SNIA ESF | ETHERNET STORAGE

2017 Ethernet Roadmap for Networked Storage

Brad Booth, Microsoft
Vittal Balasubramanian, Dell
Brad Smith, Mellanox
Fred Zhang, Intel

December 1, 2016

SNIA Legal Notice

SNIA ESF | ETHERNET STORAGE

- The material contained in this presentation is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

About SNIA

SNIA ESF | ETHERNET STORAGE

SNIA-At-A-Glance

160
unique member companies

3,500
active contributing members

50,000
IT end users & storage pros worldwide

Learn more: **snia.org/technical** **@SNIA**

Today's Presenters

Brad Booth
Microsoft

Vittal Balasubramanian
Dell

Brad Smith
Mellanox

Fred Zhang
Intel

Agenda

- Who Needs Faster Ethernet?
- New Speeds
- Cable and module options
- Roadmap to 200Gb Ethernet
- Q&A

Why Do We Need Faster Ethernet?

- Faster storage—Flash and Persistent Memory
  - Up to 28Gb/s sequential read from one NVMe SSD
- New storage models
  - Cloud, scale-out, software-defined, hyper-converged
  - More speed & more replication, generally on Ethernet
- Video, gaming, mobile, Internet-of-Things
  - 4K/8K video capture and production
  - Video surveillance, mobile, streaming, social media

New Storage Models

- New Storage Models Displacing Fibre Channel SANs
  - Cloud: file, object, iSCSI, or distributed DAS
  - Software-defined, Big Data, Scale-out
  - Hyper-converged infrastructure, virtualization, containers
- Needs faster Ethernet networking
  - More east-west traffic
  - Faster servers, faster media
  - Converged network for storage and compute traffic

New Speeds

- 100 GbE: IEEE *802.3ba* defined in 2010
  - 10 lanes of 10Gb/s – first products in 2011
  - 4 lanes of 25Gb/s – first products in 2015
- 25/50 GbE: Consortium defined in 2014
  - First products in 2015
  - 1 or 2 lanes of 25Gb/s
- 25 GbE: IEEE *802.3by* defined June 2016
- 2.5/5 GbE: IEEE *802.3bz* defined Sept 2016
  - Twisted pair—2.5GBASE-T and 5GBASE-T
  - Speed upgrade for access (office/home) networks

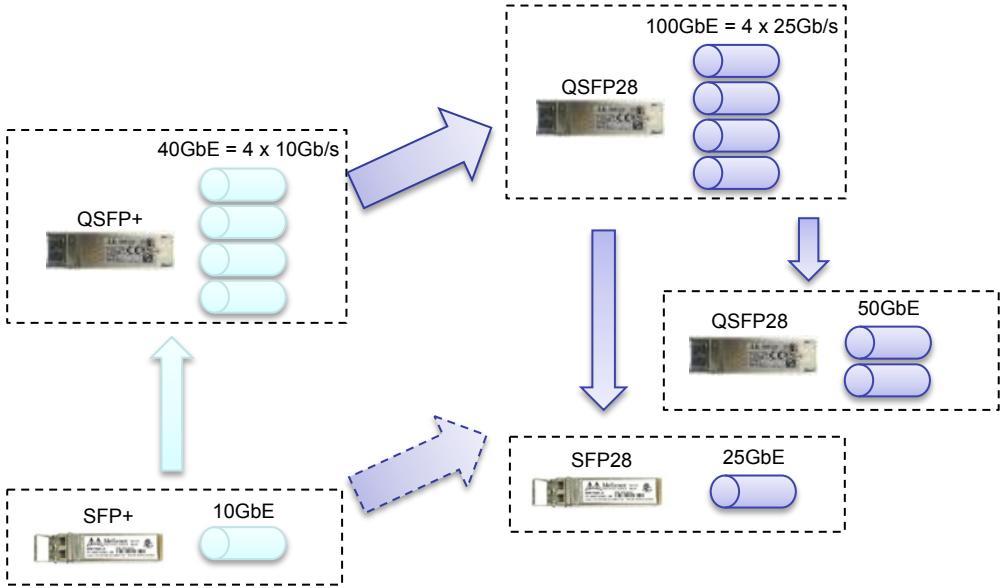
IEEE

25G Ethernet 50G Consortium

NBASE-T™ ALLIANCE

What Changed?

