Optical Trends in the Data Center

Doug Coleman
Manager, Technology & Standards
Distinguished Associate
Corning Optical Communications

2017 BICSI Fall
CONFERENCE & EXHIBITION
SEPTEMBER 24-28 | LAS VEGAS, NV

Data Center Trends
Transceiver, Server, Switch, Fiber Type, and Cable Distance

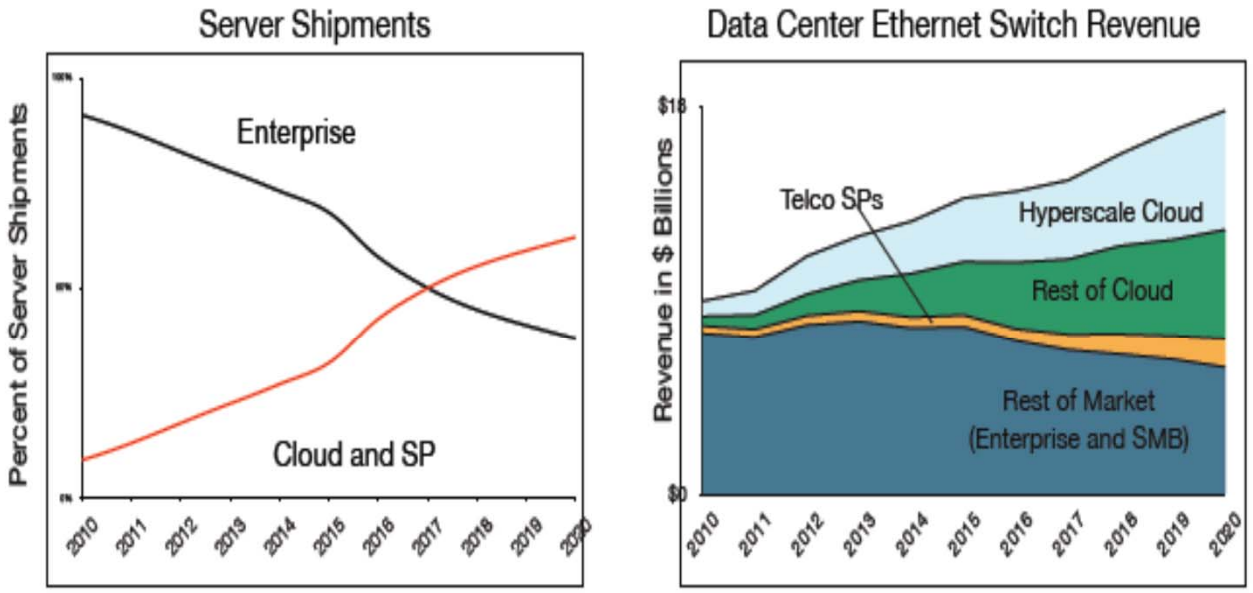
Server Access Switch Speed Migration

Network Location	2014	2016	2018
Mega Cloud server access	10 GE → 40 GE	40 GE → 25/50 GE	25/50 GE → 50/100 GE
Tier 2/3 Cloud server access	1 GE → 10 GE	10 GE → 25 GE	25 GE → 50 GE
Enterprise server access	1 GE → 10 GE	1 GE → 10 GE	10 GE → 25 GE

- Increased server access switch data rates demand higher data rate uplinks – 40/100/200/400G

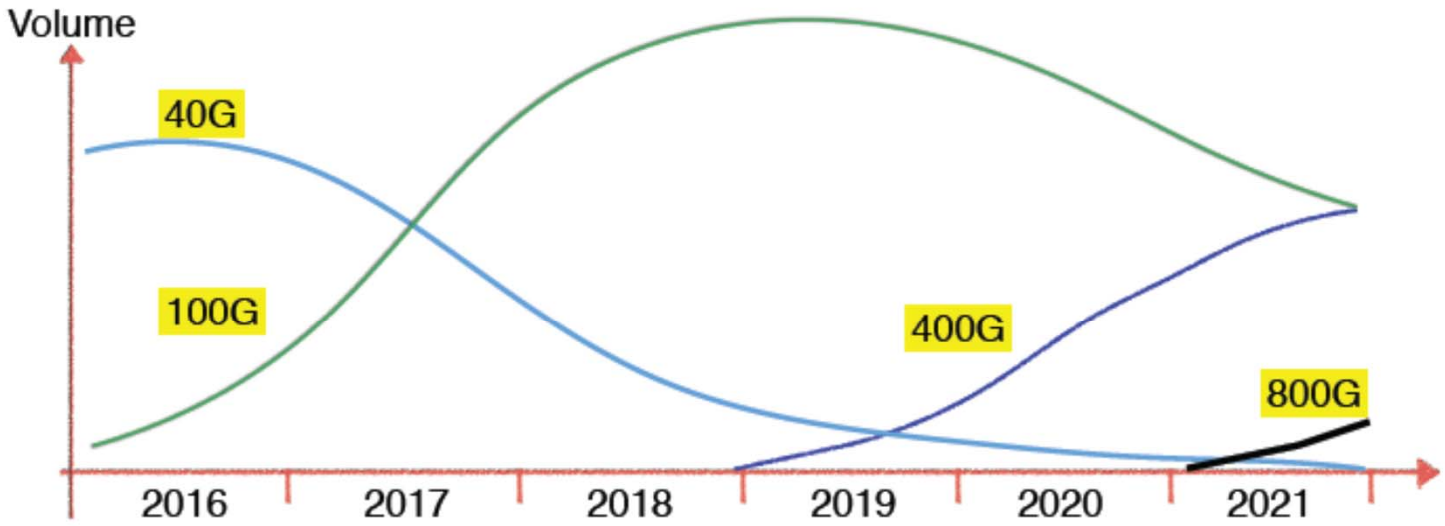
DELL'ORO GROUP

Server and Switch Trends



Source: Dell'Oro Market Research, Ethernet Switch Update, October 2016

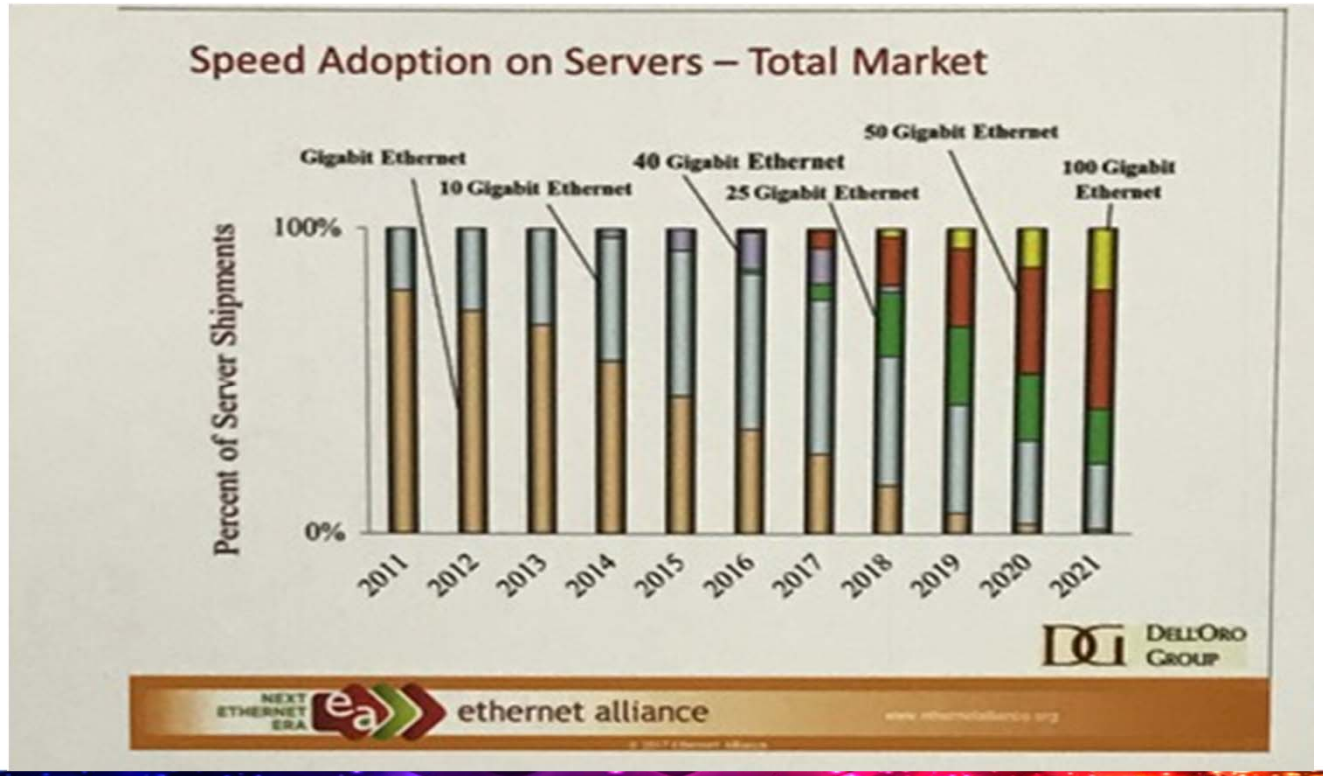
40/100/400G Switch Port Transition



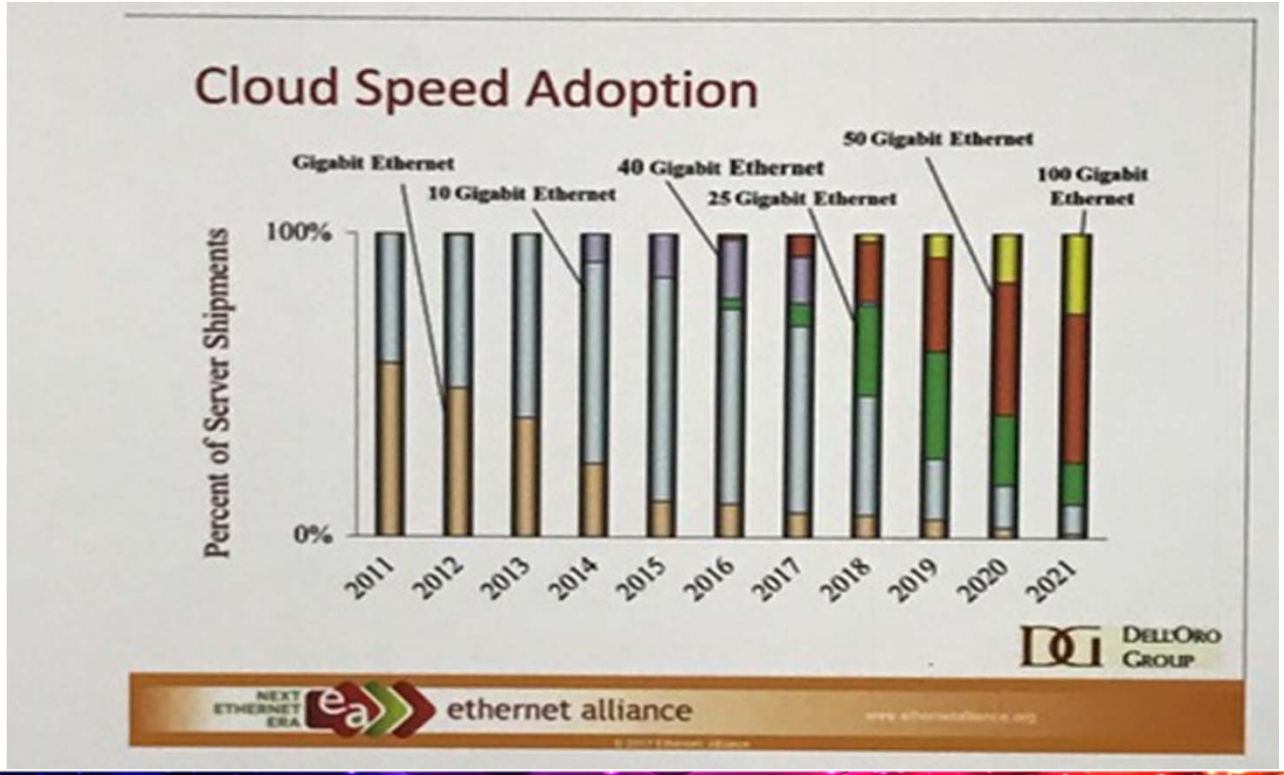
100G includes 4x25G, 400G includes 8x50G and 4x100G

