weiqiang, China Mobile

Test & Measurement OEMs

- Thananya Baldwin, Ixia
- Paul Brooks, JDSU
- Ed Nakamoto, Spirent
- Yukiharu Ogawa, Anritsu
- Jerry Pepper, Ixia
- Sergio Prestipino, Exfo
- Steve Sekel, Agilent
- Pavel Zivny, Tektronix

System OEMs

- Andreas Bechtolsheim, Arista
- Martin Bouda, Fujitsu
- Zeljko Bulut, NSN
- Cornelius Cremer, NSN
- Jörg-Peter Elbers, Adva
- Katsumi Fukumitsu, Fujitsu
- Rob Hayes, Intel
- Scott Kipp, Brocade
- Masashi Kono, Hitachi
- John McDonough, NEC
- Mounir Meghelli, IBM
- Pravin Patel, IBM
- Petar Pepeljugoski, IBM
- Rick Rabinovich, ALU
- Oren Sela, Mellanox
- Ted Sprague, Infinera
- Hidehiro Toyoda, Hitachi
- David Warren, HP

Supporters, cont.

System OEMs, cont.

- Chengbin Wu, ZTE
- Qingmin Zhang, Extreme

Optics Suppliers

- John Abbott, Corning
- Jon Anderson, Oclaro
- Chris Bergey, Luxtera
- Jens Fiedler, U2T
- Kiyo Hiramoto, Oclaro
- Hideki Isono, Fujitsu
- John Johnson, Cyoptics
- Jonathan King, Finisar
- David Lewis, JDSU
- Arlon Martin, Kotura
- Beck Mason, JDSU
- Tom Palkert, Molex
- John Petrilla, Avago
- Stefan Rochus, Cyoptics
- Steve Swanson, Corning
- Nathan Tracy, TE Connectivity
- Eddie Tsumura, Sumitomo
- Ed Ulrichs, SourcePhotonics
- Winston Way, NeoPhotonics