- Old: 10Gb/s per lane
  - 1/4 lanes = 10/40 GbE
  - 10 lanes = 100 GbE
- New: 25Gb/s per lane
  - 1/2 lanes = 25/50 GbE
  - 4 lanes = 100 GbE
- 2 wires or fibers / lane*
  - Copper or optical
  - Can re-use existing fiber



*WDM allows multiple lanes per fiber pair

Why Faster Lanes Are Good

- 25GbE vs. 10GbE
  - 2.5x BW at 1.5x the price
  - Compatible with 10GbE
- 50GbE vs. 40GbE
  - 1.25x BW at same price
  - Half the lanes/fibers
  - 2x switch port density
- 100GbE for switch links
  - 60% fewer uplinks

Same optical cable supports 10 and 25GbE

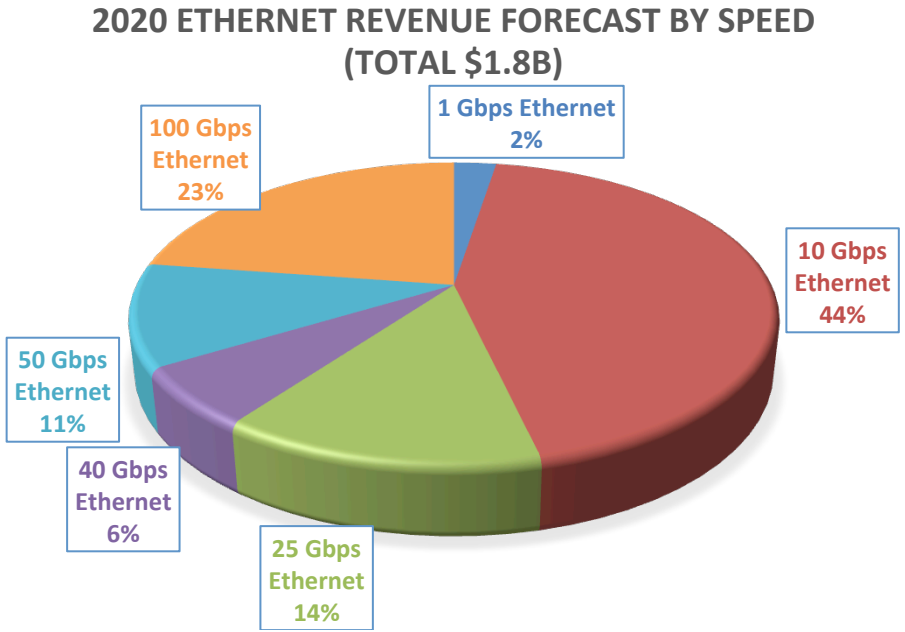
Typical new generation switch

- Up to 32 ports at 40/100GbE
- Up to 64 at 50GbE (with breakout cables)
- Up to 128 ports at 10/25GbE (with breakout cables)

25/50/100 GbE Adoption

- Rapid adoption of new speeds

98% of total Ethernet revenue will be from speeds >10GbE by 2020



July 2016, CREHAN RESEARCH, Long-range Forecast Server-class Adapter & LOM– Total Ethernet (including FCoE) Server-class Adapters & LOM Controllers

Two Paths To Faster Ethernet Storage—A Generalization

- Considerations for Cloud, SDS, Hyper converged, Scale-Out
  - Storage capacity looking to move up
    - 40GbE since 2012
    - 25/50GbE to endpoints
  - Upgrade 40GbE to 50GbE
  - 100GbE switch links
- Traditional Enterprise Arrays
  - Added 10GbE in 2012-2014
  - Adding 40GbE in 2016/2017
  - 25/50/100 GbE in 2017/2018
  - New switches support 10, 25, 40, 50, & 100

New Ethernet Speeds Summary

- New storage designs need faster Ethernet
- 25/40/50/100 GbE speeds available now
- Faster lanes = less cabling
  - Denser switches
  - More cost-effective networking
- Cloud moving 10/40 → 25/50/100, while enterprise moving 10GbE → 40GbE.
  - SDS, servers, & startups support new speeds
  - Very large enterprise moving → 25/50/100GbE

Cables and Transceivers

Brad Smith
Director of Marketing,
LinkX Interconnect Team, Mellanox
BradS@Mellanox.com
DAC, AOCs, Optical Transceivers, Ethernet & InfiniBand Networking

