50G and 100G Use Cases

Brad Booth

Azure Networking, Microsoft

Supporters

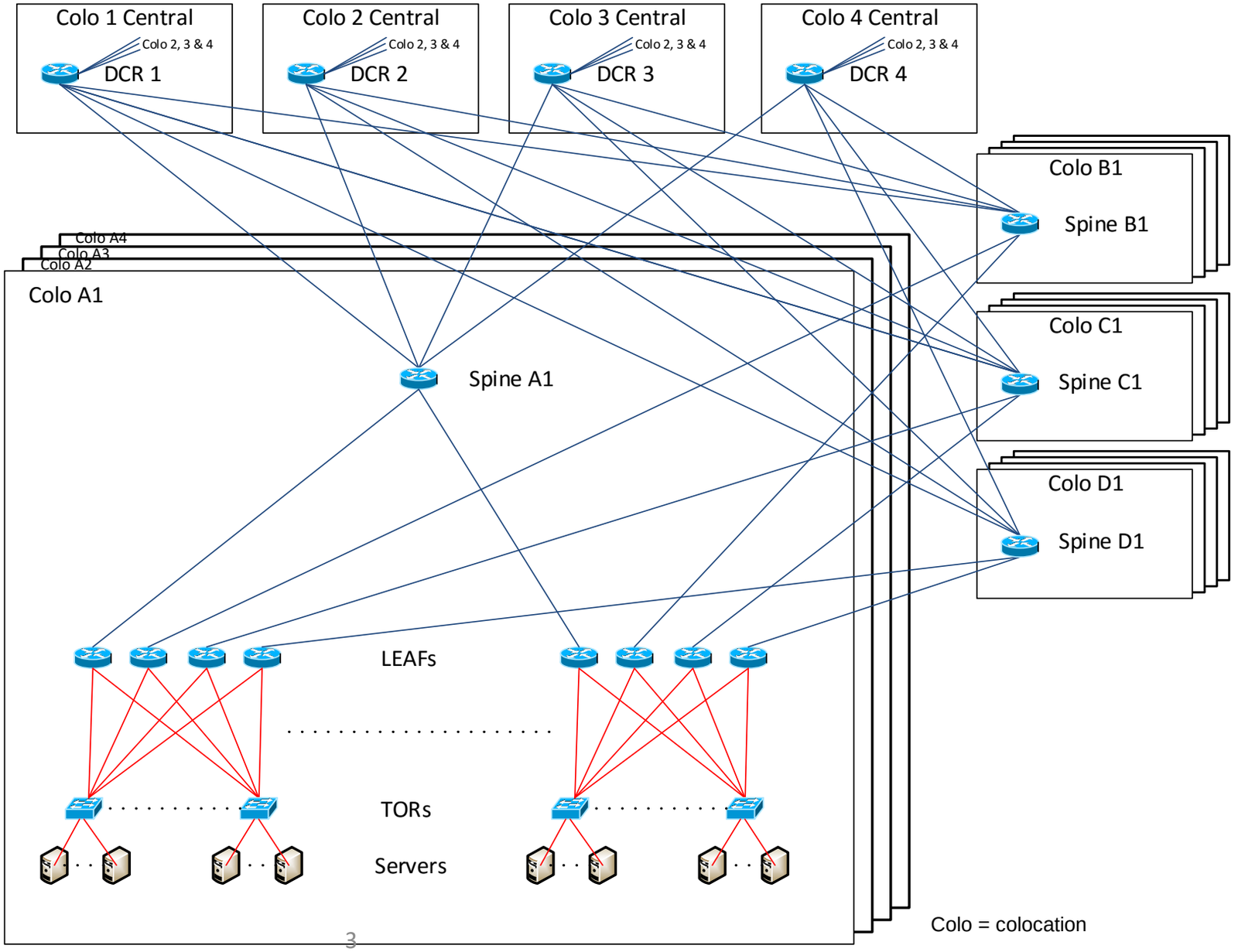
- Mike Andrewartha, Microsoft
- Phil Sun, Credo
- Tom Issenhuth, Microsoft
- Haoli Qian, Credo
- Jeff Twombly, Credo
- Zhenyu Liu, Credo
- Yasuo Hidaka, Fujitsu