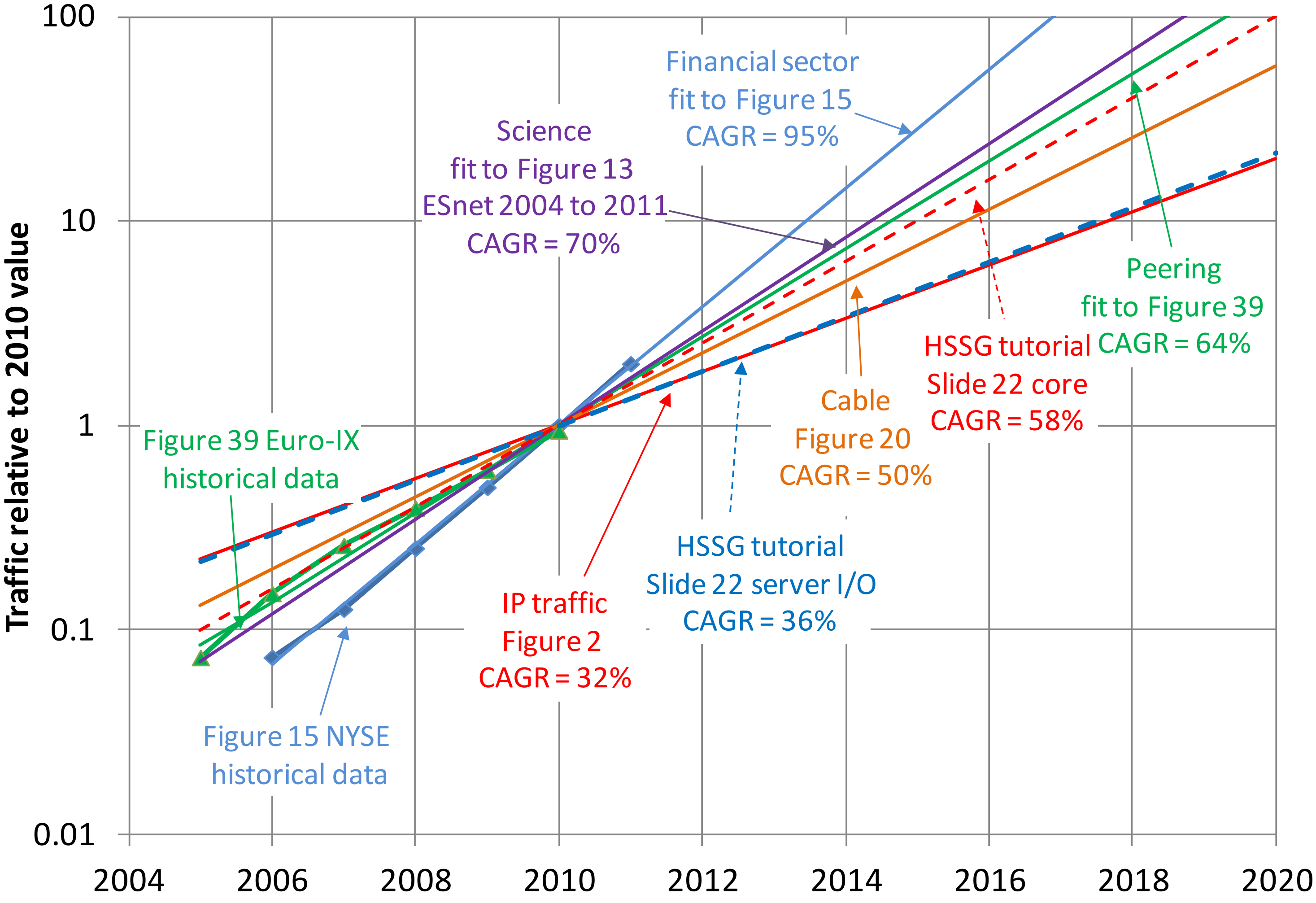
Suppliers

- Liav Ben Artsi, Marvell
- Sudeep Bhoja, InPhi
- Carlos Calderon, Cortina
- Frank Chang, Vitesse
- Ryan Latchman, Mindspeed
- Karl Muth, TI
- Greg McSorley, Amphenol
- Winston Mok, PMC-Sierra
- Venky Nagapudi, APM
- Takeshi Nishimura, Yamaichi
- Song Shang, SemTech