"Call the Cable Guy"

Connecting Storage

- Storage and compute need to be interconnected
- Cabling affects…
  - Cost, performance, reliability
  - Power consumption, rack density, upgrade paths
- Cables & transceiver costs in modern DCs are escalating
- Large installations cost approaching 40-50% of total CapEx
- NVME FLASH subsystems driving high speed interconnects
  - Only 3 NVME cards can consume a 100Gb/s link

3 Main Types of DC High-Speed Interconnects

SNIA ESF | ETHERNET STORAGE

Direct Attach Copper (DAC)

Copper Wires
Key feature = Lowest Priced Link
25/50/100GbE: 3m-5m reach

Copper Cables

Active Optical Cables

2 Transceivers w/optical fiber bonded inside
Key feature = Lowest Priced Optical Link
100m/200m Reaches

Transceivers with Integrated Fibers

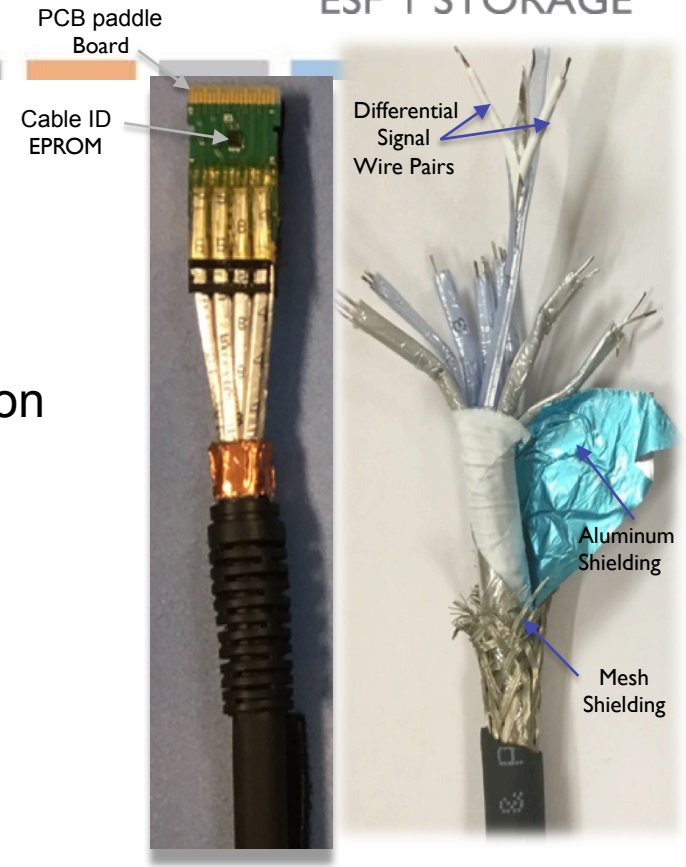
Optical Transceivers

Converts electrical signals to optical laser light sent over optical fibers
Key features = Connectors & Long Reaches