Cloud Data Center Interconnection

Optical Modules

Active Optical Cables

Direct Attach Copper Cabling



100G Ecosystem Server to Tier 1 (Leaf)

- DAC from server to Tier 0 (TOR)
  - 3 meters max
  - Longer reach requires FEC
  - FEC impacts system design and latency
  - 802.3by provides a no-FEC option
- AOC from Tier 0 to Tier 1 (EOR or MOR)
  - 20 meters max
  - Relies upon CAUI-4 specification
  - Medium is irrelevant
  - Latency not as critical

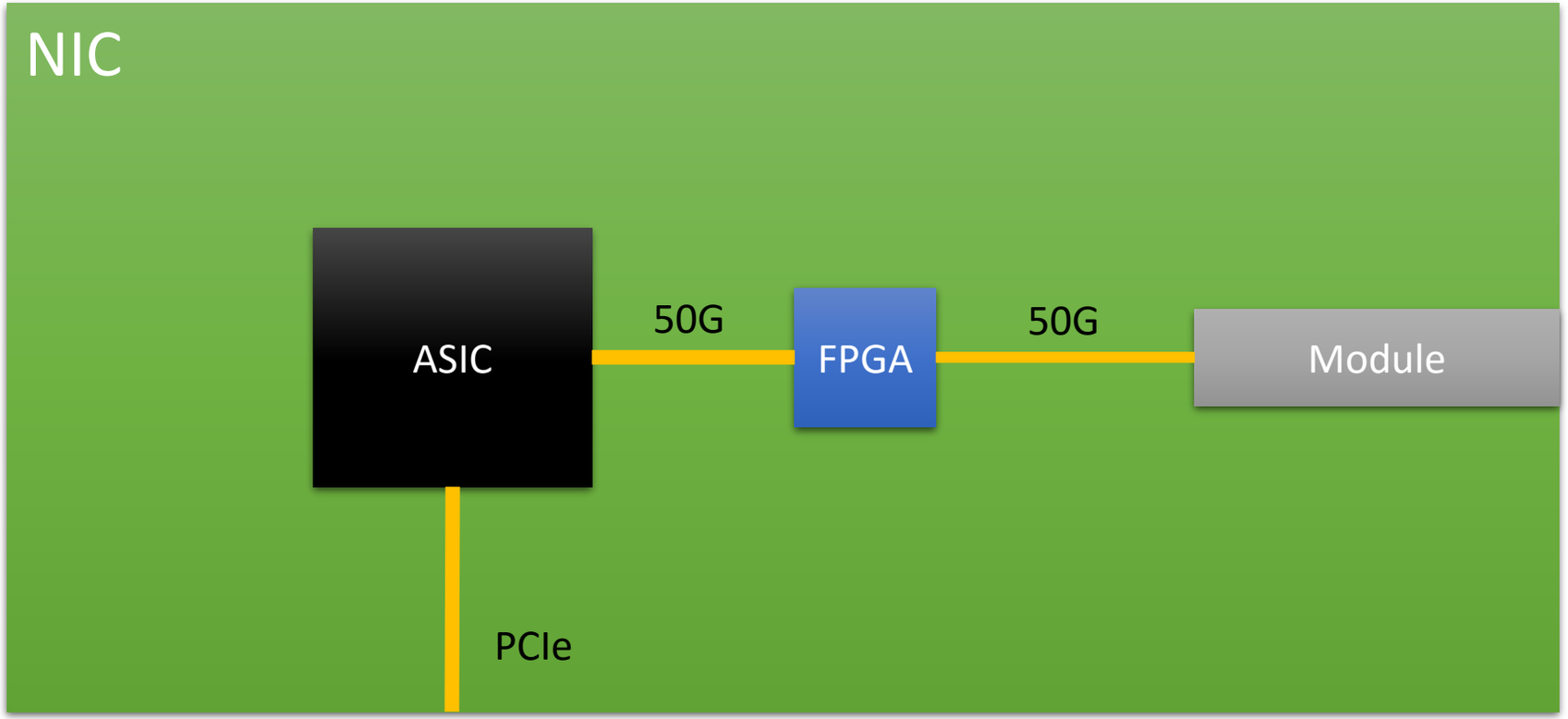
400G Ecosystem Server to Tier 1

- DAC from server to Tier 0
  - 2 meters maximum
  - Would like end-to-end FEC latency to be under 200 ns
  - Existing 50G PAM4 in 802.3bs requires KP4 FEC
    - About 170-200 ns per hop (sun_030216_50GE_NGOATH_adhoc.pdf)
- AOC from Tier 0 to Tier 1
  - Still 20 m maximum reach
  - Will rely upon the CDAUI-8 specification
  - FEC is more tolerable
  - Medium is still irrelevant

Latency Sensitivity

- Fungibility