Outline

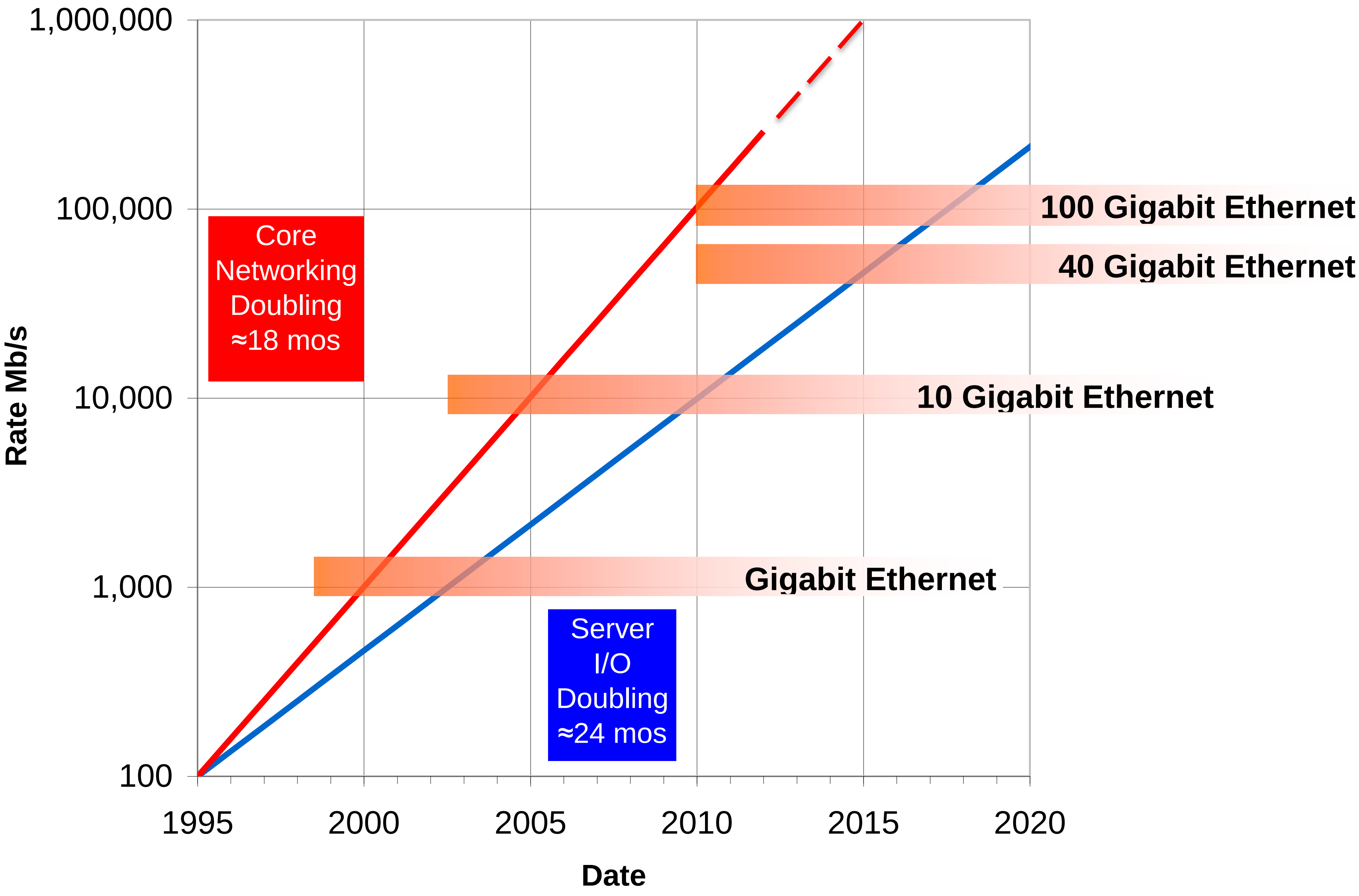
- Need for 400Gb/s Ethernet
- Near-term Applications
- Near-term Alternatives
- Near-term Technical Viability
- Straw Poll