"Transceiver"
4-channels Transmiter
4-channels Receiver

Transceivers with Detachable MPO or LC Connectors

DAC Value Proposition

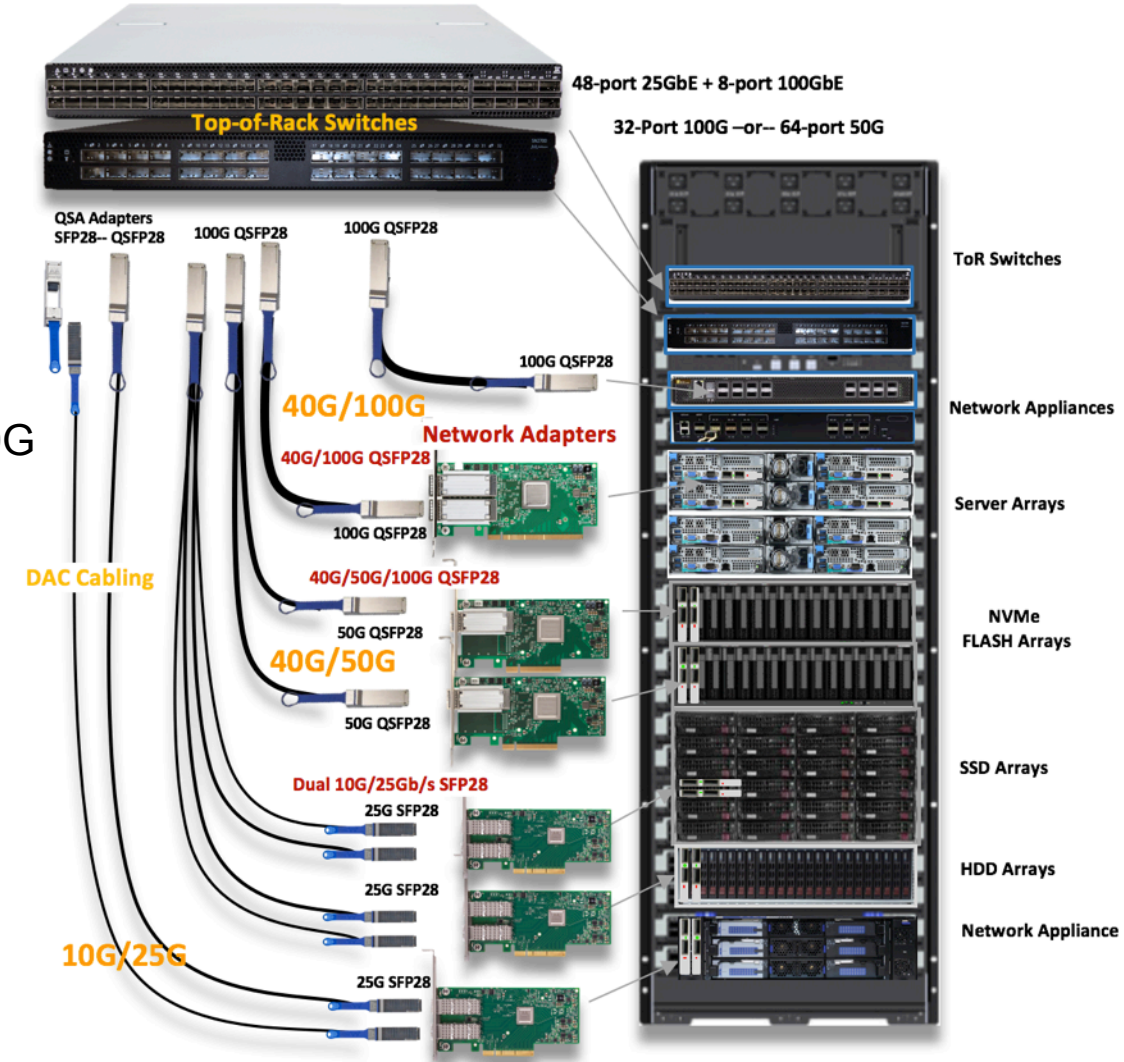


- Low cost
- High reliability – fewest elements to fail
  - (wire, shielding, EPROM, PCB, solder ball)
- No active electronics or optics – simplest construction
- Zero power consumption – no active elements
- Lowest latency & Ultra-low cross talk
- Reaches 3-5m at 25/100G –and- 7m at 10/40G

