The Case for 200GbE

John D’Ambrosia, Dell

July 14, 2015

Acknowledgement

- Thanks to the following individuals for providing content
  - Kapil Shrikhande, Dell
  - Scott Kipp, Brocade
  - Ali Ghiasi, Ghaisi Quantum
  - Chris Cole, Finisar

From Nowell Deck, Flash Mob Meeting, 5/15

What: Scope of CFI
50 GbE (plus adjacent interests?)

Clear Scope:

- Single lane 50 GbE, server IO

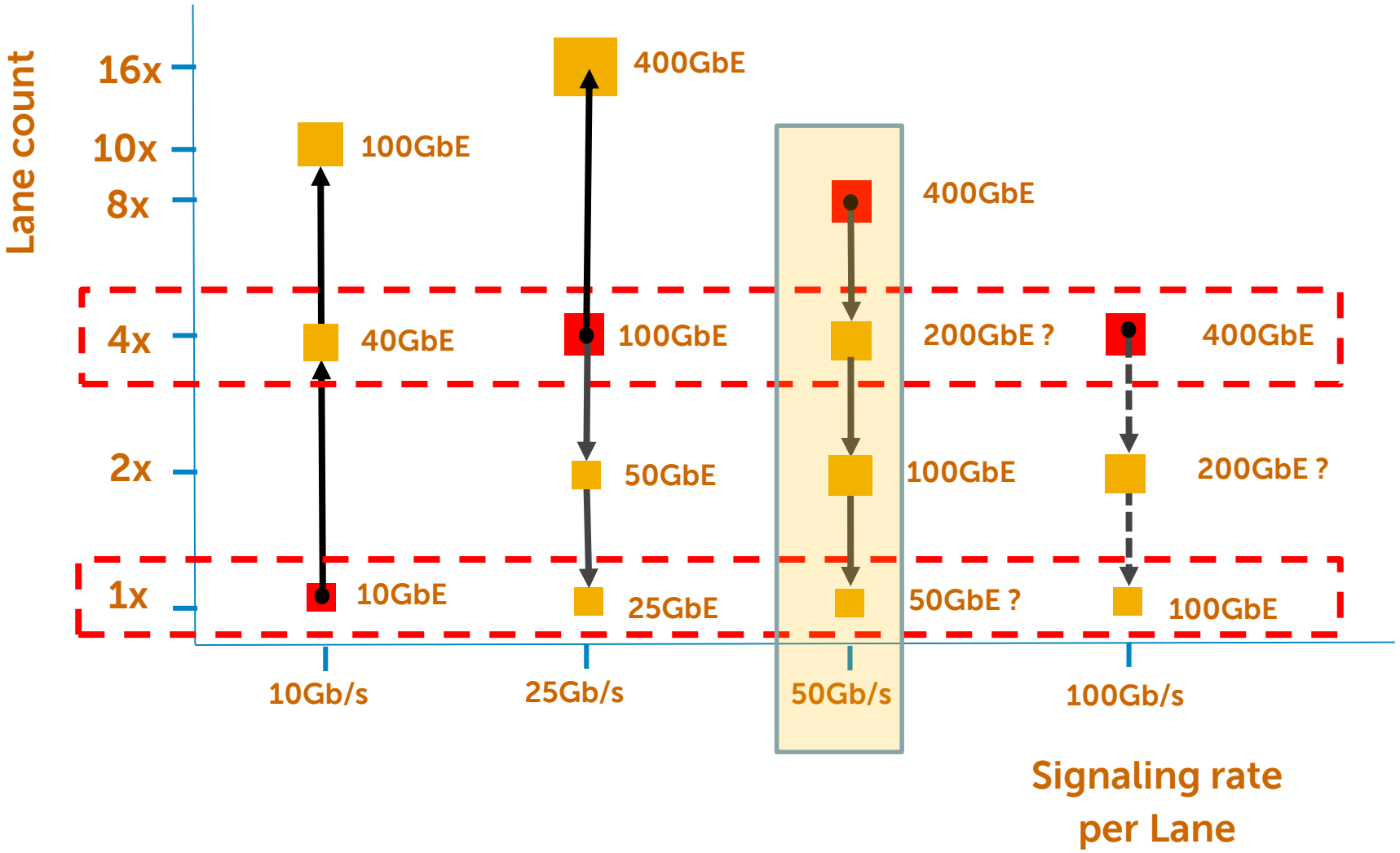
Potential adjacent interests that may incline some to want to broaden scope:

- Single lane 40 GbE – lower cost
- Other single lane options (25 GbE SMF, 100 GbE)
- 200 GbE
- Other nx50Gb/s options (100 GbE)

6

REASONS USED TO EXPLAIN 200 GbE

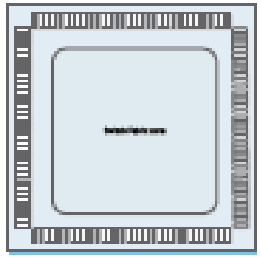
The New Rate Paradigm



Evolution using 50G SerDes

Next-gen switch ASIC

N x 50Gb/s SerDes chip



Radix

E.g. N = 128

128 x 40/50GbE

64 x 100GbE

32 x 200GbE

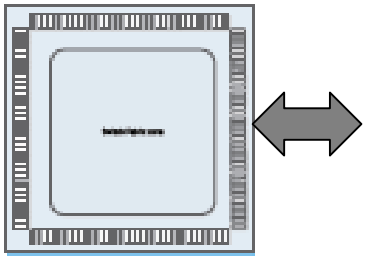
16 x 400GbE

Speed

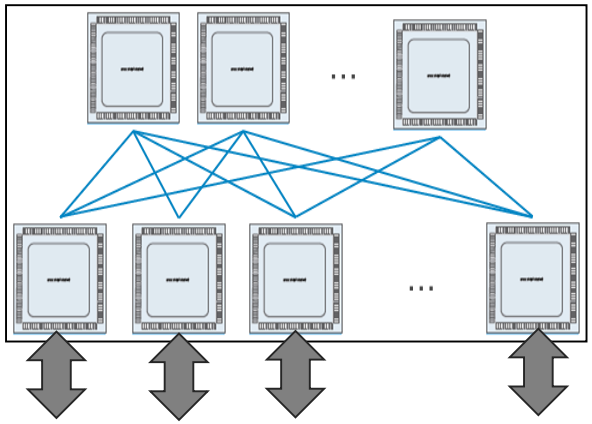
- 50GbE Server I/O
  - Single-lane Speed > 25GE
- 200GbE Network I/O
  - Four-lane Speed > 100GE
  - Balances Radix v. Speed
- 200GE on same fiber optic cables as 100GE possible
- 4x50GE breakout possible
- DC scalability – same as 25/100GE, 10/40GE designs

Ethernet ports using 50G SerDes

- 128x50Gb/s switch ASIC



128x50GbE
32x200GbE
16x400GbE



- E.g. TOR configuration
  - 96x50GE + 8x200GE

Large port count Spine switch

= N*N/2, where N is switch chip radix

N=32 → <= 512x200GE Spine switch

N=16 → <= 128x400GE Spine switch

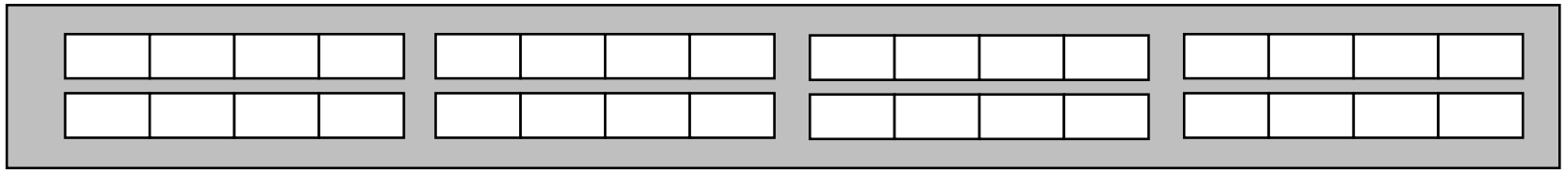
- High port count of 200GE better suited for DC scale-out