Source: Arista

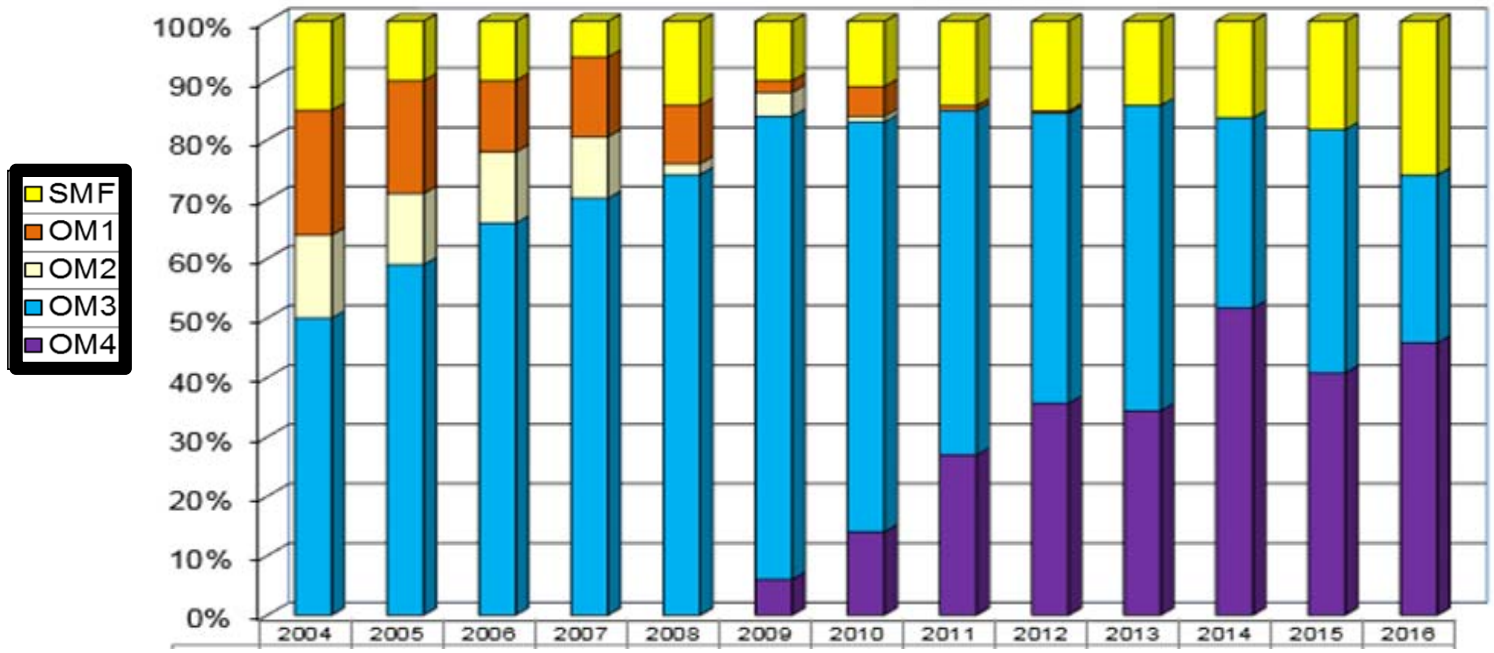
Speed Migration on Servers – Total Market



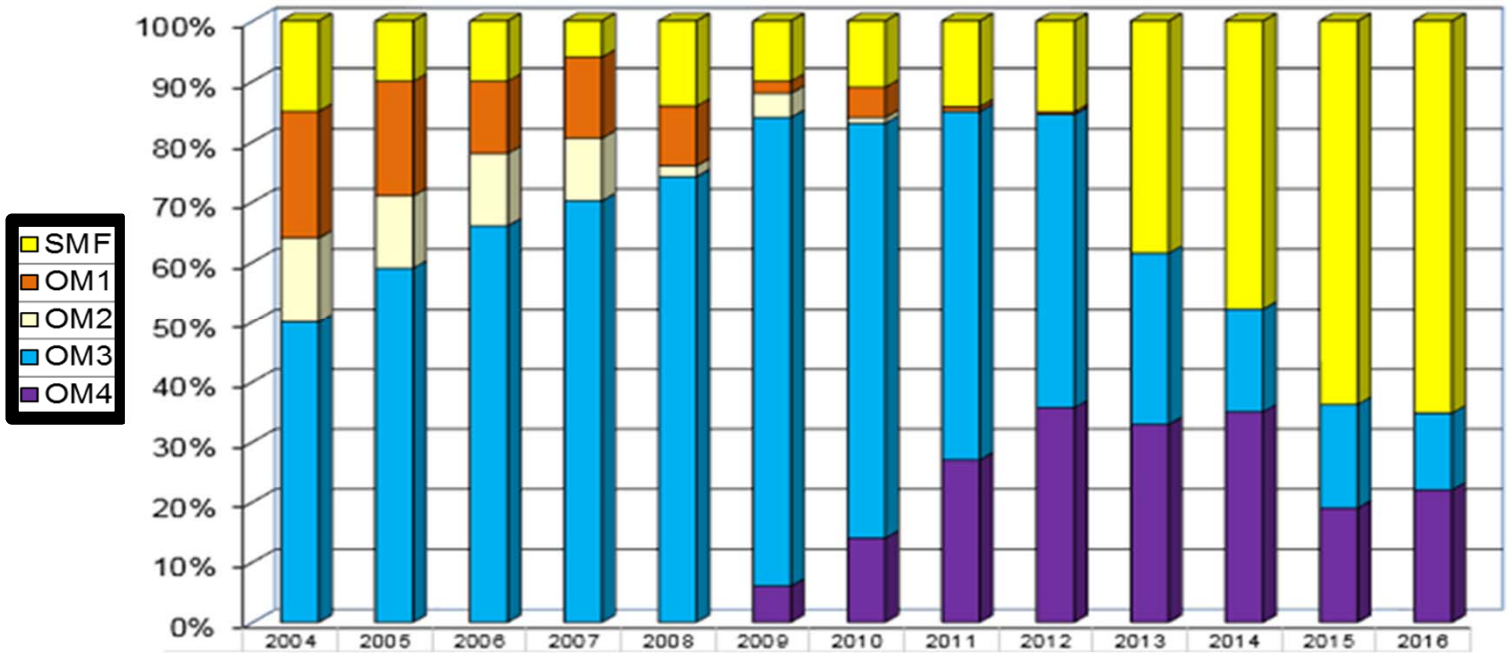
Speed Migration on Servers – Cloud Market