  - Ability to use servers for any application
  - Storage servers may need to support SQL, Office 365, OneDrive, etc.
  - Compute servers may need to support HPC, Azure Compute, etc.
- Minimal latency permits broader application support
  - Reduces number of server SKUs
  - Provides greater economic feasibility
- Latency
  - Increased due to PAM4 modulation (compared to 10G, 25G)
  - Additional increase due to FEC requirement??

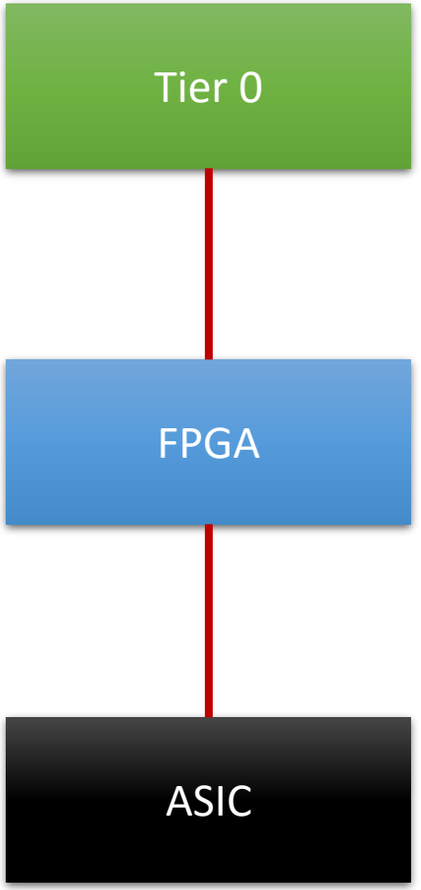
50G Bump In The Wire Example



research.microsoft.com/en-us/projects/catapult

How Latency Adds Up

- Assume 50G PAM4 interfaces w/ 200ns KP4 FEC latency
- From server to Tier 0
  - ASIC to FPGA = 200 ns
  - FPGA to Tier 0 = 200 ns
  - Return trip (Tier 0 to FPGA to ASIC) = 400 ns
  - Total FEC latency impact = 800 ns
- Need to have the latency under 200 ns server to Tier 0
  - 400 ns is a non-starter
  - Eliminating C2C FEC would help reduce to 200 ns
- Can we go lower?