Used within the rack or to adjacent racks

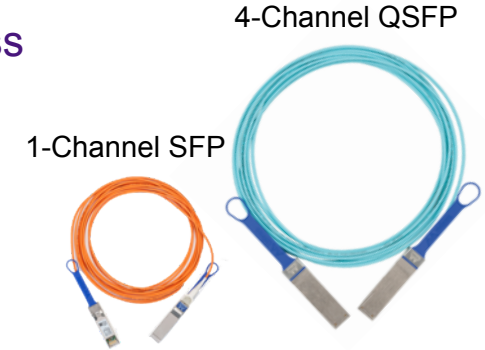
DAC in the Rack

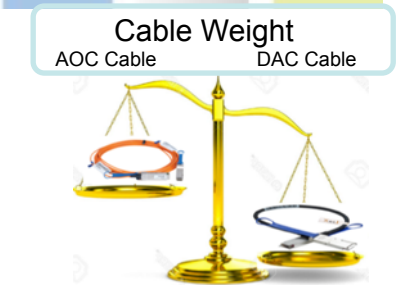
3-5m at 25/50/100G
7m at 10/40G

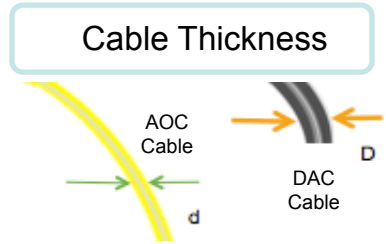


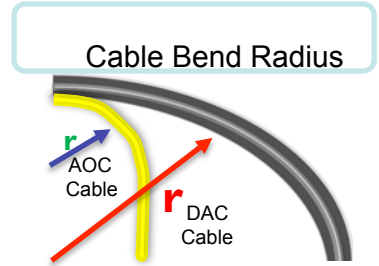
AOC Value Proposition

- **Lowest priced optical solution**
- **3m-to-100m reach**
- "Plug & Play" complete solution
- Dramatically lighter & thinner cable than DAC
  - Increased rack air flow; less "rack cable mess"
  - Tighter cable bends, Easier system access
- Enclosed optics
  - No connector cleaning or reliability issues