Bandwidth Growth

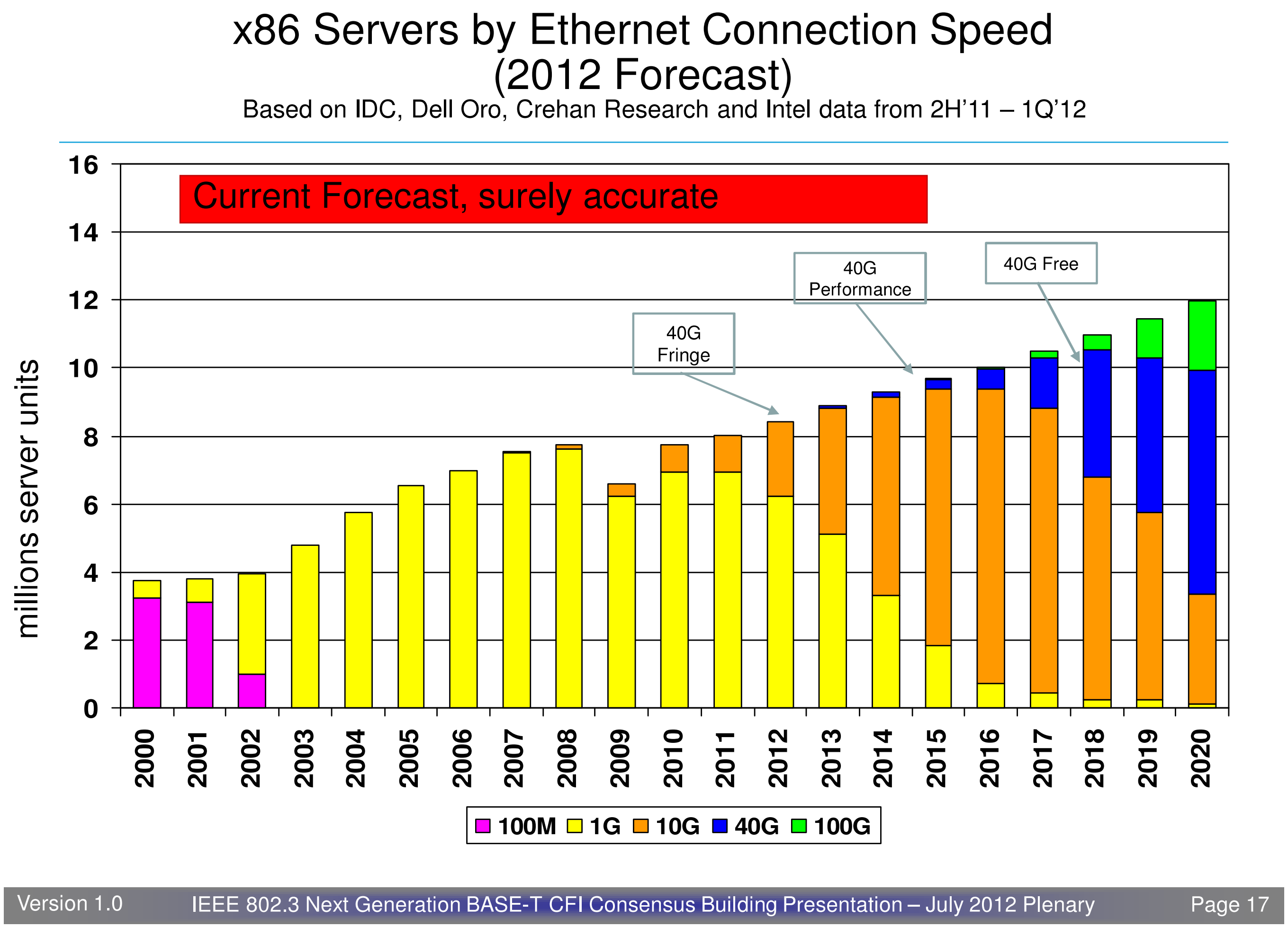


Bandwidth Assessment Ad-hoc (BWA) Summary

Computing & Networking Growth



Server Growth



x86 Servers by Ethernet Connection Speed
(2012 Forecast)