Latency Is Important for BMP

Interconnect in the 400G Ecosystem

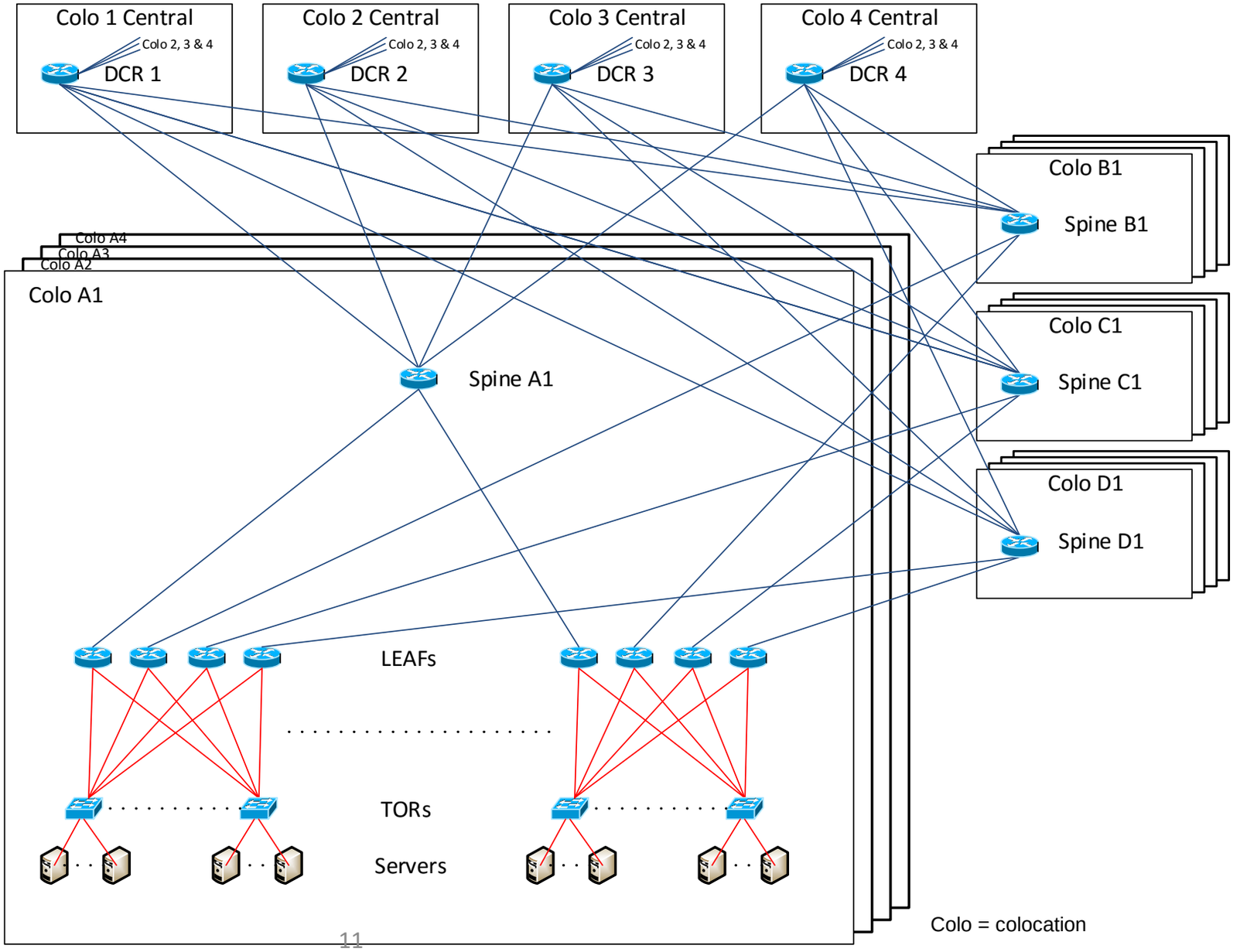
- Today's Server to Tier 0
  - Interconnect is based on 25G technology
  - Links are 50G Ethernet - 2x25G based on 25G Ethernet Consortium spec
    - Bandwidth growth drove us to use 50G
    - Don't require an 802.3 specification here
- Tomorrow's Server to Tier 0
  - Interconnect will be based upon 50G PAM4 technology
  - Expect links will be 100G Ethernet (2x50G)
  - Choice for 802.3:
    - Create the specification
    - Let a consortium do it

Cloud Data Center Interconnection

Optical Modules (400G)

Active Optical Cables 400G

Direct Attach Copper Cabling (100G)