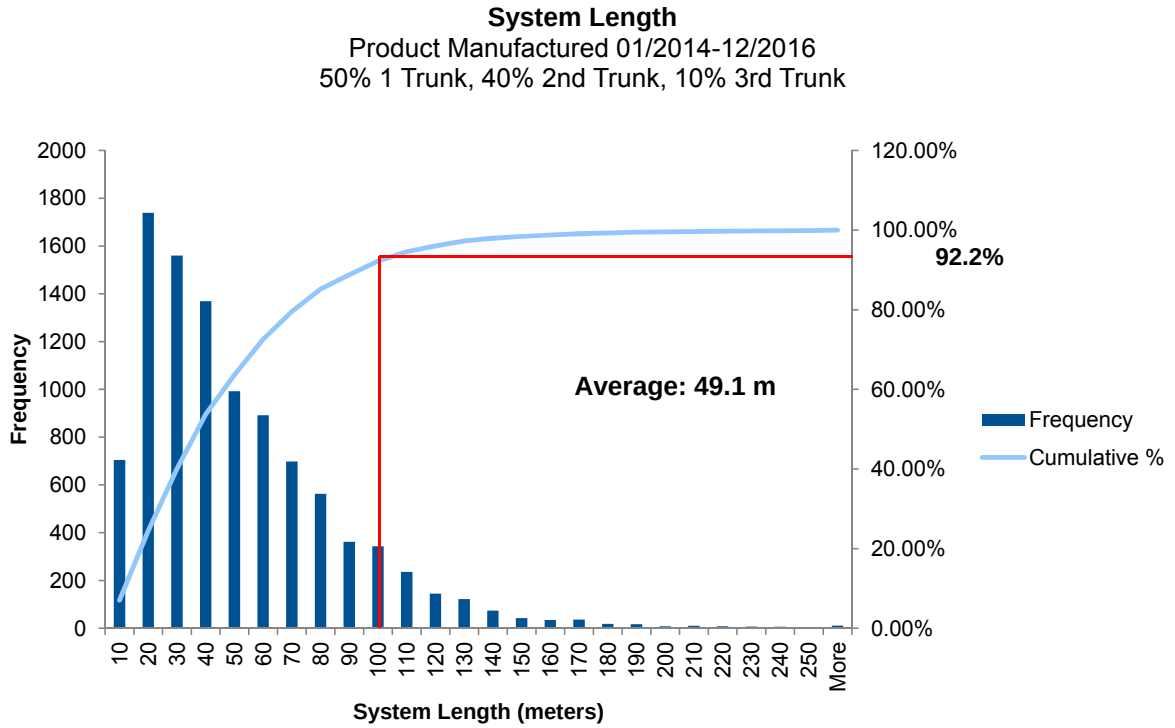
Enterprise Only



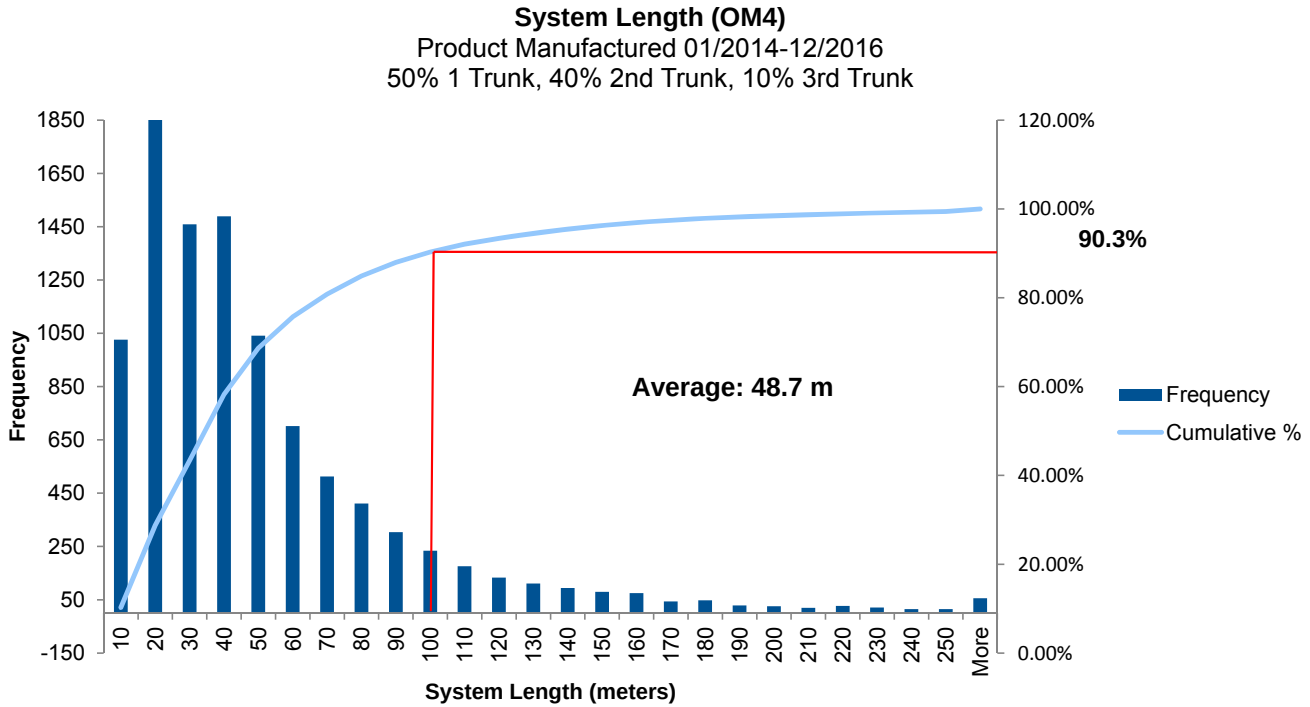
Enterprise and Hyperscale



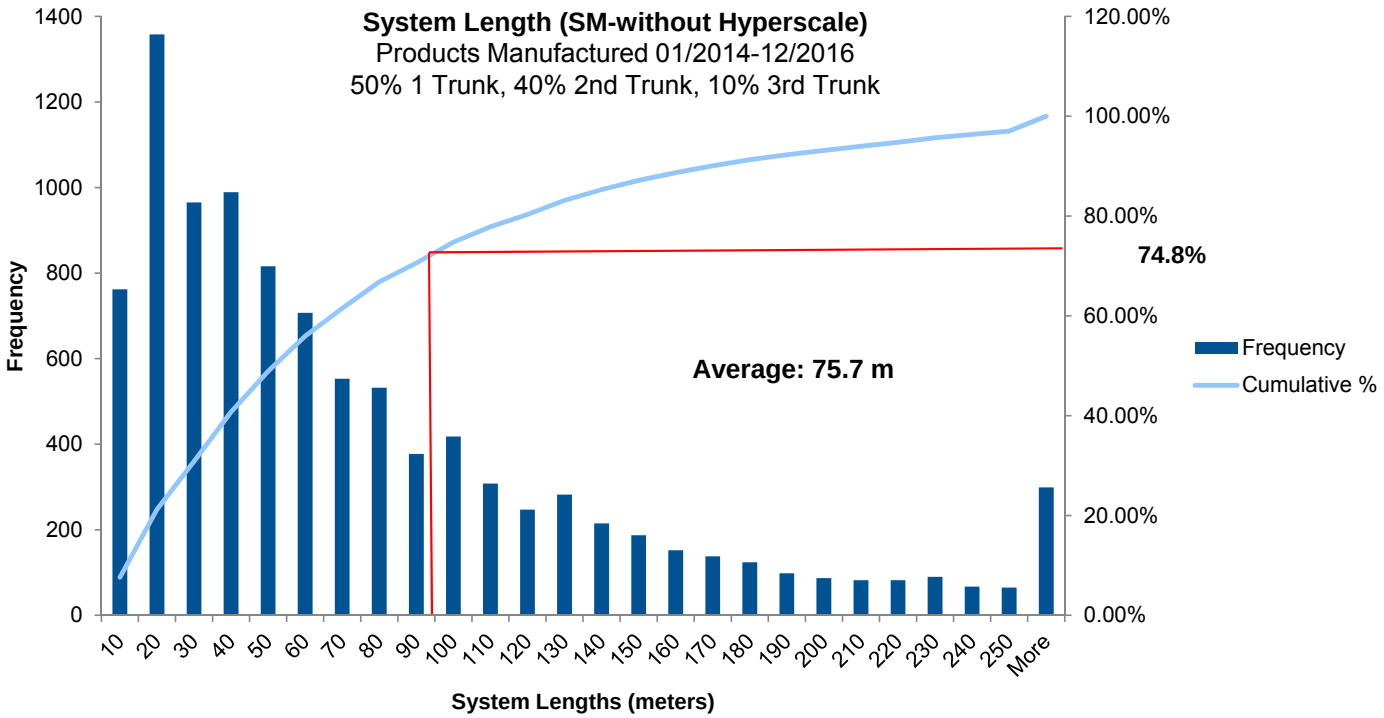
OM3 System Length



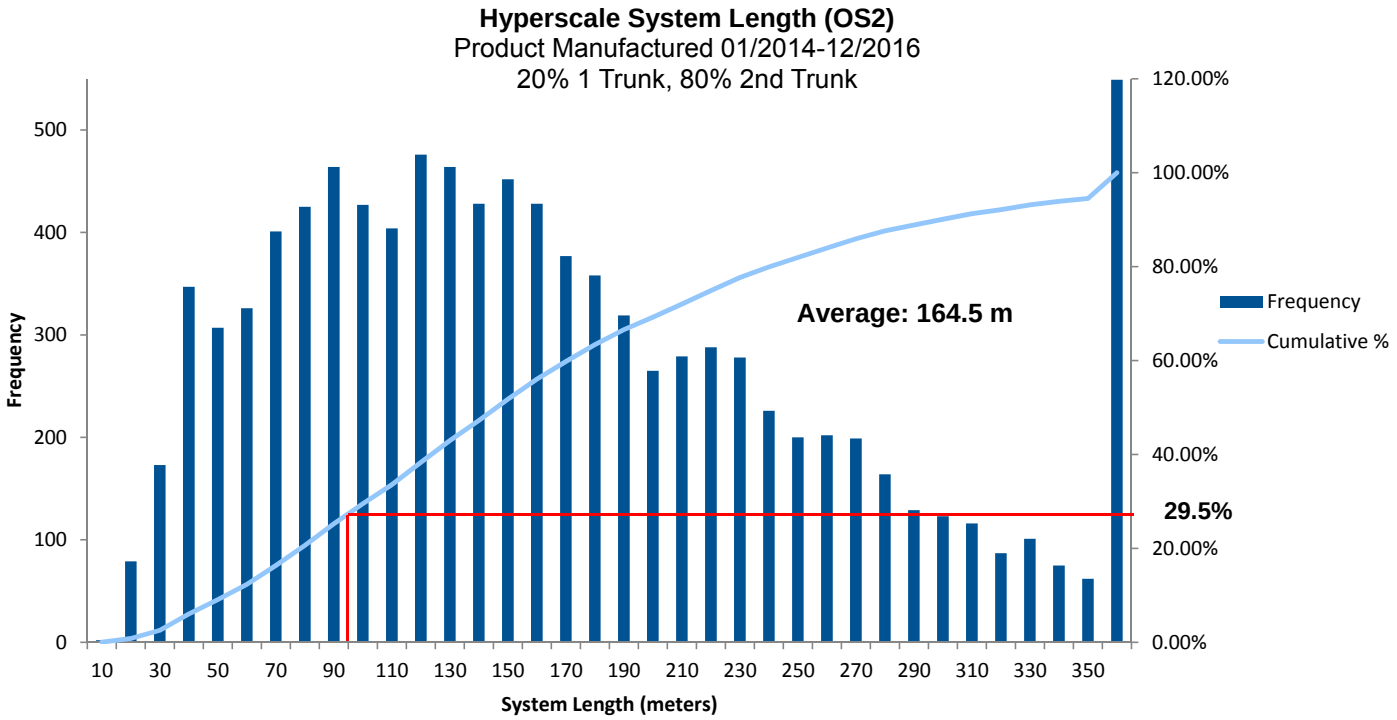
OM4 System Length



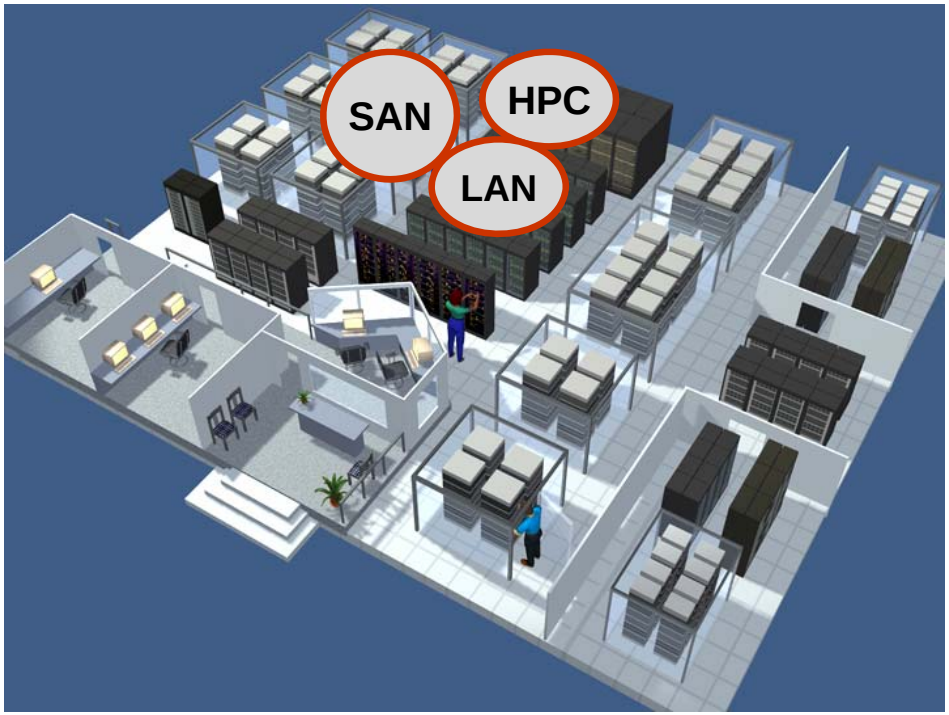
Single-Mode Enterprise



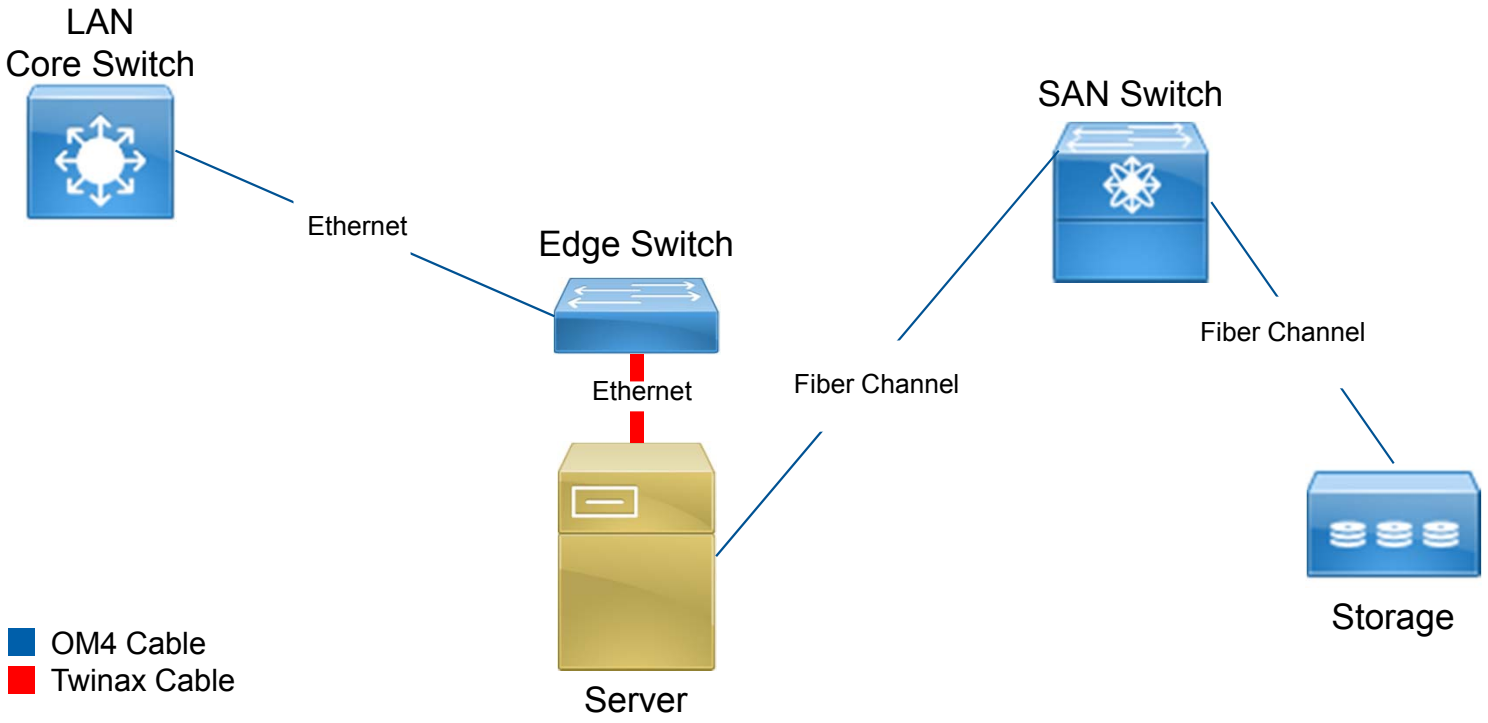
Single-Mode Hyperscale



Data Center Fabrics



Data Center Logical Data Flow



40GbE Port Roadmap

Physical Medium Dependent Sublayers for 40GbE

PMD Name	Electrical Interface to Optical Module	Reach	Medium	Date Standard Ratified
40GBASE-CR4	Not Applicable	7 m	Twinax	2010
40GBASE-SR4	XLAUI / XLPPI	100/150 m	OM3/OM4	2010
40GBASE-LR4	XLAUI / XLPPI	10 km	OS1/OS2	2010
40GBASE-FR	XLAUI	2 km	OS1/OS2	2011
40GBASE-ER4	XLAUI	40 km	OS1/OS2	2015
40GBASE-T	Not Applicable	30 m	Cat 8	2016

Source: Ethernet Alliance

40GeSR4 – Extended Reach Parallel Optics

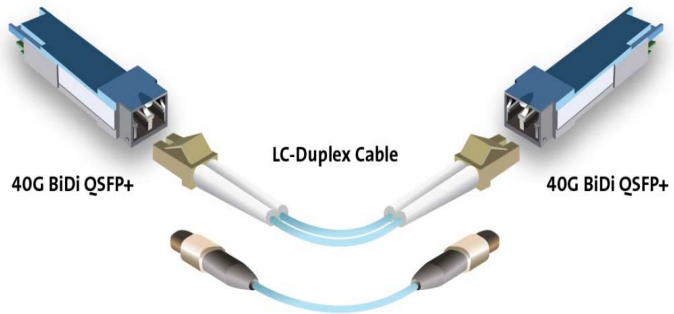
- 40G eSR4 Parallel Optics Extended Reach
  - OM3/OM4:
    - 300 m / 400 m (Standard Assumptions)
    - 330 m / 550 m (Corning Connectivity)

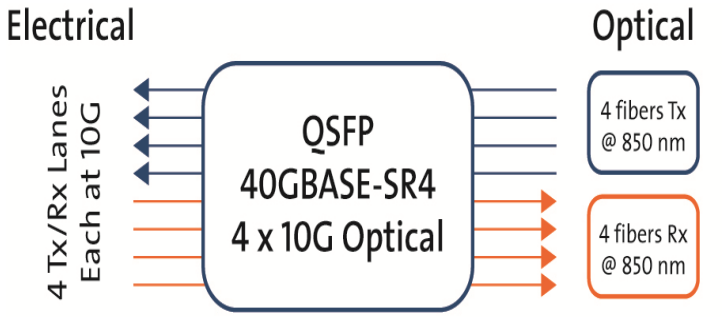
- Corning collaborated with FIT (Avago) to verify the QSFP+ eSR4 modules met link distance specifications for 40G Ethernet applications
  - Internal testing demonstrated 1250m+ with random transceiver and OM4 fiber

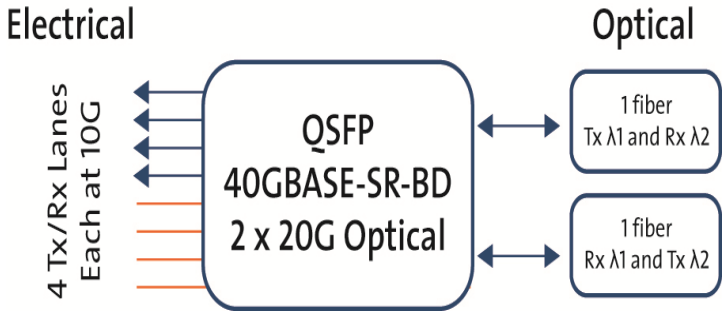
- Commercially available now

Source: Avago

40G BiDi MMF Transmission







Performance

- Duplex OM3/OM4/OM5 LC Connectivity/Polarity
- QSFP Transceiver Form Factor (3.5w)
