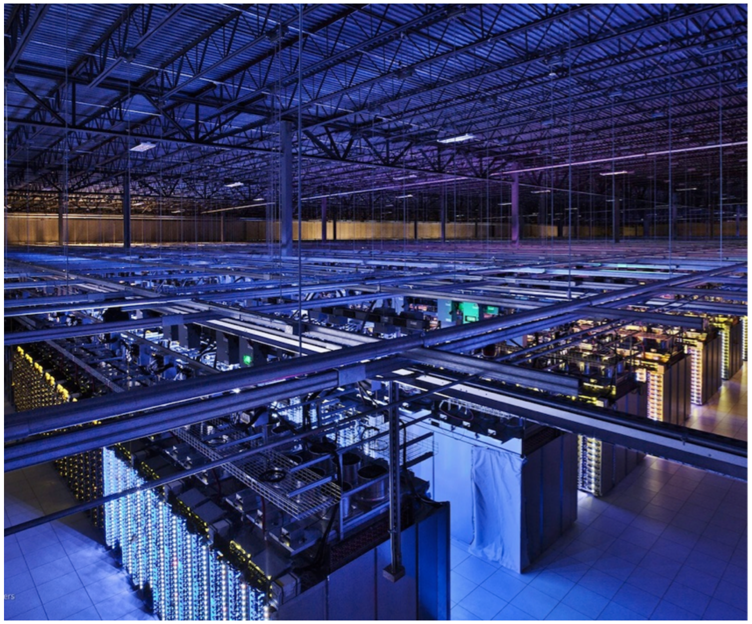
Scaling the Cloud Network

Andreas Bechtolsheim
Chairman, Arista Networks Inc

The World has Moved to the Cloud

Billions of Smartphones



Millions of Servers in the Cloud

Creating the Hyper-scale Datacenter Era

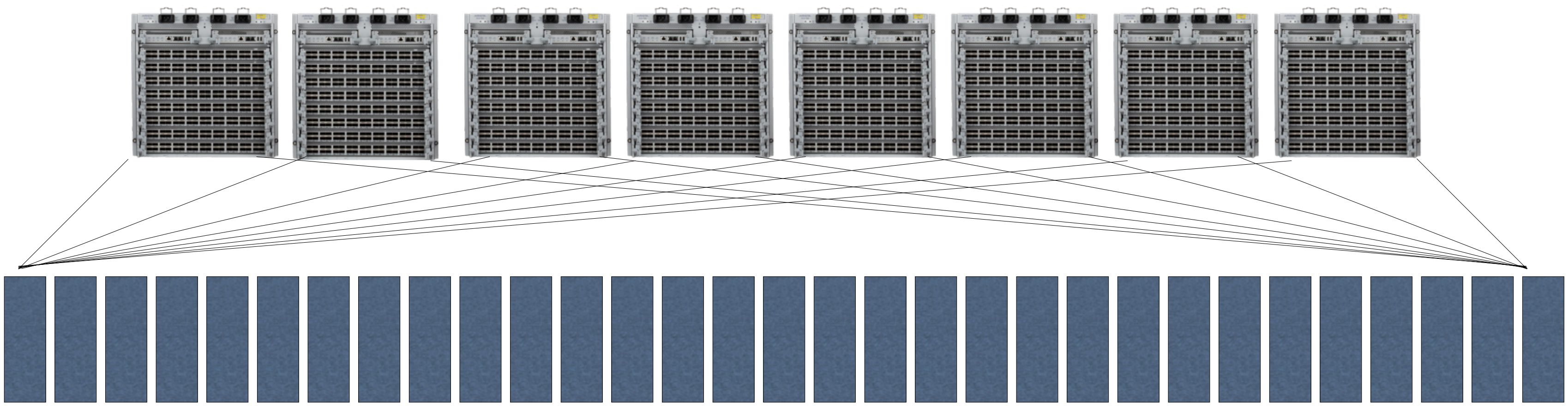
Hyper-scale Cloud Network Challenge

How do you interconnect 100,000s of servers such that cloud applications can easily scale?