Something to Note

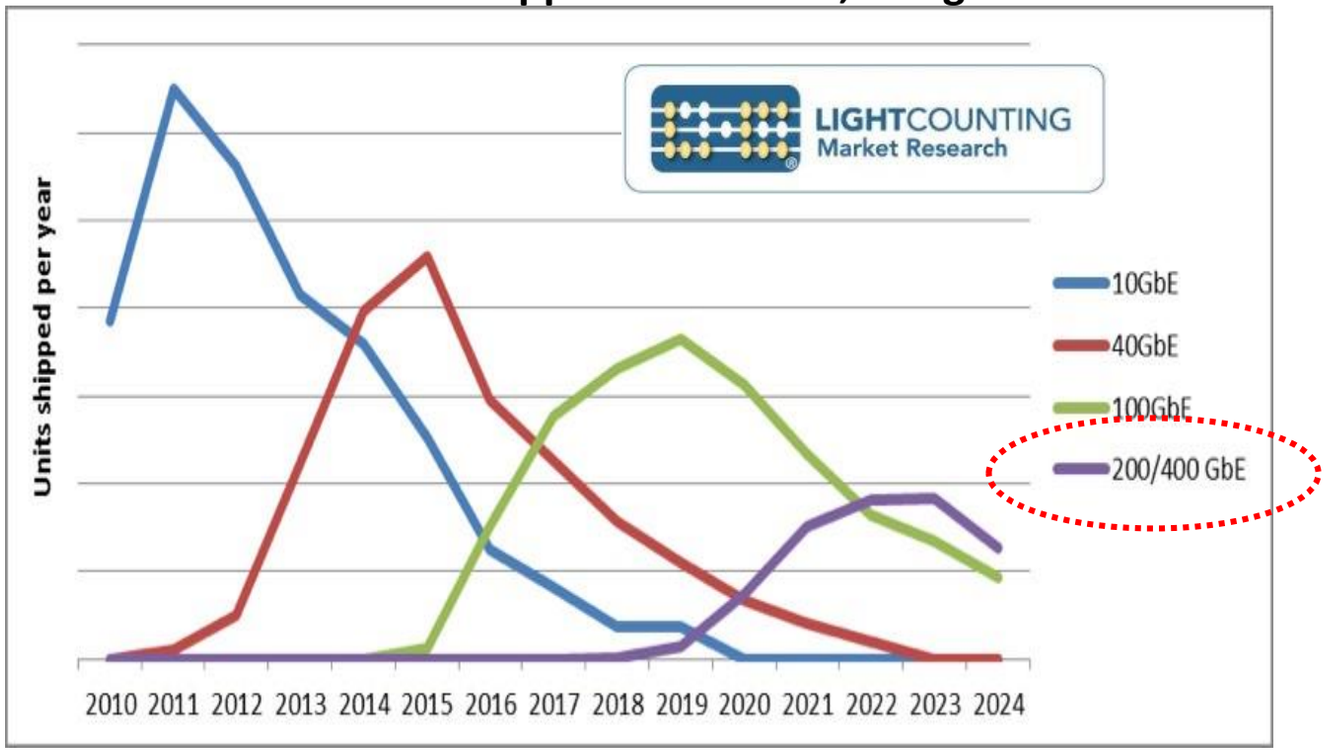
- Infrastructure supports parallel medium
  - True for today's 40G ecosystem
  - True for 100G ecosystem
  - True for 400G ecosystem
- Therefore!!
  - 50G will become the new "base" technology
  - Supplying specifications for all medium up to 500 m simplifies end user technology selection
  - Providing parallel derivatives for physical interconnect is goodness
- MAC rates don't need to equal PHY rates

NGTH (Next Gen Two Hundred)