- Up to 100/200/200m on OM3/OM4/OM5
- Tap Modules Available

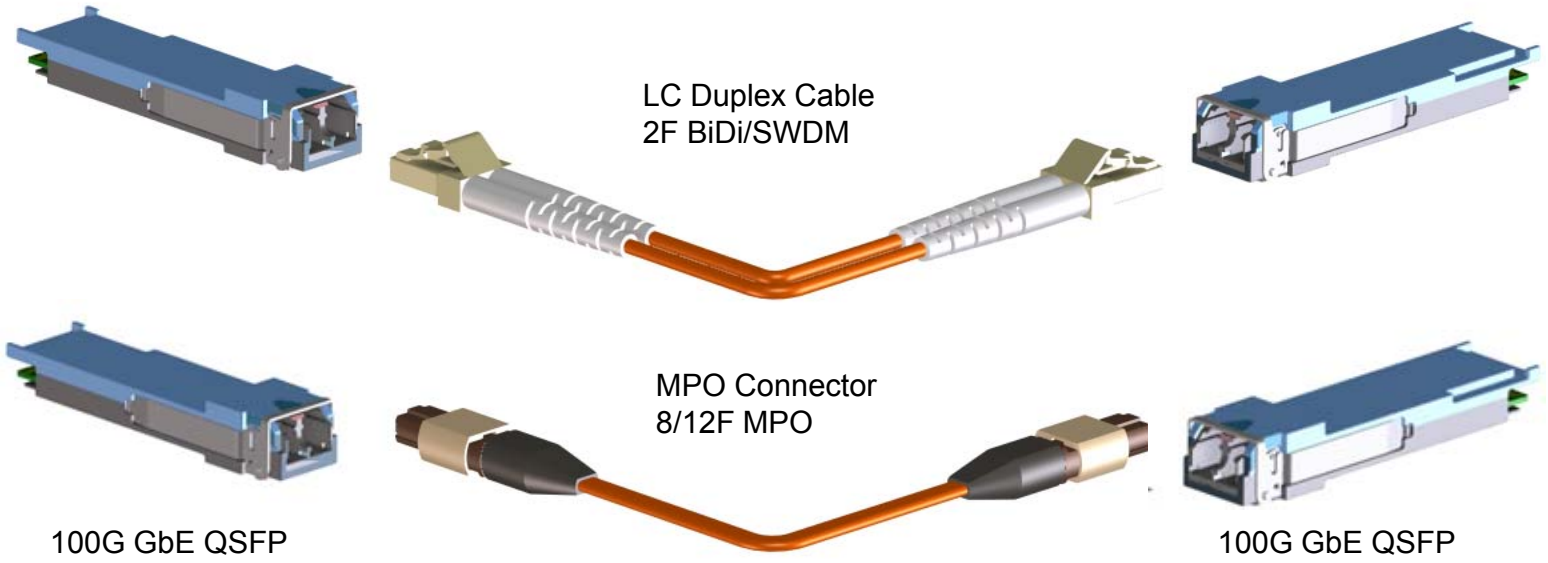
40G Duplex OM3/OM4/OM5 SWDM Transmission

- 40G SWDM QSFP transceiver
  - 4x10G WDM
  - Distance:
    - OM3/OM4/OM5: 240/350/440m
  - Expected Commercial Availability: 2017

- SWDM Manufacturers
  - Finisar and Lumentum

Source: sWDM Alliance

High Level Differences Between SR4 and BiDi/SWDM



100GbE Port Roadmap

PMD Name	Electrical Interface to Optical Module	Reach	Medium	Date Standard Ratified
100GBASE-CR10	N/A	7 m	Twinax	2010
100GBASE-SR10	CAUI-10	100/150 m	OM3/OM4	2010
100GBASE-LR4	CAUI-10	10 km	OS1/OS2	2010
100GBASE-ER4	CAUI-10	40 km	OS1/OS2	2010
100GBASE-CR4	N/A	5 m	Twinax	2014
100GBASE-SR4	CAUI-4	70/100 m	OM3/OM4	2015
100GBASE-LR4	CAUI-4	10 km	OS1/OS2	2015

Source: Ethernet Alliance

802.3bm 100GBASE-SR4 Guidance

- Identical Parallel Optical Lane Assignments to 40GBASE-SR4
- Identical 8F MPO Connectivity for 40GBASE-SR4 and 100GBASE-SR4
  - 24F MPO Connectivity not Compatible
  - Eight OM3 or OM4 Multimode Fiber Compatible

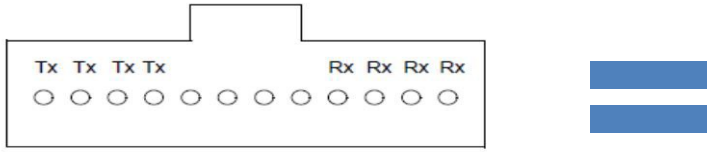


Figure 95–7—100GBASE-SR4 optical lane assignments

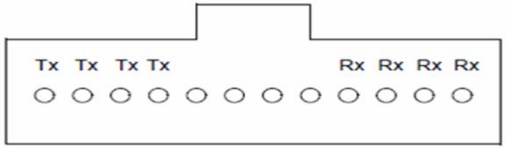
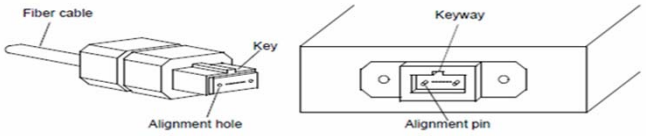


Figure 86–6—40GBASE-SR4 optical lane assignments

Source: IEEE Draft 802.3bm and 802.3ba

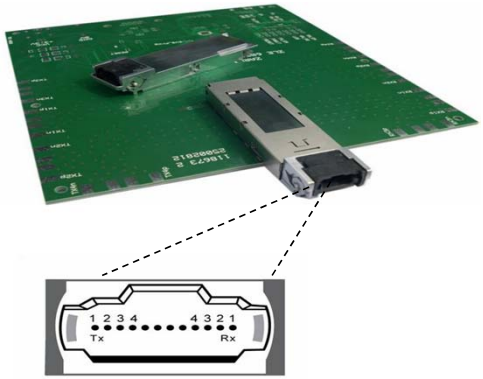
802.3bm 100GBASE-SR4 Guidance

- 8 or 12F MPO
  - Maximum link distance is based on an allocation of 1.5 dB total connection
  - The loss of a single connection shall not exceed 0.75 dB
  - The maximum discrete reflectance shall be =< –20 dB