Progression in Speeds

1.28T Throughput

40GbE

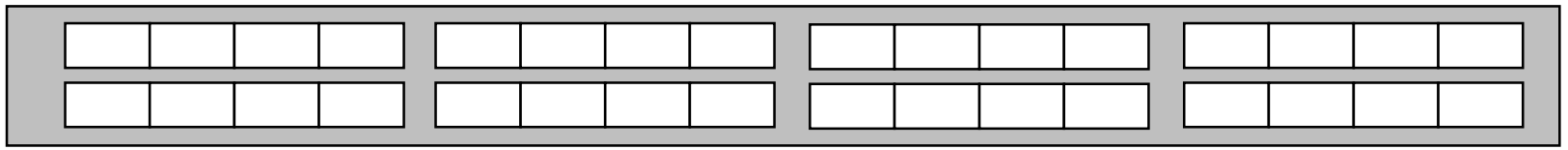
10G Lanes



3.2T Throughput

100GbE

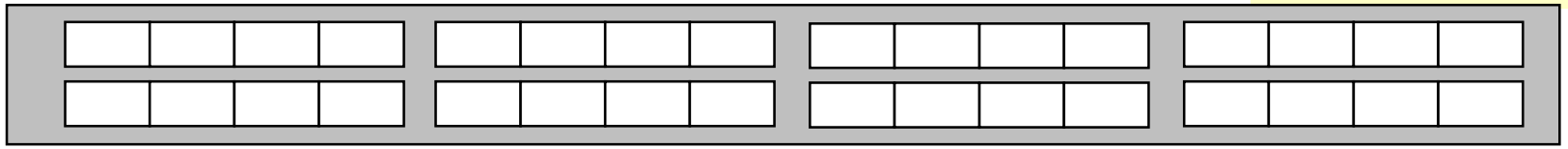
25G Lanes



6.4T Throughput

200GbE

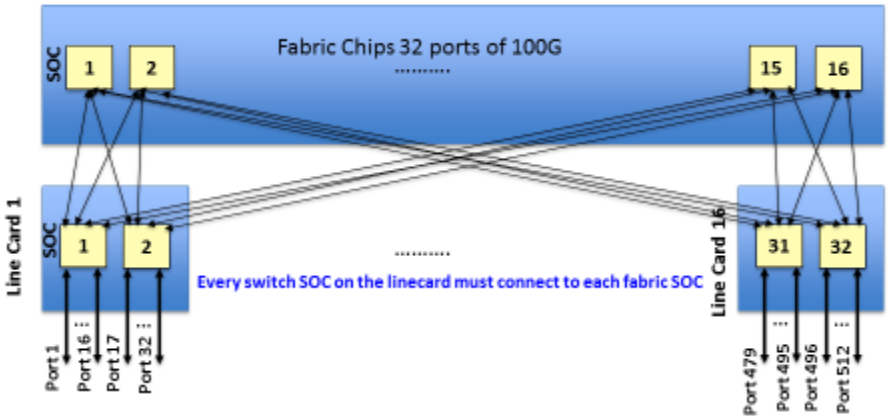
50G Lanes



Used with permission from Scott Kipp, Brocade

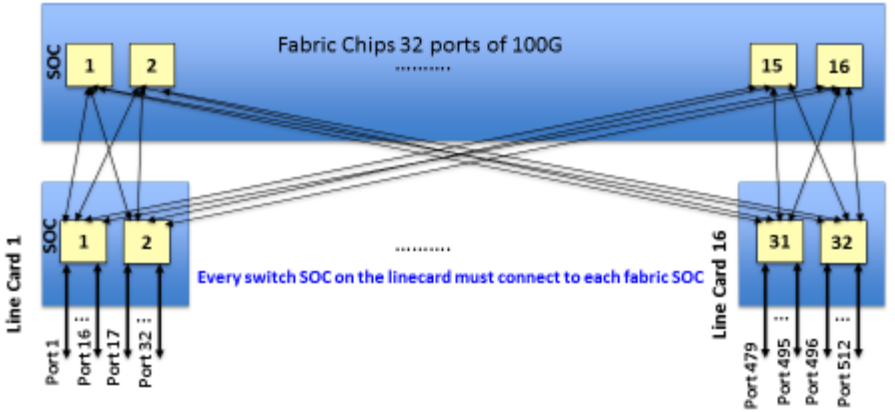
3 Stage Folded Clos with 100 GbE Supports Radix of 512