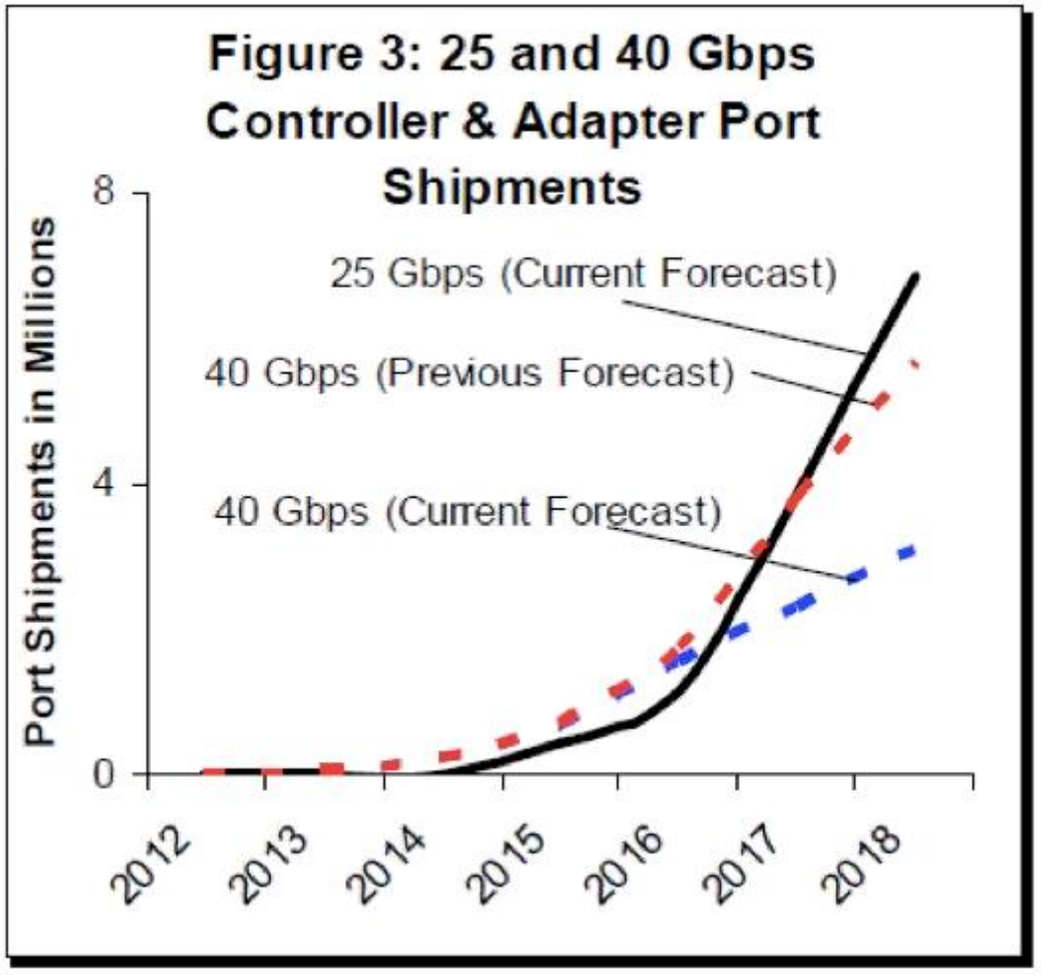
- Feedback received: 200G MAC to MAC provides no value
  - If switch radix is important, 100G (2x50G) is the better alternative
  - If bandwidth is important, 400G will win
  - 200G MAC to MAC may bifurcate the market
    - Is there impact to 400G BMP?
    - Is the investment in 200G worth it if end users really want 400G?
    - Impact to network architectures/OS? Will companies make the investment?
- 200G PHY, what is it really?
  - Is it 4x50G? Covered by a 50G specification
  - Is it 2x100G? Could be covered by 100G specification
  - Or is it just a packaging option (QSFP)??

NGTH – BMP Impact

Ethernet Transceivers shipped to Amazon, Google & Microsoft



LightCounting – Mega Data Center Optics (preview – Mar. '16)



Source: Dell'Oro Controller and Adapter Forecast Report -- July 2014

NGOH (Next Gen One Hundred)

- MAC is already done… network OS's understand it
- If 50G is the next base, then NGOH will be critical
  - Need more than just a copper cabling PMD
- Data point
  - An end user (not MSFT) is using the 50G (2x25G) from Leaf to Spine
  - Servers can support 100G worth of bandwidth
  - One and two lane of base technology is popular
- Gearboxes negatively impact TCO
- Need to support MMF and PSM variants for BMP

NGOH is an Opportunity

Summary

- 50G will become the next generation base technology
- Parallel variants are seeing broad deployment with 25G technology
  - Trend likely to continue with 50G
- Latency is critical
  - Can we eliminate or significantly reduce FEC's latency?
- NGOH is a market opportunity
  - Need broad medium support
- NGTH
  - Value and market impact have not been established

Recommended Objectives

- Support an optional end-to-end FEC latency of ≤ 100 ns
- Define a 2-lane 100 Gb/s PHY for operation over:
  - At least 2 meters on copper twin-axial cabling
  - Up to at least 100 meters on MMF
  - Up to at least 500 meters on parallel SMF
- Define a 50 Gb/s PHY for operation over up to at least 500 meters on SMF

Thank You