12F MPO **12F MPO Transceiver Interface**

Figure 95–8—MPO female plug with flat interface and MDI



Source: IEEE 802.3bm and 802.3a

Extended Link eSR4
QSFP28 - 100GBASE-eSR4 Type MMF Transceiver

- Identical build standard to 100G QSFP28 SR4
- 300m (200m) target OM4 (OM3) MMF transmission
- FEC per IEEE 802.3bm for both SR4 and eSR4
- 4 x 25.78Gb/s VCSEL TOSA/ ROSA
- Operating temperature 0ºC to 70ºC
- 3.0W max power consumption (CDRs on)
- Compliant to QSFP28 MSAs,
- Technology scalable to 25GE SFP28 eSR l

Source FIT

Available 4th Quarter 2017

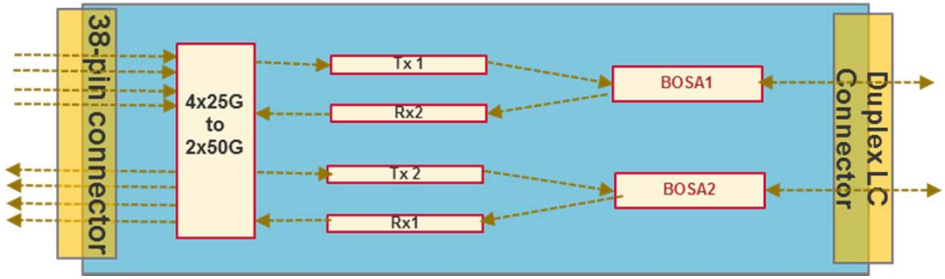
100G BiDi MMF Transmission

Performance

- Duplex OM3/OM4/OM5 LC Connectivity/Polarity
- 2x50G PAM4 or 2x20G NRZ
- 850nm/900nm wavelengths
- FEC Integrated into Transceiver
- QSFP Transceiver Form Factor Up to 70/100/150m on OM3/OM4/OM5
- Backward compatible with 40G BiDi
- Available late 2017

100GE PAM4 BiDi
In QSFP28

Source FIT



Source: FIT

100G Duplex OM3/OM4/OM5 SWDM Transmission

- 100G SWDM QSFP transceiver
  - 4x25G WDM
  - Distance:
    - OM3/OM4/OM5: 75/100/150m
  - Expected Commercial Availability: 2017
- SWDM Manufacturers
  - Finisar and Lumentum

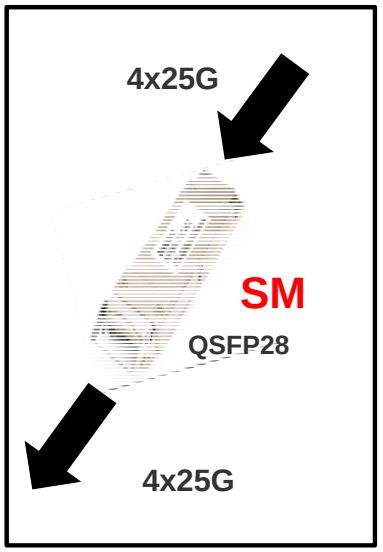
Source: sWDM Alliance

100G SM Activity

- IEEE failed to acquire 75% membership approval for a 100G SM short reach (500m to 2km) variant
  - Proprietary Alliance/MSA Activities
  - Data Center Focus

- CWDM4 MSA
  - 1310nm, 4x25G cWDM4
  - 500m – 2km

- PSM4 MSA
  - 1310nm and 1550nm 4x25G parallel
  - =< 500m



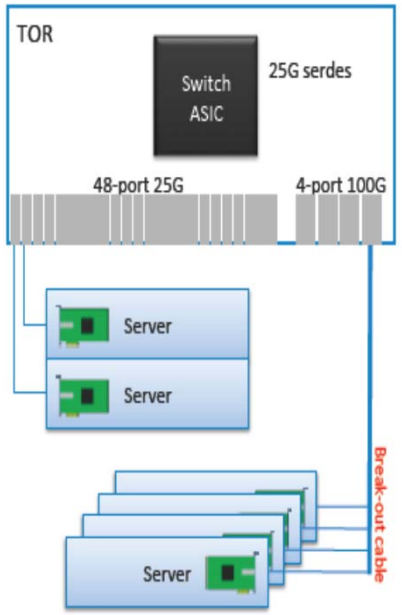
Source FIT

IEEE 802.3 25 Gb/s Ethernet

- 802.3by Task Group
  - ToR to Server interconnect focus
  - 25G Objectives:
    - MMF at least 100m (2F)
    - Twinax at least 3m and 5m
  - Standard completed July 2016

- 802.3cc Task Group
  - ToR to Server interconnect and Carrier focus
  - 25G Objectives
    - SM at least 10km (2F)
    - SM at least 40km (2F)
  - Estimated Standard completion October 2017

AOC, Pluggable Transceiver Modules and Twinax Will be Primary Server Interconnect Method



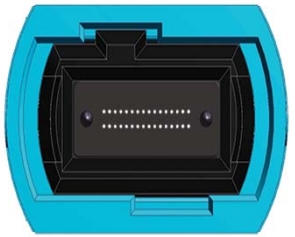
Source: IEEE 802.3 Call For Interest – 25Gb/s Ethernet over a single lane for server interconnect – July 2014 San Diego

IEEE 802.3bs 400G Ethernet Task Force

- Study Group Adopted November 2013
  - 400G Objectives:
    - At least 100 m over MMF
    - At least 500 m over SMF
    - At least 2 km over SMF
    - At least 10 km over SMF
- Hyperscale Data Center Focus
- IEEE 802.3bs 400G Task Force Approved March 2014
  - Expected standard completion date: December 2017

400G MMF Variants

- First Generation Standards Guidance 32F Parallel Optics (2017)
  - 400GBASE-SR16
  - Tx 16x25G and Rx 16x25G (NRZ)
  - 32F MPO Connector, Double Array
  - Minimal traction expected with AOC primary use



2x16 format

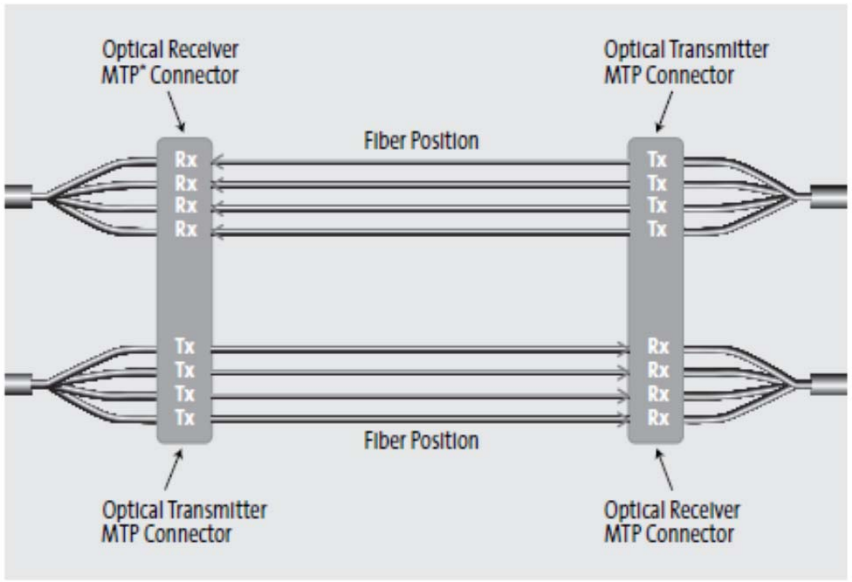
- (Pending) Second Generation Expected to Utilize 8F Parallel Optics with WDM (2019)
  - Tx 4x100G and Rx 4x100G
    - Two wavelengths at 25Gbaud PAM4
    - Four wavelengths at 25Gbaud NRZ
  - 8/12F MPO Connector
  - Deployed as pluggable module / embedded optics for core to edge uplinks

1x12 or 8 format

Source: USConec

IEEE 802.3bs adopts 200G SM

- 200Gb/s PHYs
  - SMF at least 2km (2F)
    - 200GBASE-FR4
  - SMF at least 10km (2F)
    - 200GBASE-LR4
  - SMF at least 500m (8F)
    - 200GBASE-DR4
- 50G PAM4 Modulation



200G SMF Parallel 4x50G (PAM4)

400G SMF Variants

- At least 500m - 8F Parallel Optics
  - 400GBASE-DR4
  - Tx 4x100G and Rx 4x100G (50Gbaud PAM4)
  - 8F or 12F MPO Connector
- At least 2km - 2F Serial Optics
  - 400GBASE-FR8
  - 8x50G WDM (25Gbaud PAM4)
  - 2F LC Connector
- At least 10km - 2F Serial Optics
  - 400GBASE-LR8
  - 8x50G WDM (25Gbaud PAM4)
  - 2F LC Connector

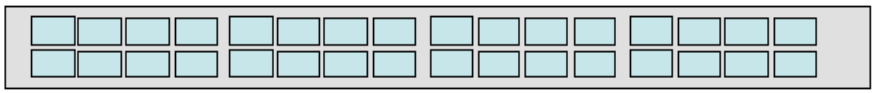
WDM Lane Assignments

Lane	Center Frequency THz	Center Wavelength nm	Wavelength Range nm
L0	235.4	1273.55	1272.55 to 1274.54
L1	234.6	1277.89	1276.89 to 1278.89
L2	233.8	1282.26	1281.25 to 1283.28
L3	233.0	1286.66	1285.65 to 1287.69
L4	231.4	1295.56	1294.53 to 1296.59
L5	230.6	1300.05	1299.02 to 1301.09
L6	229.8	1304.58	1303.54 to 1305.63
L7	229.0	1309.14	1308.09 to 1310.19

- 200/400G SM Connector RL Requirements : MPO: -45 dB, LC : -35 dB

400G Pluggable Modules

- QSFP-DD and OSFP
- 32 ports per 1U line card
- 100/200G breakout capability
- Module power 10-12w
- MPO or LC Connector
- MM and SM Capable
- 2018-2019 Availability

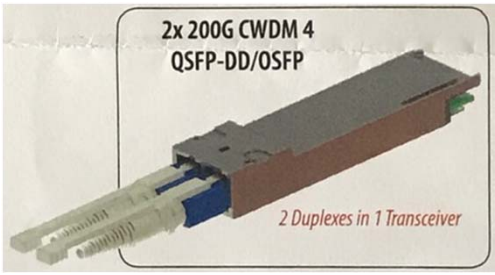


32 OSFP or 32 QSFP-DD Ports
12.8 Tb / 1U

OSFP

QSFP-DD

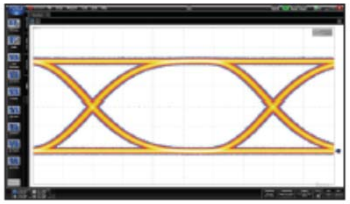
Micro-LC



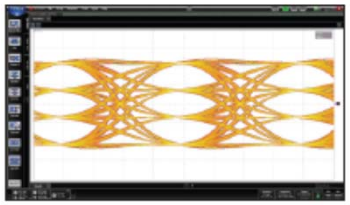
Multi-Level Encoding 50G (PAM4)
MMF and SMF

- Non-Return to Zero (NRZ)
  - Typical modulation method to 25G at 1 bit / symbol
  - Electrical channel loss and noise issues at =>25G
  - Two amplitude levels
  - Potential MMF VCSEL NRZ limit
- Pulse Amplitude Modulation (PAM4)
  - Evolving systems increase the amount of information (bits) sent with each symbol
  - Requires more complex transceivers
  - Doubles the BW with 2 bits/symbol
  - Four amplitude levels
  - 4-6 dBo optical penalty
    - FEC power gain compensates
  - Extends MMF VCSEL lifecycle