- **Building block 3.2 Tb/s SOC (32x100GbE/128x25GbE)**
  - 3 stage Clos with 51.2 Tb capacity



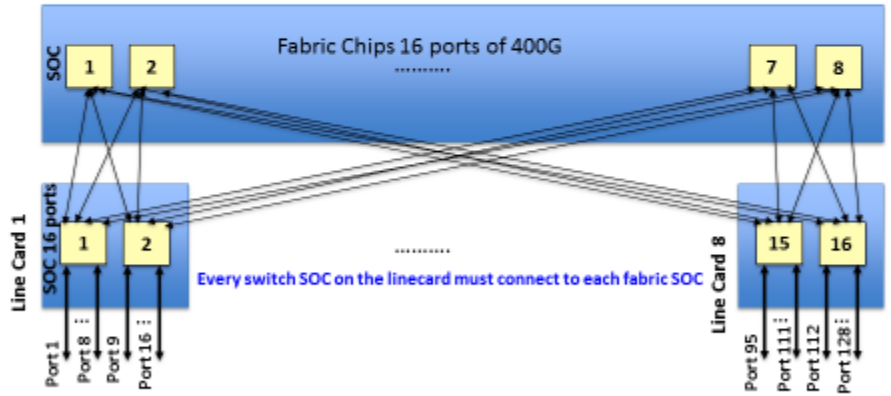
3 Stage Folded Clos with 200 GbE Supports Radix of 512

- **Building block 6.4 Tb/s Switch SOC (32x200GbE/128x50GbE)**
  - 3 stage Clos delivers 104.4 Tb twice as if one would uses 400G links



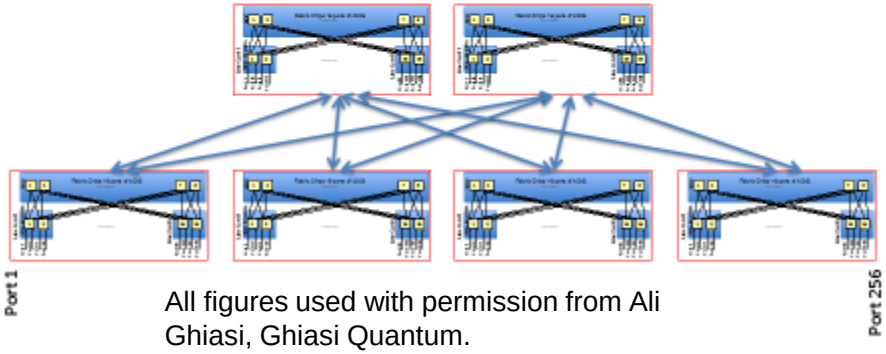
3 Stage Folded Clos with 400 GbE only Supports Radix of 128

- **Building block 6.4 Tb/s SOC (16x400GbE/128x50GbE)**
  - 3 stage Clos fabric has capacity of only 51.2 Tb!



400 GbE Requires 5 Stage Clos to have Identical Capacity as 200 GbE

- **Building block 6.4 Tb/s SOC (16x400GbE/128x50GbE)**
  - 400 GbE require 3x the switch SOC and 3x the interconnect to achieve 104.4 Tb capacity
  - Implementation is multi-chassis requiring large number of SR-16/PSMx links

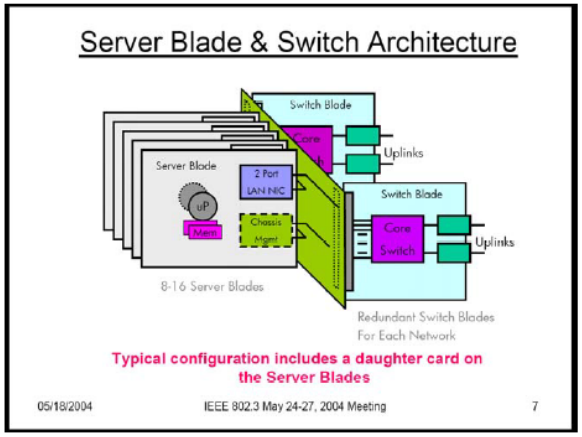


All figures used with permission from Ali Ghiasi, Ghiasi Quantum.