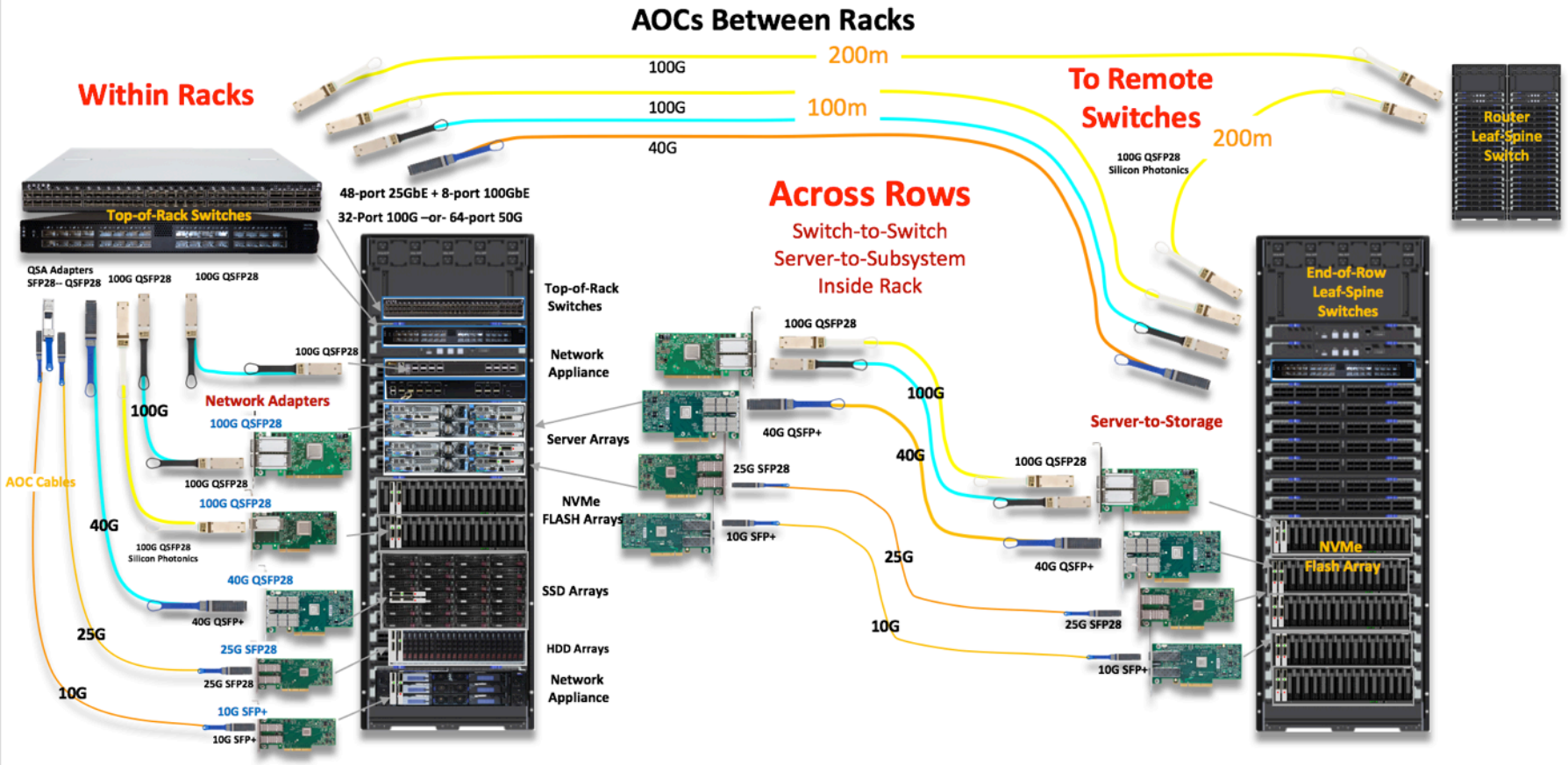






AOCs Across The Top

SNIA ESF | ETHERNET STORAGE



Optical Transceivers – 2 Main Types

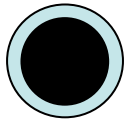
SNIA ESF | ETHERNET STORAGE

SR1

2 fibers

SR4

8 fibers



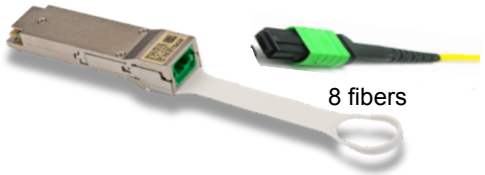
Large Dia fiber

Multi-mode transceivers
Reaches to ~100m

10G SFP+
25G SFP28
Short Reach 1-channel (**SR1**)
VCSEL Laser

40G QSFP+
100G QSFP28
Short Reach 4-channel (**SR4**)
VCSEL Laser

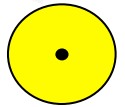
PSM4



8 fibers

CWDM4/LR4

2 fibers



Small Dia fiber

Single-mode transceivers
Reaches to 2-10Km

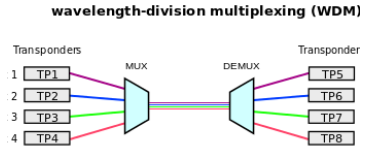
40G QSFP+
100G QSFP28
Parallel Single Mode 4-Ch **(PSM4)**
InP & Silicon Photonics-based

40G QSFP+
100G QSFP28
Long Reach, 4-ch **(LR4)** Coarse WDM, 4-ch **(CWMD4)**

Value Proposition

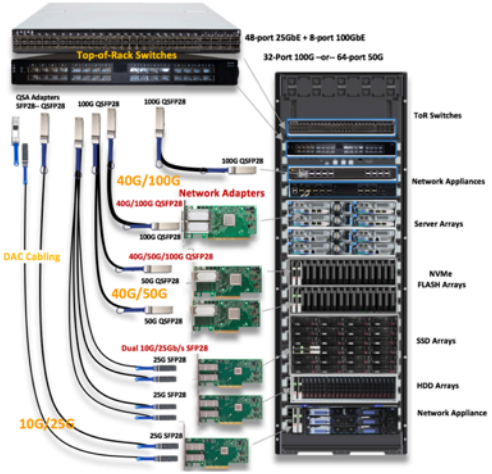
- Long reach 100m-2km and 10km
- Multiple choices of features and costs
- Disconnect-able optical connectors

WDM maps signals to different wavelengths and multiplexes all into a single fiber