Based on IDC, Dell Oro, Crehan Research and Intel data from 2H'11 – 1Q'12

Current Forecast, surely accurate

40G Fringe

40G Performance

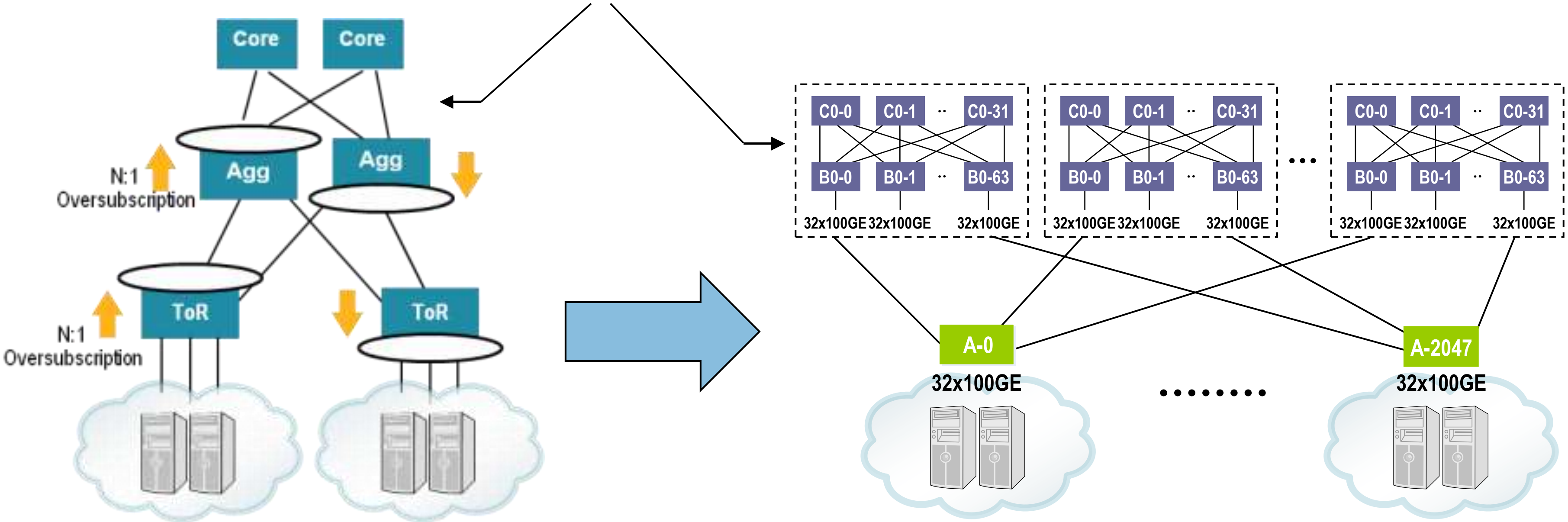
40G Free

Version 1.0 IEEE 802.3 Next Generation BASE-T CFI Consensus Building Presentation – July 2012 Plenary Page 17

Higher bandwidth uplinks needed as server ports transition to 10GbE, 40GbE & 100GbE this decade

Data Center Architecture Trend

400GbE need



Hierarchical Fat Tree architecture

Non-blocking architecture

400Gb/s Near-term Applications

- Core ⇔ Transport (400Gb/s Transport demonstrated)
- Core ⇔ Core
- Datacenter ⇔ Datacenter
- Datacenter upper layer switch interconnect
  (shown on previous slide)