Backplane Applications

Backplane Ethernet & blade server architectures

- IEEE Std 802.3™-2008 defines GbE and 10GbE operation over a modular platform backplane (1 m objective)
  - 1000BASE-KX (GbE)
  - 10GBASE-KX4 (10 GbE, 4 x 3.125 GBd)
  - 10GBASE-KR (serial 10 GbE)
- Blade servers: 2nd Gen backplanes
  - Based on 10GBASE-KX4 architecture...
  - ...but satisfy 10GBASE-KR channel requirements
  - IEEE Std 802.3ba™-2010 introduced 40 Gb/s operation on backplanes: 40GBASE-KR4 (4 x 10.3125 GBd)
- Blade servers: 3rd Gen backplanes
  - Backwards compatibility needed
  - 100GbE must support 4 lane approach



Source: Koenen, "Channel Model Requirements for Ethernet Backplanes in Blade Servers", May 2004.

13

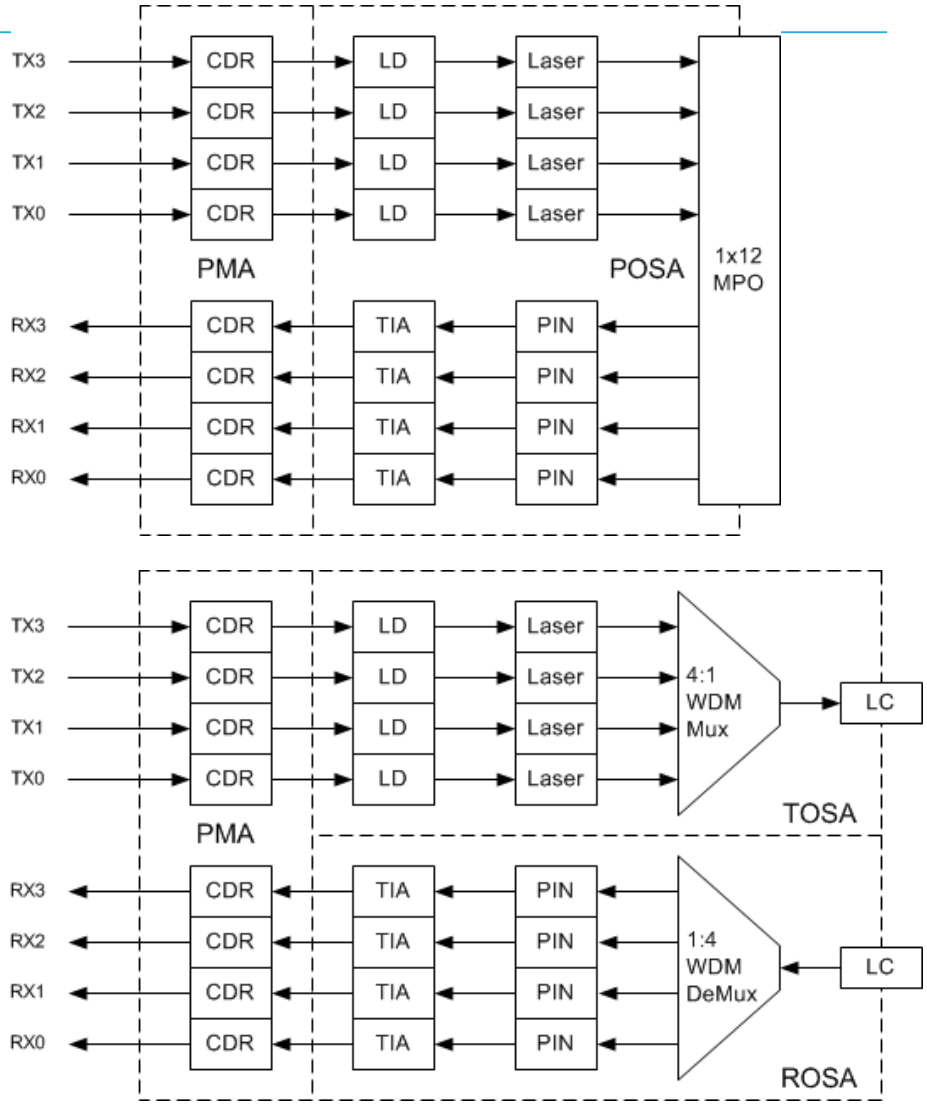
100GbE Electrical Backplane/Cu Cable CFI
IEEE 802 Plenary, Dallas, TX, Nov 2010