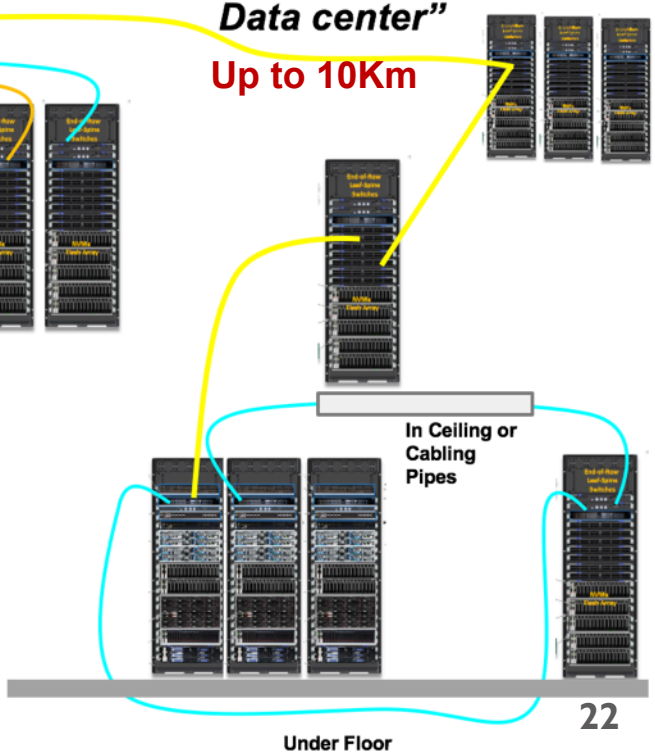
Linking it all Together

"DAC in the Rack
3m-7m



"AOCs Across the Top"
3m-200m

"Transceivers Across the Data center"
Up to 10Km



Ethernet Cabling Summary

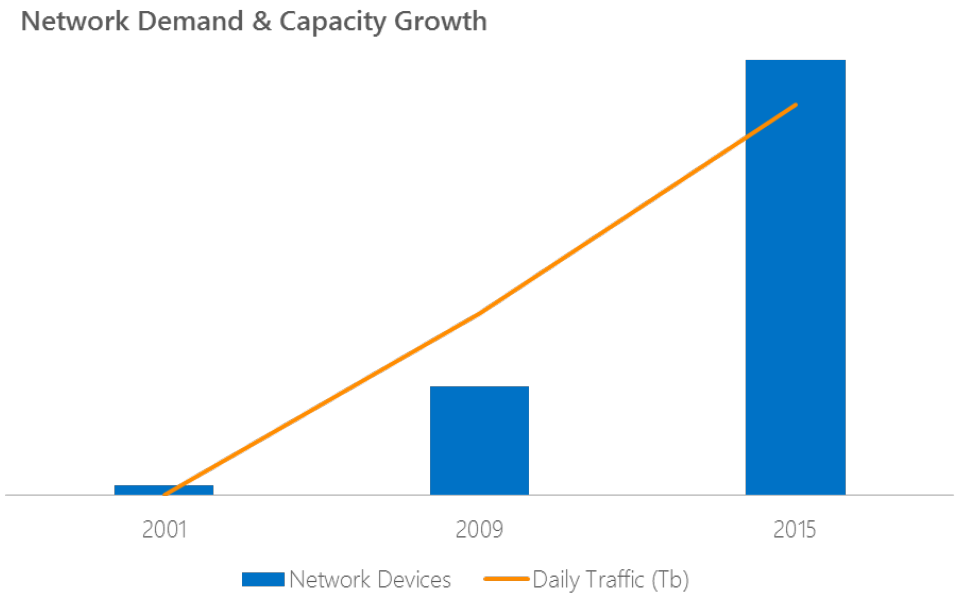
- 3 Major types of cables
- DAC for short distances (3-7m)
- AOC for medium distances (3-40m)
- Transceivers for long distances or re-using existing optical cable (up to 10Km)
- WDM puts multiple lanes on 1 fiber for long reach
- Multiple option choices to minimize costs and maximize performance

Pathway to 200G, 400G and Beyond

Brad Booth
Principal Engineer
Microsoft Azure
brbooth@microsoft.com

The need for speed…

Hyper-scale Network Growth



Scaling Beyond 100GbE

- Contributing factors
  - Data growth and need for replication
  - Increase in VMs
- Supporting technologies
  - PCIe Gen4, NVMe SSDs, Persistent Memory
  - IEEE 802.3 standards projects
  - OIF common electrical interface projects
  - New module form factors: QSFP-DD, OSFP, CFP8, COBO

200GbE and 400GbE Status

- 400GbE
  - Initial discussions of 400GbE starting in late 2012
  - In March 2013, IEEE 802.3 formed a 400GbE study group
  - In March 2014, IEEE P802.3bs (400GbE) task force created
- 200GbE
  - Nov 2015, study group formed 50G, next gen 100G & 200G
  - May 2015 the 200GbE SMF effort merged into P802.3bs
- P802.3bs has entered Sponsor Ballot