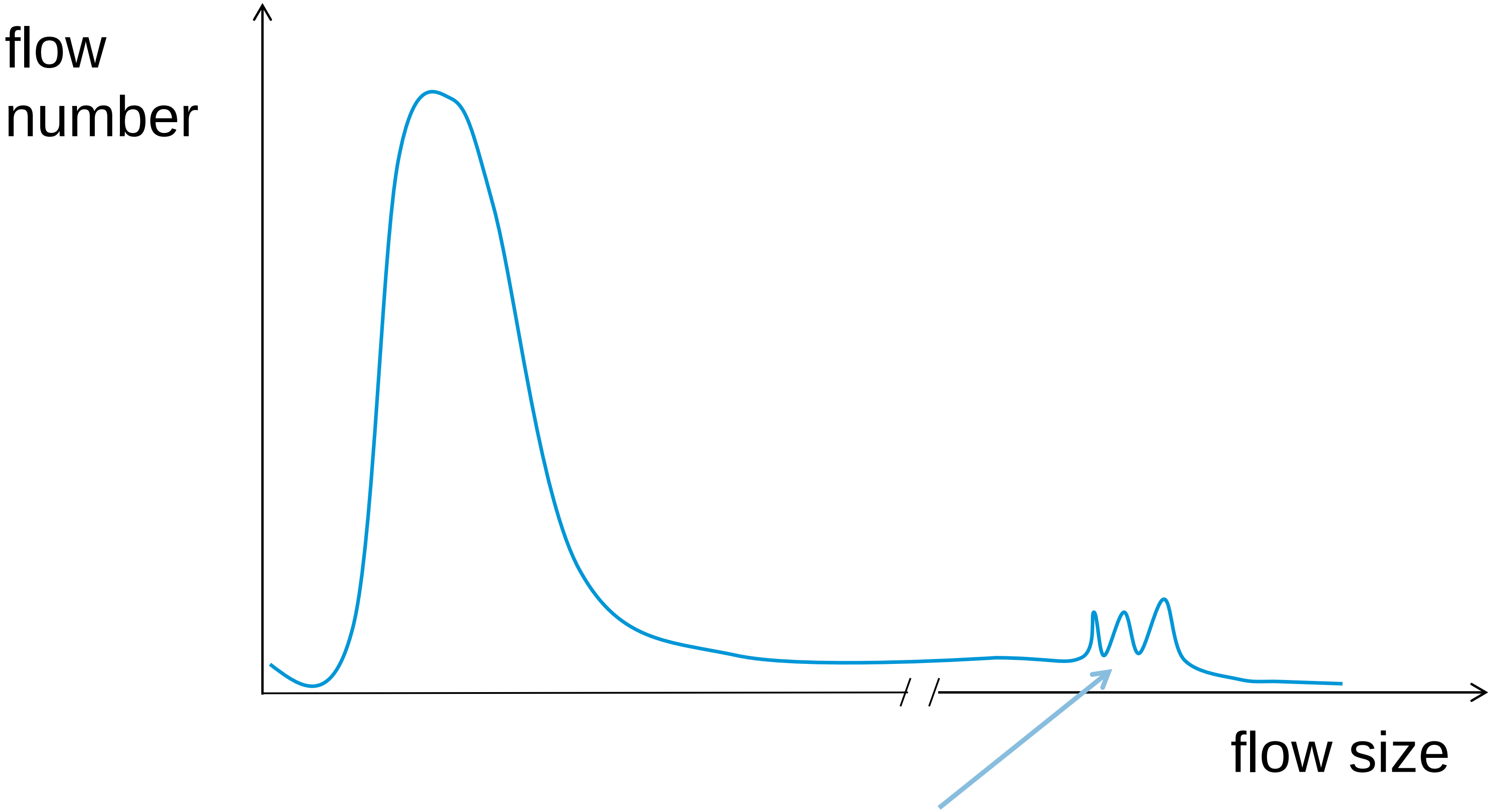
400Gb/s vs. Higher Rates

- Customers want parity in W/bit, $/bit, and bits/system
- Faster interface rates require exotic implementations
    - Not yet competitive per W, per $, or density
    - Higher R&D investment
    - Longer time to market
- 400GbE can reuse 100GbE building blocks
- 400GbE fits in the dense 100GbE system roadmap
- Data rates beyond 400Gb/s require an increasingly impractical number of lanes if 100GbE technology is reused

400Gb/s vs. 4 x 100Gb/s Link Aggregation

- Traffic is often trunked into large tunneled flows
  - Insufficient entropy to do hashing efficiently
  - Link Aggregation (LAG) is inefficient
  - BW not considered which leads to flow imbalance
  - A faster interface provides predictable performance
- Sources of large flows:
  - Content distribution
  - Secure traffic
- Fewer items to manage provides operational efficiency
  - Bandwidth is growing exponentially
  - Without faster links, link count grows exponentially therefore management pain grows exponentially

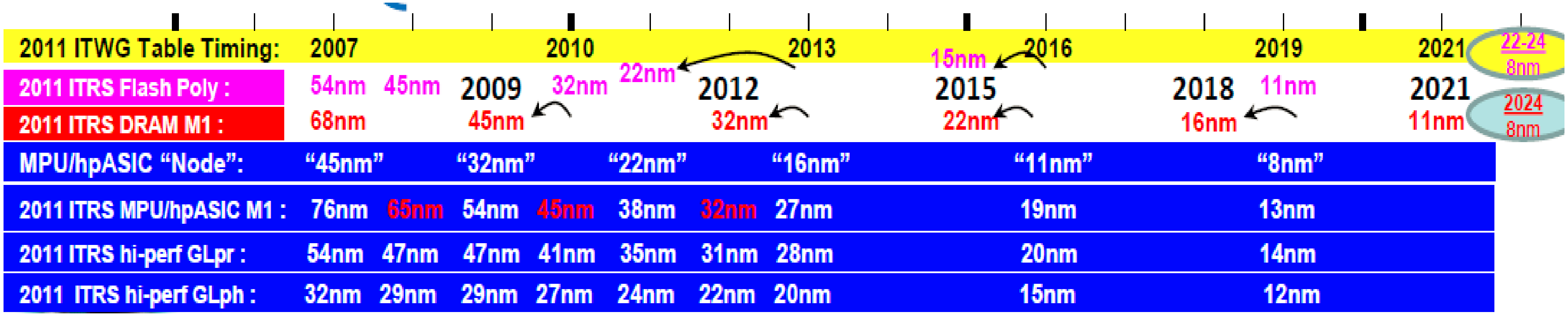
400Gb/s vs. 4 x 100Gb/s LAG, cont.