NRZ (PAM-2)

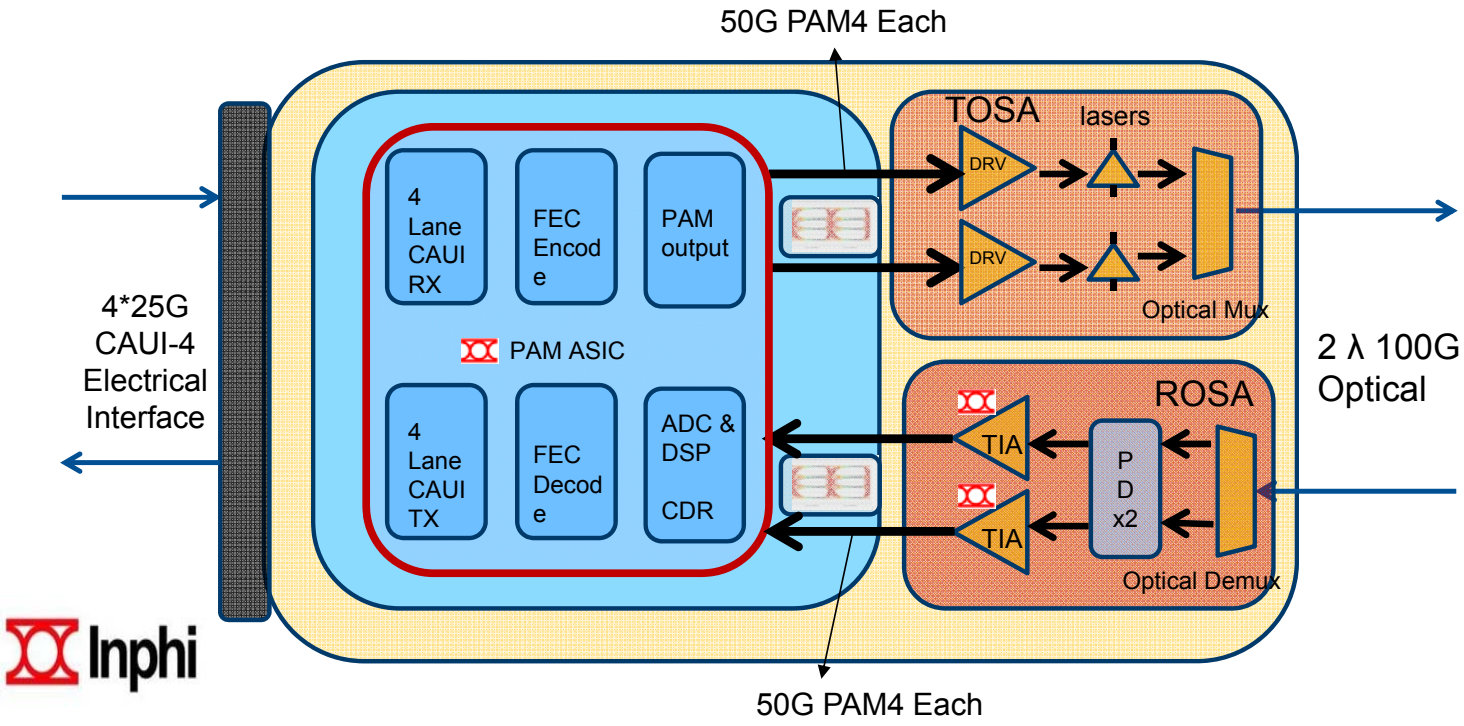


PAM-4



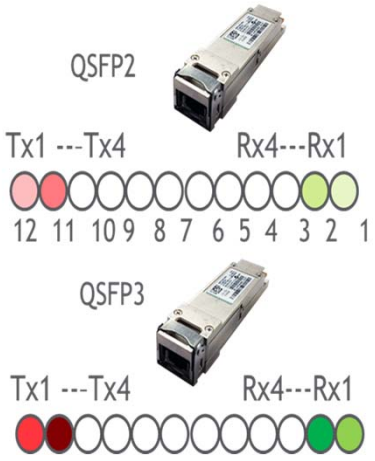
Source: KEYSIGHT Technologies

PAM4 ASIC: 100G Module Implementation Example



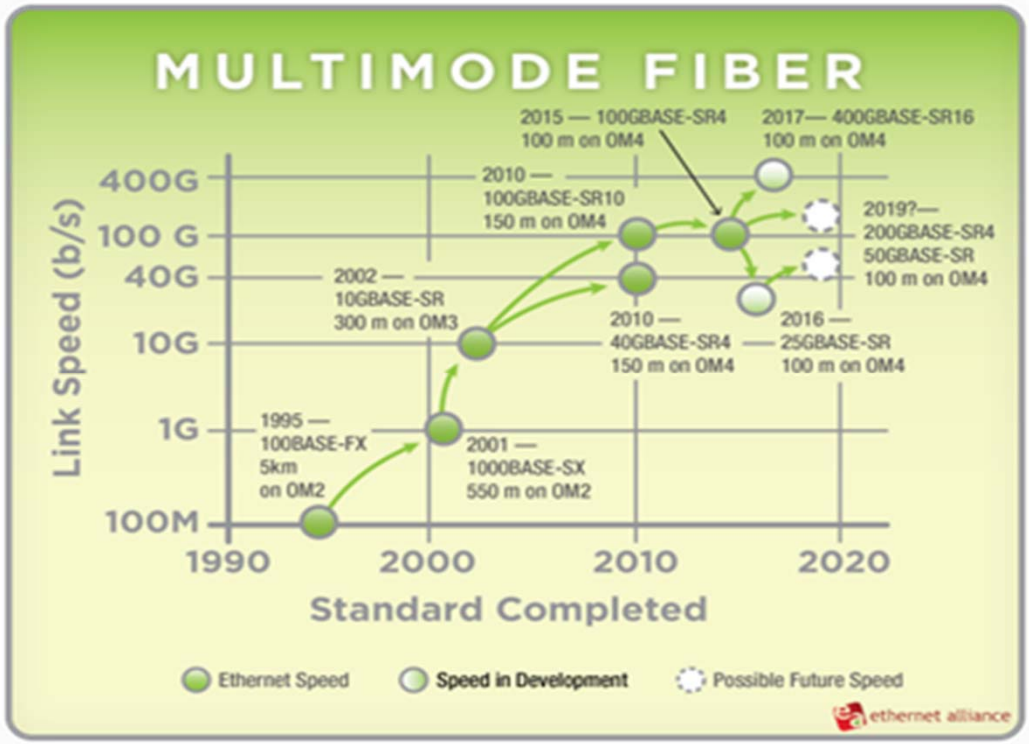
IEEE 802.3cd 50/100/200G Task Force

- 50Gb/s
  - MMF at least 100m (2F)
    - 50GBASE-SR - 25Gbaud PAM4
  - SMF at least 2km (2F)
    - 50GBASE-FR - 25Gbaud PAM4
  - SMF at least 10km (2F)
    - 50GBASE-LR – 25Gbaud PAM4
- 100 Gb/s,
  - MMF at least 100m (4F)
    - 100GBASE-SR2 – 25Gbaud PAM4
    - 25Gbaud PAM4
  - SMF at least 500 m (2F)
    - 100GBASE-DR – 50Gbaud PAM4
- 200 Gb/s
  - MMF at least 100m (8F)
    - 200GBASE-SR4 – 25Gbaud PAM4

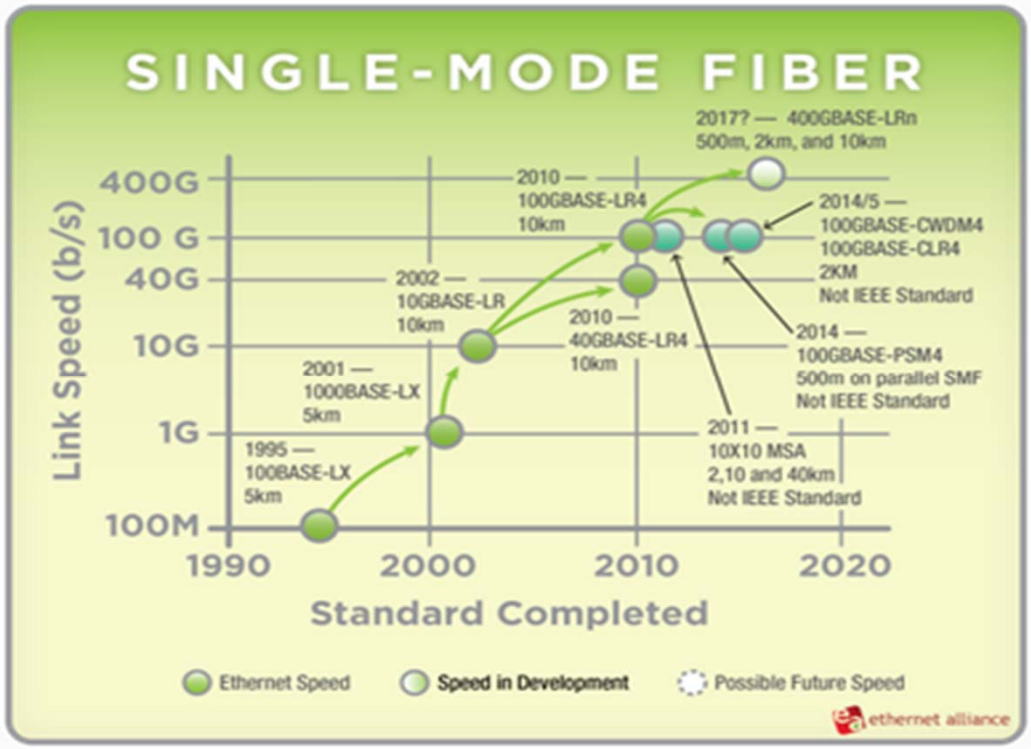


Source: Brocade

Ethernet Speed Roadmap



Ethernet Speed Roadmap



Fiber Channel Speed Roadmap

Product Naming	Throughput (Mbytes/s)	Line Rate (Gbaud)	T11 Specification Technically Complete (Year)*	Market Availability (Year)*
1GFC	200	1.0625	1996	1997
2GFC	400	2.125	2000	2001
4GFC	800	4.25	2003	2005
8GFC	1,600	8.5	2006	2008
16GFC	3,200	14.025	2009	2011
32GFC	6,400	28.05	2013	2016
128GFC	25,600	4X28.05	2014	2016
64GFC	12,800	56.1	2017	2019
256GFC	51,200	4X56.1	2017	2019
128GFC	25,600	TBD	2020	Market Demand
256GFC	51,200	TBD	2023	Market Demand
512GFC	102,400	TBD	2026	Market Demand
1TFC	204,800	TBD	2029	Market Demand

FC

Source: FCIA Speedmap v20

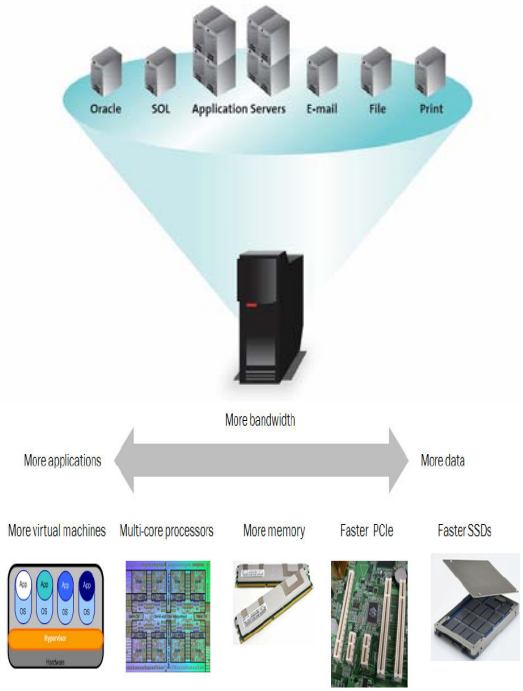
FC-PI7 64G / 256G Objectives

- Backward compatibility to 32GFC and 16GFC
- Same external connectors as present connector
  - LC and SFP+ for 64GFC
  - MPO and QSFP56 for 256GFC
- Distances
  - 100m on OM4 cables for 64GFC (2F)
  - 10km on SMF cables for 64GFC (2F)
  - 100m on OM4 cables for 256GFC (8F parallel)
  - 2km on SMF cables for 256GFC (8F parallel, 2F WDM)
- PAM4 multi-level encoding (28 GBaud (56GFC))
- Server virtualization and flash storage require higher FC data rates
- Standard Completion 2018