Building Blocks

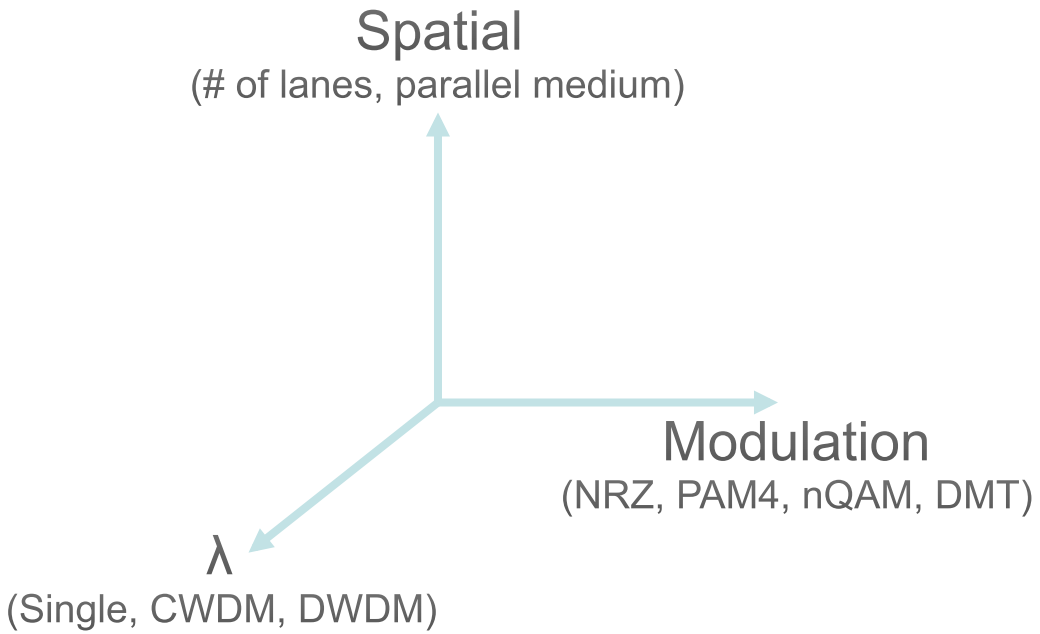
- 200GbE and 400GbE using primarily 50 Gb/s technology
  - 16 x 25 Gb/s for 400GBASE-SR16
  - 4 x 100 Gb/s for 400GBASE-DR4
- Distances and medium supported
  - Up to 3m over copper (200G-CR4)
  - Up to 100M over multimode fiber (200G-SR4, 400G-SR16)
  - Up to 500 m over parallel single-mode fiber (DR4)
  - Up to 2km on single-mode fiber (200G-FR4, 400G-FR8)
  - Up to 10km on single-mode fiber (200G-LR4, 400G-LR8)

Building Blocks (cont.)

- Electrical interfaces
  - 16 lane interface operating at 25 Gb/s NRZ
  - 8 lane interface operating at 50 Gb/s PAM4 (25 Gbaud)
- Optical interfaces
  - 25 Gbaud NRZ
  - 25 Gbaud PAM4 (50 Gb/s)
  - 50 Gbaud PAM4 (100 Gb/s)
- NRZ has lower latency

Increased Bandwidth

- Steps underway
  - 100G per lambda
  - Heavy use of PAM4
  - Parallel medium
- Under consideration
  - 400G per lambda
  - Coherent
  - All optical



Pathway Summary

- 200G/400G path
  - Based primarily on 50 Gb/s technology
  - 100 Gb/s technology used sparingly
  - New module form factor required
- Beyond 400G
  - 100 Gb/s enables interface and optical technology
  - 400 Gb/s per lambda enables WDM and parallel medium to scale to 1.6T+

Overall Summary

- New Ethernet Speeds are Here
  - 25, 50, 100GbE for data center; 2.5/5GbE for access
  - Supports flash and new storage architectures
- Different Cables for Different Use Cases
  - Copper, AOC, and transceivers support different deployments
- 200GbE and 400GbE are Coming
  - Using 50Gb/s and 4x or 8x lanes

After This Webcast

SNIA ESF | ETHERNET STORAGE

- Please rate this Webcast and provide us with feedback
- This Webcast and a PDF of the slides will be posted to the SNIA Ethernet Storage Forum (ESF) website and available on-demand
- www.snia.org/forums/esf/knowledge/webcasts
- A full Q&A from this webcast, including answers to questions we couldn't get to today, will be posted to the SNIA-ESF blog: sniaesfblog.org
- Follow us on Twitter @SNIAESF
- Need help with these terms? Download the 2016 SNIA Dictionary http://www.snia.org/education/dictionary

Thank You