November 9, 2010

- 50G Serial would be used in backplane enclosures that would need to be backwards compatible with Backplane Ethernet Family, including10/40/100GBASE-KR4 (KP4)
- x4 architecture supports x1, x2, and x4

100G & 200G (MMF & SMF) QSFP Comparison

Block	100G	200G	New?
PMA IC	4x25G NRZ	4x50G PAM4	**Yes**
PMA Package	25G	25G	No
Laser Driver	Limiting	Linear	**Yes**
Laser	25G	25G	No
TOSA optics	4xWDM	4xWDM	No
TOSA package	25G	25G	No
PIN	25G	25G	No
TIA	Limiting	Linear	**Yes**
ROSA optics	4xWDM	4xWDM	No
ROSA package	25G	25G	No



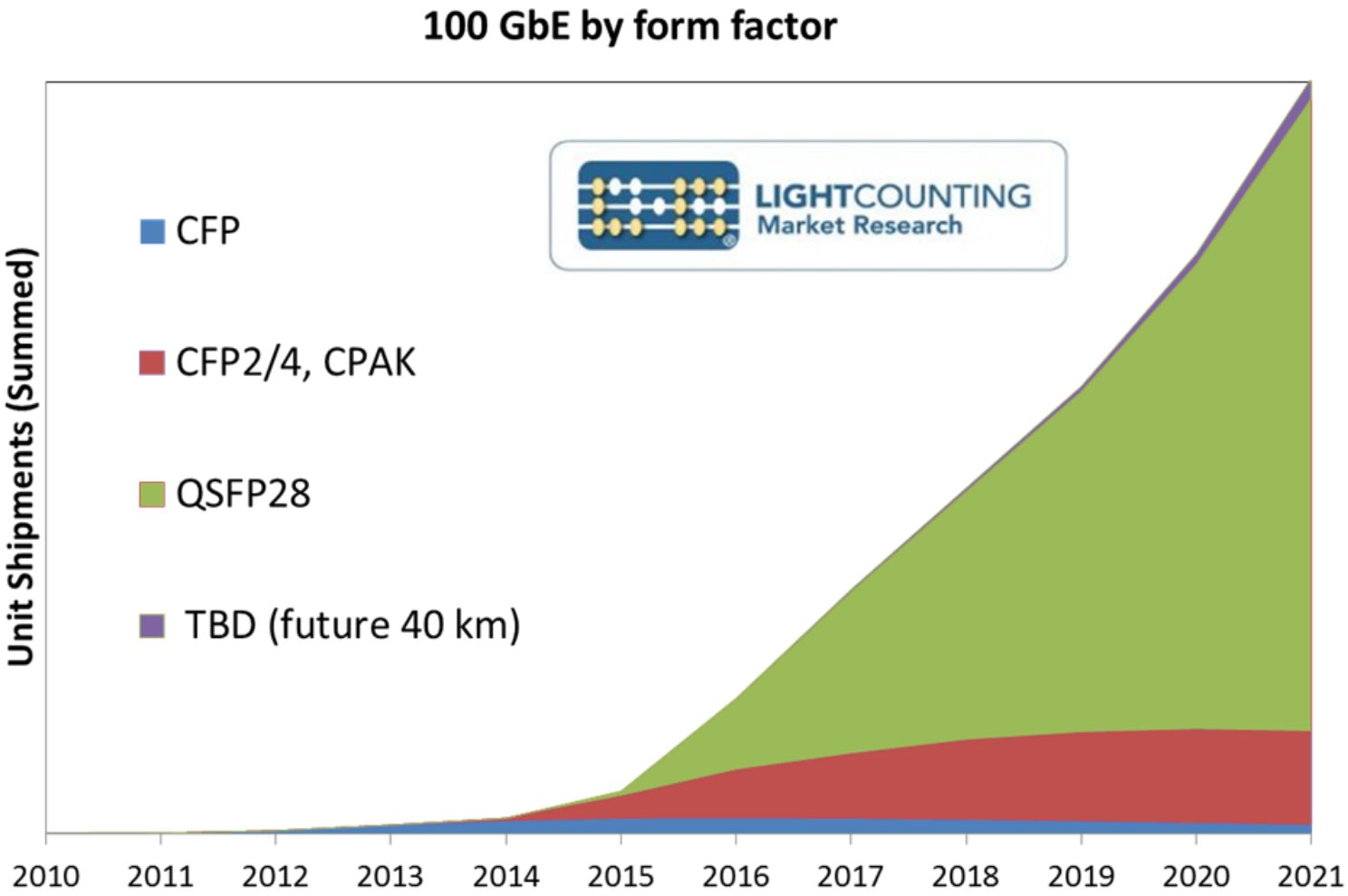
Provided by Chris Cole, Finisar

100G & 200G (MMF & SMF) QSFP Comparison

- The major change in going from 100G to 200G QSFP (MMF or SMF) is the ICs: PMA, Driver, and TIA.
- Cost of 4x50G PAM4 ICs will approach cost of 4x25G NRZ ICs over time, as a function of volume and process shrinks
- Over time cost delta between 100G and 200G optics will be driven by delta in optical margin between 25GBaud NRZ & PAM4
- 200G optics will benefit from the volume of 100G optics because the packing and optical components are the same
- The cost of Gb/s of 200G optics will eventually be lower than cost of 100G optics

Provided by Chris Cole, Finisar

QSFP as a Recurring Thought



Competing Industry Efforts

- IBTA Roadmap for 2017 (http://www.infinibandta.org/content/pages.php?pg=technology_overview)
  - 50GbE (HDR-Single Lane)
  - 200GbE (HDR – x4 lane)
- Fiber Channel Roadmap (T11 Spec / Mktg Availability)
- (http://fibrechannel.org/fibre-channel-roadmaps.html)
  - 64GFC 56.1G x 1 (2017 / 2019)
  - 256GFC 56.1G x 4 (2017 / 2019)