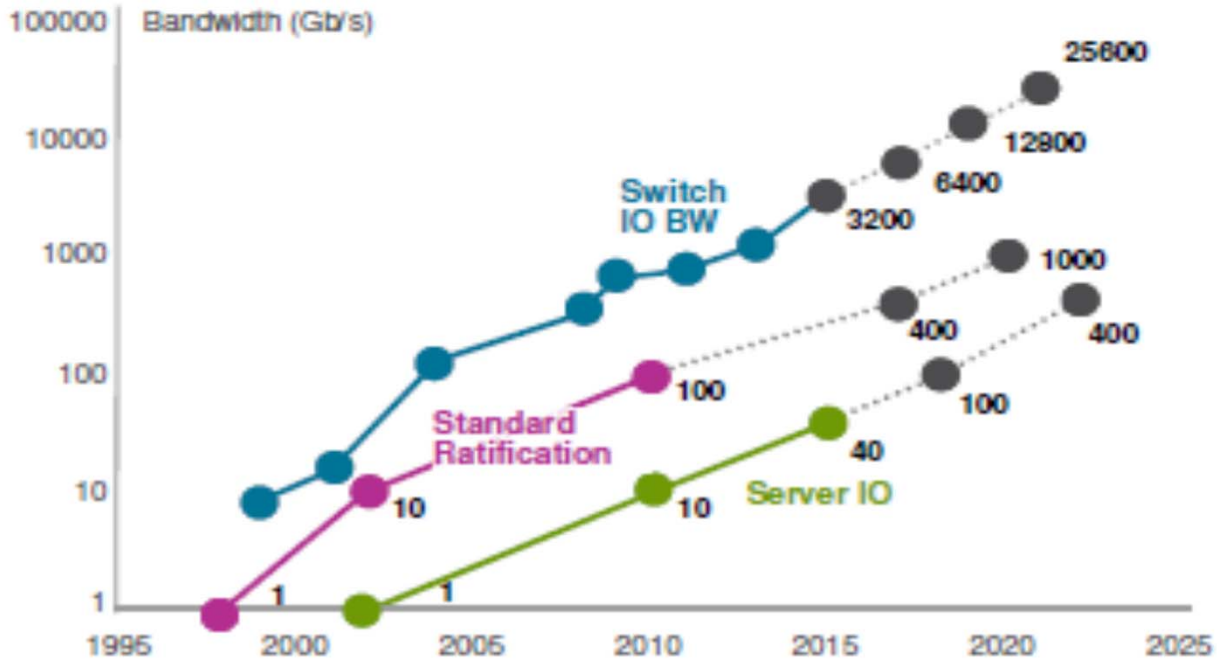
Server Virtualization Drives Higher Data Rates

- Multiple applications run in parallel on one server.
- Increases utilization efficiency from < 20% to >70%
- CPU cores (4,8,18..28)
  - Typical Enterprise / Hyperscale server:4-12 c
  - Typical HPC server: 18-28 cores
- Typical 2-3GHz/core
- Increased Memory (6-12 TB)
- PCIe2: 8 lanes @ 5G 8b/10b
- PCIe3: 16 lanes @ 8G 128b/130b
- PCIe4: 16 lanes @ 16G 128b/130b (2018)

- Drives utilization of high data rate optical connectivity (40/100/400G) to support increased sever I/O speeds (10/25/50/100G)



Switch ASIC Roadmap



Source: Broadcom

Source: Broadcom

Ethernet Lane Speed Transition

Lane Speed	10Gbps	25Gbps	50Gbps	100Gbps
1X	10G	25G	50G	100G
2X	—	50G	100G	200G
4X	40G	100G	200G	400G
8X	—	—	400G	800G
Availability	2011	2015	2018 (E)	2020 (E)

Source: Arista

Source: Arista

Switch Silicon Manufacturing Status

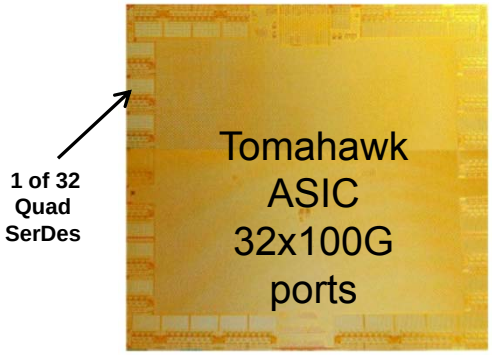
Switch Vendor	Product	Bandwidth	Serdes	Notes
Broadcom "The incumbent"	Trident II	1.28 Tb	128 x 10G	Volume 40GbE
	Tomahawk	3.2 Tb	128 x 25G	Volume 100GbE
	Tomahawk2	6.4 Tb	256 x 25G	Sampling now
Barefoot Networks 2013 Startup	Tofino family	1.9Tb	96x 10G	User programmable Edge-Core Networks building into Wedge switch
		2.5 Tb	100 x 25G	
		3.3 Tb	132 x 25G	
		6.5 Tb	260 x 25G	
Cavium	Xpliant	3.2 Tb	128 x 25G	Programmable, updateable Chosen by Arista, Brocade
Centec Networks	Golden Gate	1.2 Tb	96 x 10G	Chinese semiconductor and white box vendor Open Compute Switch Abstraction Layer
Innovium 2014 startup	TeraLynx family	3.2 Tb	128 x 25G	First to announce switch with 50G PAM4 serdes 12.8 Tb in single 16nm die Programmable
		6.4 Tb	128 x 50G PAM4	
		9.6 Tb	192 x 50G PAM4	
		12.8 Tb	256 x 50G PAM4	
Mellanox	Spectrum	3.2 Tb	128 x 25G	8th generation switch silicon Support Open Ethernet Also offer complete switches
Nephos Mediatek 2016 spin-off	Taurus	1.2 Tb	96 x 10G	Supporting OCP specs 10G switch shipping 6.4 Tb switch demo
		3.2 Tb	128 x 25G	
		6.4 Tb	256 x 25G (2 die)	

Source: Lightcounting

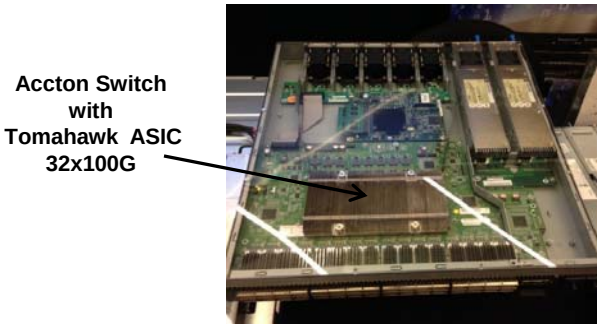
Source: Lightcounting

Switch ASIC Environment

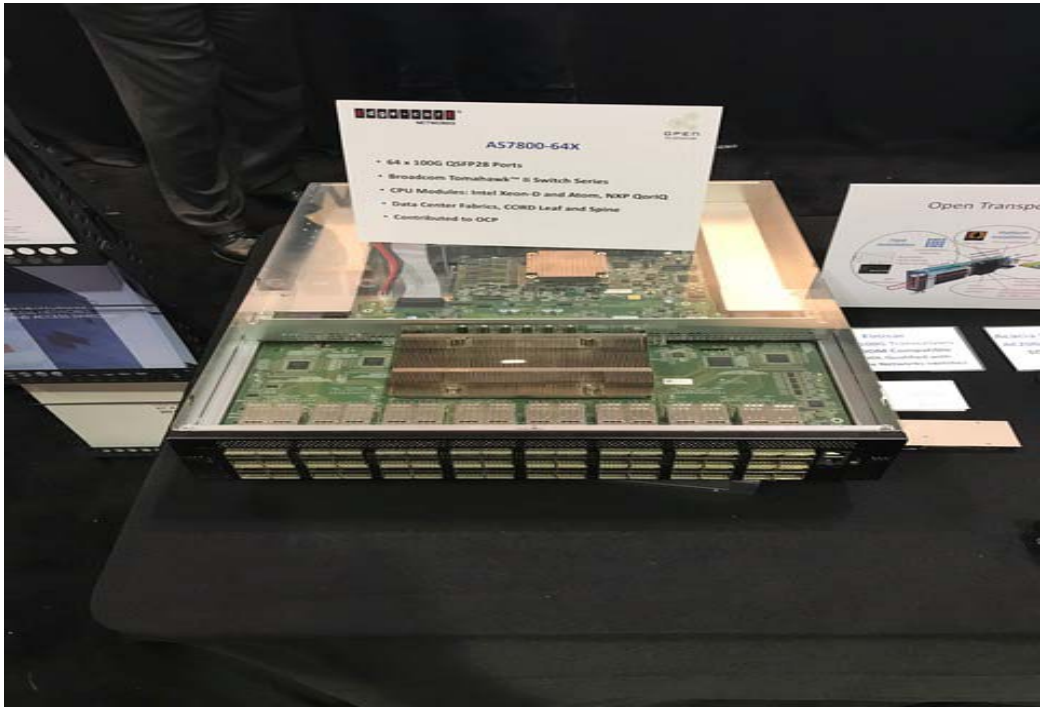
- Broadcom Tomahawk 3.2 Tb/s ASIC
  - 128x25G SerDes lanes
  - 7 Billion Transistors
  - ~200W power dissipation
  - 32x100G Ports, 128x25G ports, 64x50G ports
- Expected Industry ASIC Roadmap
  - 3.2 Tb/s (25G per lane, 100G/port)
  - 4.8Tb/s (25G per lane, 100G/port)
  - 6.4 Tb/s (25/50G per lane, 100G/port)
  - 12.8 Tb/s (50/100G per lane, 400G/port
  - 25.6 Tb/s (100G per lane, 800G/port)