Large flows result in individual links becoming congested and bundles losing efficiency

400Gb/s MAC Technical Feasibility

- CMOS IC features have shrunk by ~2x since 100Gb/s MAC/PCS was defined in 802.3ba
- CMOS International Technology Roadmap for Semiconductors, 2011 Revision Overview:



2011 ITWG Table Timing:	2007			2010			2013			2016			2019		2021	
2011 ITRS Flash Poly :	54nm	45nm		32nm	22nm				15nm				11nm			22-24 8nm
			2009			2012			2015			2018			2021	
2011 ITRS DRAM M1 :	68nm		45nm			32nm			22nm			16nm			11nm	2024 8nm
MPU/hpASIC “Node”:	“45nm”		“32nm”		“22nm”		“16nm”			“11nm”			“8nm”			
2011 ITRS MPU/hpASIC M1 :	76nm	65nm	54nm	45nm	38nm	32nm	27nm			19nm			13nm			
2011 ITRS hi-perf GLpr :	54nm	47nm	47nm	41nm	35nm	31nm	28nm			20nm			14nm			
2011 ITRS hi-perf GLph :	32nm	29nm	29nm	27nm	24nm	22nm	20nm			15nm			12nm			

- ITRS Sponsoring Industry Associations (IAs): European Semiconductor IA, Japan Electronics and Information Technology Association, Korea Semiconductor IA, Taiwan Semiconductor IA, (US) Semiconductor IA

400Gb/s MAC Technical Feasibility, cont.

- Typical 100Gb/s MAC/PCS ASIC:
  - 45/40nm CMOS
  - 160b wide bus
  - 644MHz clock
- Potential 400Gb/s MAC/PCS ASIC:
  - 28/20nm CMOS
  - 400b wide bus
  - 1GHz clock
- 400Gb/s MAC/PCS FPGA will be feasible with wider buses and slower clocks

400Gb/s Study Group Topics

- Elements of 400Gb/s Study Group:
  - 400Gb/s MAC/PCS layer
  - Electrical I/O
  - SMF PMD
  - MMF PMD
- There is a strong desire to reuse 802.3ba, 802.3bj, and 802.3bm technology building blocks, which may include:
  - MAC/PCS architecture
  - FEC
  - CAUI-4
  - 100GBASE-LR4 or 100GBASE-nR4
  - 100GBASE-SR4

What Happens After This 400Gb/s Project

- Supported 400Gb/s apps. need lower cost PMDs
- Unsupported 400Gb/s apps. need new PMDs
- Bandwidth keeps growing (see BWA graph on page 6)
- As before, there will be follow-on projects
- Possible follow-on CFI(s) time frame: 3 to 6 years
- Possible follow-on Study Group Topics
  - New 400Gb/s PMD(s) to reduce lane count and cost
  - and/or next higher speed MAC/PCS and PMD(s) (ex.1: 1Tb/s, ex.2: 1.6Tb/s)

Straw Poll

Support the following data rate as the basis for near term CFI:

- 400Gb/s ____
- 1Tb/s ____
- 400Gbs and 1Tb/s ____
- Rate TBD in SG ____
- No CFI ____

400Gb/s Ethernet CFI Proposal

Thank you