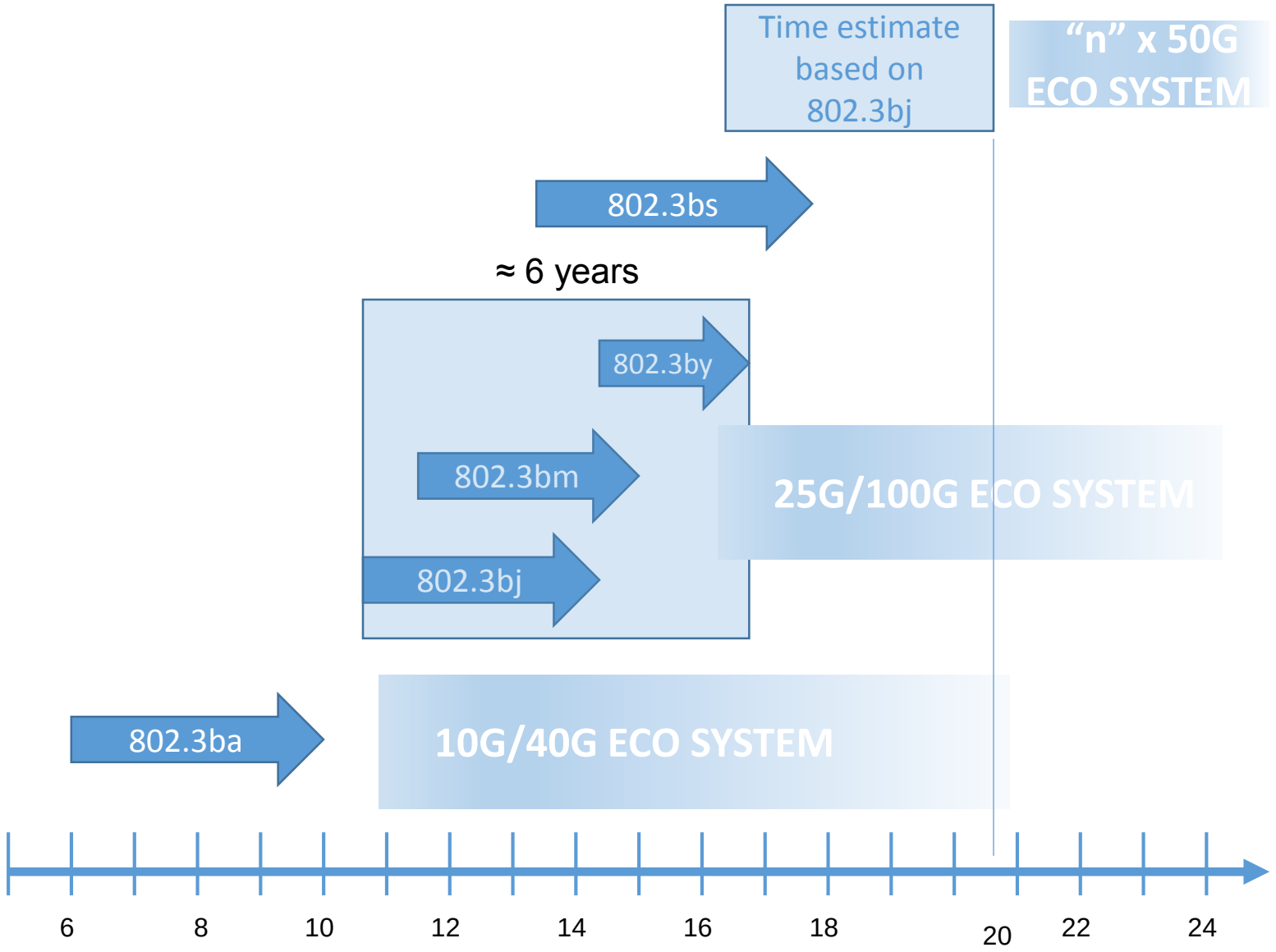
Considering the Road Ahead

IEEE 802.3 Ethernet Technology Overview (JD Assessment)

		Existing Rates									New Rates	
Media		10G	25G	40G	40G (G2)	100G (G1)	100G (G2)	100G (Gn)	400G (G1)	400G (G2)	50G	200G
PCB Traces		1x10G	1X25G	4X10G	1x40G	10x10G	4X25G	2x50G 1x100G ?	16X25G 8x50G		1x50G	4x50G
BP		1x10G	1X25G	4x10G	1x40G		4X25G				1x50G	4x50G
Cu Cable		1x10G	1X25G	4x10G	1x40G	10x10G	4X25G				1x50G	4x50G
MMF		1x10G	1X25G	4x10G	1x40G	10x10G	4X25G		16X25G	8x50G 4x100G ?	1x50G	4x50G
SMF	500m		1x25G		1x40G				4x100G (PAM4)		1x50G	4x50G
	2km		1x25G	1x40G	1x40G				TBD (8x50G WDM PAM4)	4x100G ?	1x50G	4x50G
	10km	1x10G	1x25G	4x10G WDM	1x40G	4x25G WDM			8x50G WDM (PAM4)	4x100G ?	1x50G	4x50G
	40km	1x10G		4x10G WDM		4x25G WDM						

Std or in progress | In Debate | Future

Time Frame Considerations (Rough Estimate)



Where do We Go?

- Enough examples to justify 200GbE's inclusion.
  - 200Gb solutions are happening, so 200GbE will happen
  - "200GbE Consortium" very plausible
  - For those of you who say no interest – remember 40G for networking?
- Lots of things to do
  - How we "pile on" will have an impact on the schedule?
  - How we "divide and conquer" will have an impact on schedule?
  - How do we do everything faster?
- Develop a "n" x 50G family
  - Initial focus – nAUI, backplane, twin-ax
  - Other initial focus for consideration – MMF?
  - SMF – initial work under way in 802.3bs