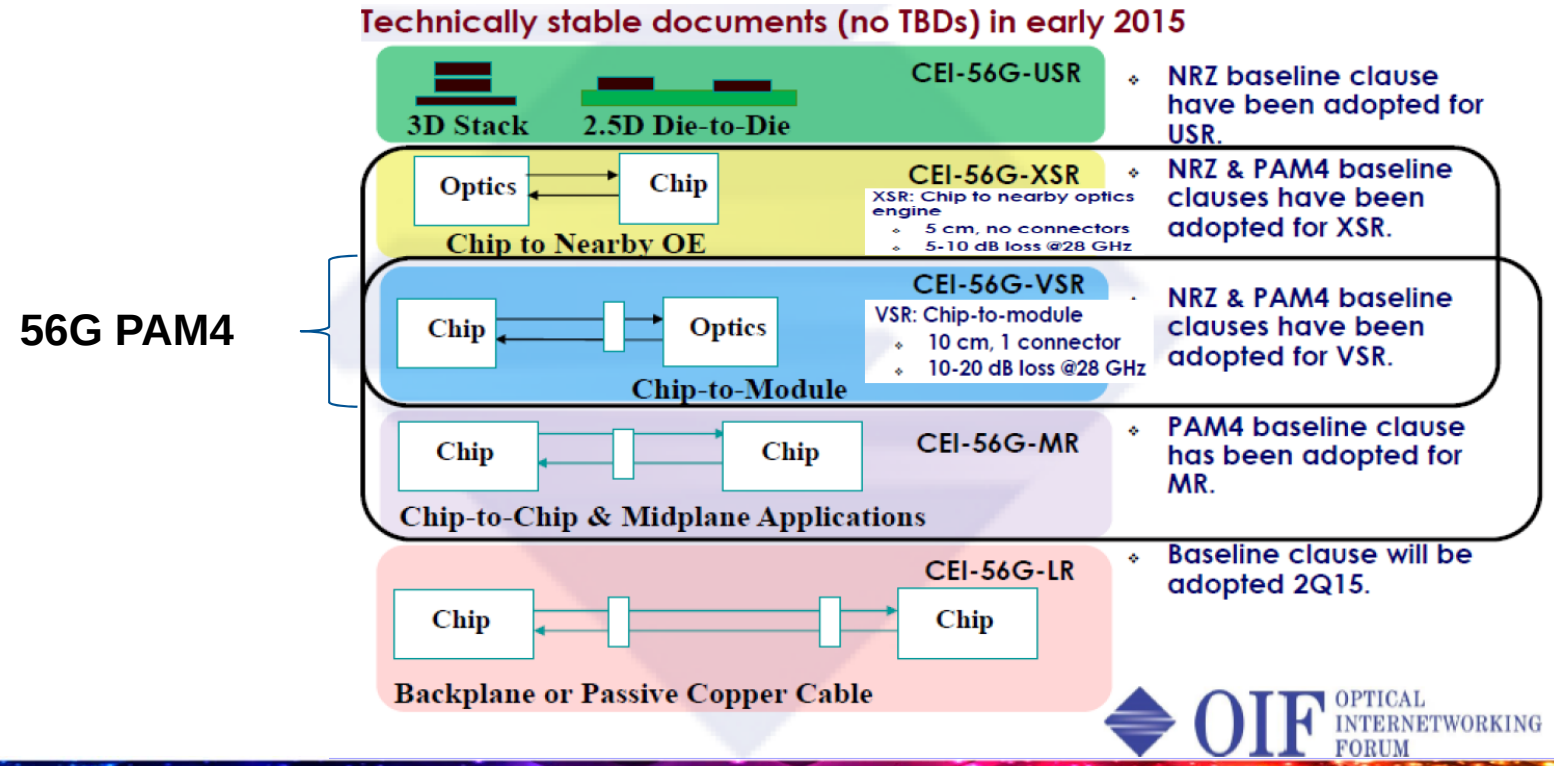
Source: Broadcom



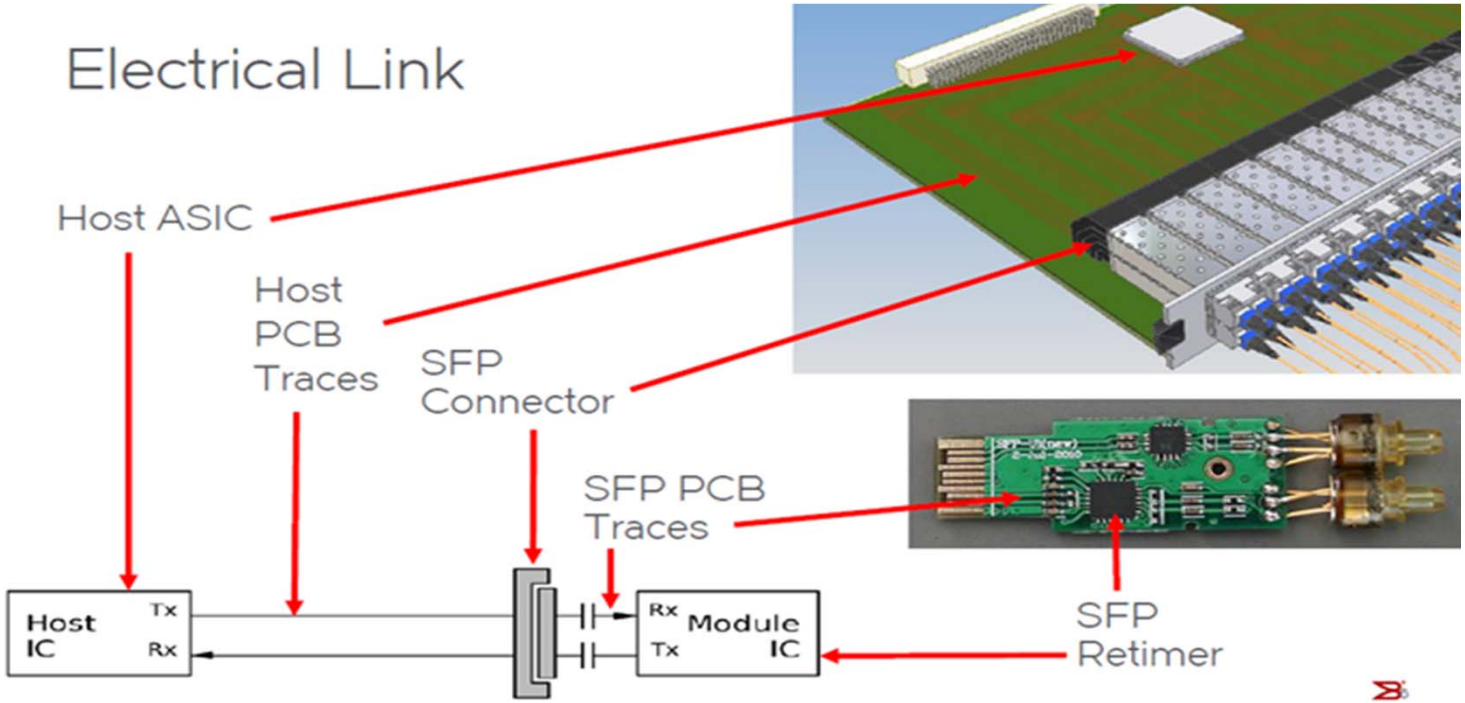
ACCTON 64x100G 2U Switch (Tomahawk I 256x25G lanes)



OIF CEI-56G Baseline Clauses

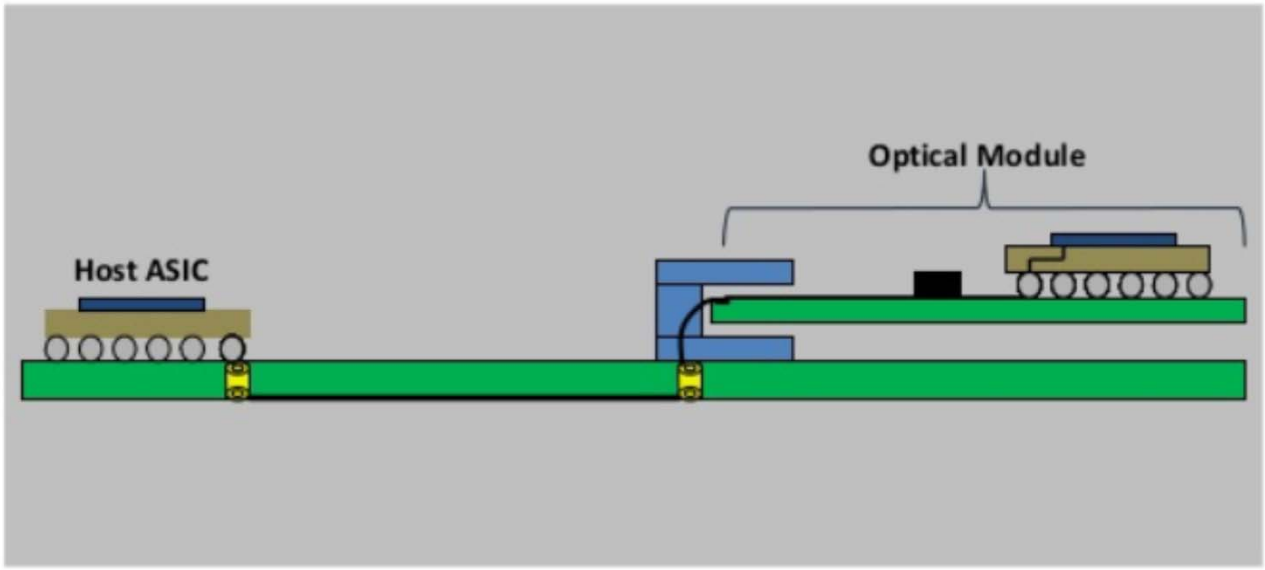


Host Card Electrical Link



Source: Brocade

OIF CEI-112-VSR Electrical Chip to Module Interfaces – 56Gbaud (PAM4)

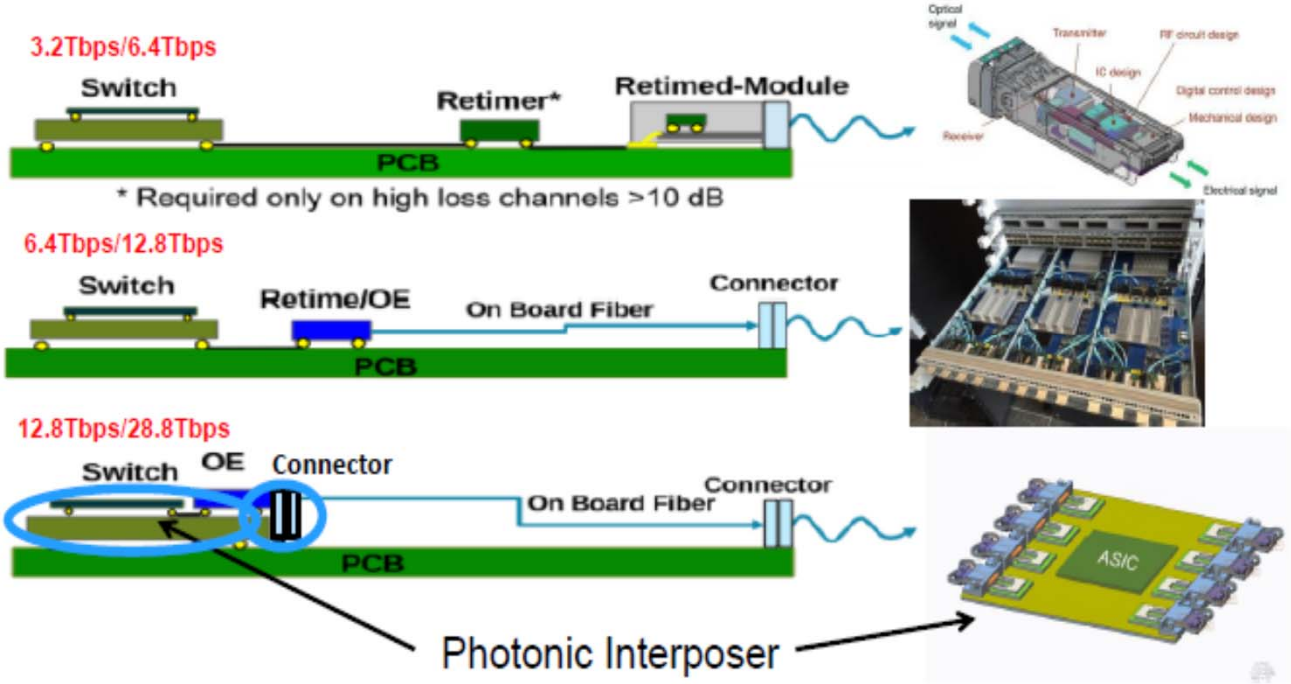


Source: Arista

Source: Arista

Evolution of Optical Interconnects

Legacy



Source: Corning

Contact Info

Doug Coleman
800 17th Street NW
Hickory, NC 28601
828-901-5580
doug.coleman@corning.com