Idealized Cloud Network

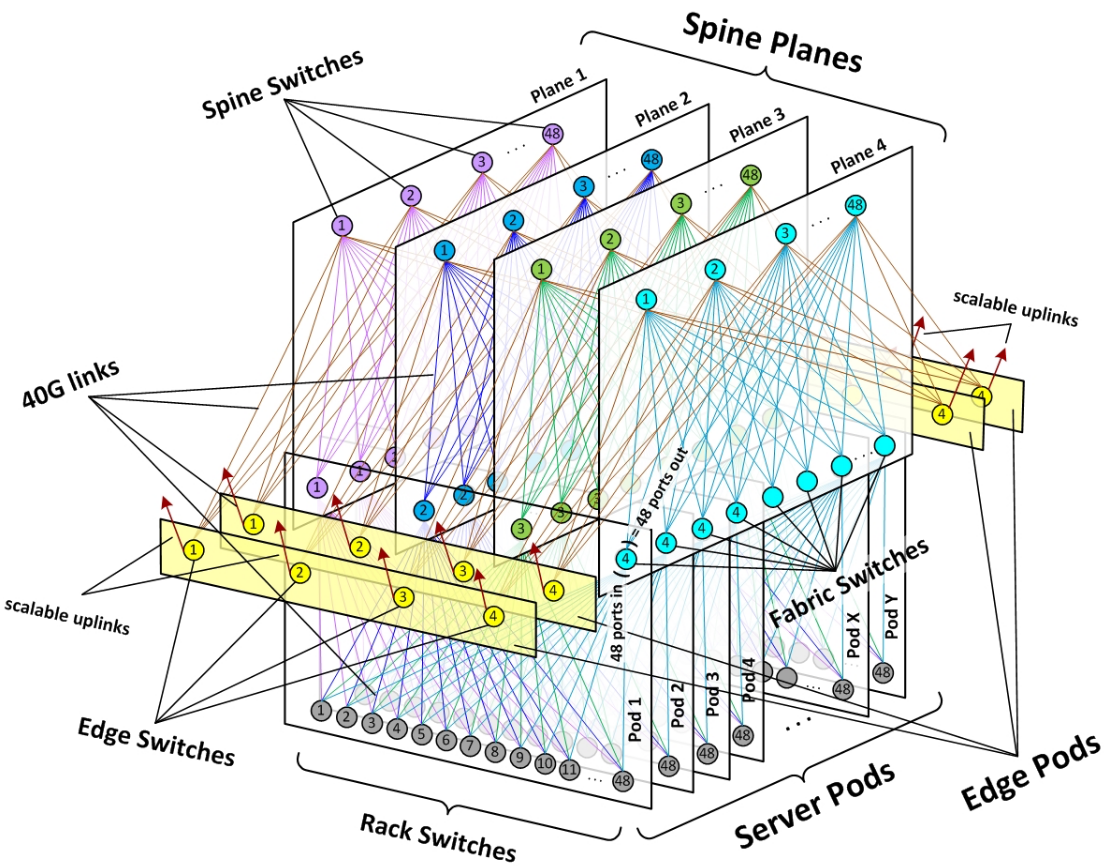
- **Ideal cloud network is truly transparent to applications**
  - Predictable bandwidth and low latency between all servers
  - 10+ Gbps Bandwidth/server, a few microseconds latency
- **This avoids the need for data placement**
  - Compute can be anywhere, data can be anywhere
  - Location does not matter since all servers are equal distant
- **Old approach was to divide datacenter into clusters**
  - Creates a significant burden on application developers
  - It was clear quickly that this was not practical

Leaf-Spine Network Architecture



Consistent bandwidth and latency from any server to any server, allowing applications to scale across the entire data center

Facebook Multi-Level Leaf-Spine Fabric



Layer3 From ToR to Edge

ECMP Load Balancing

Flow based Hashing

Large number of flows

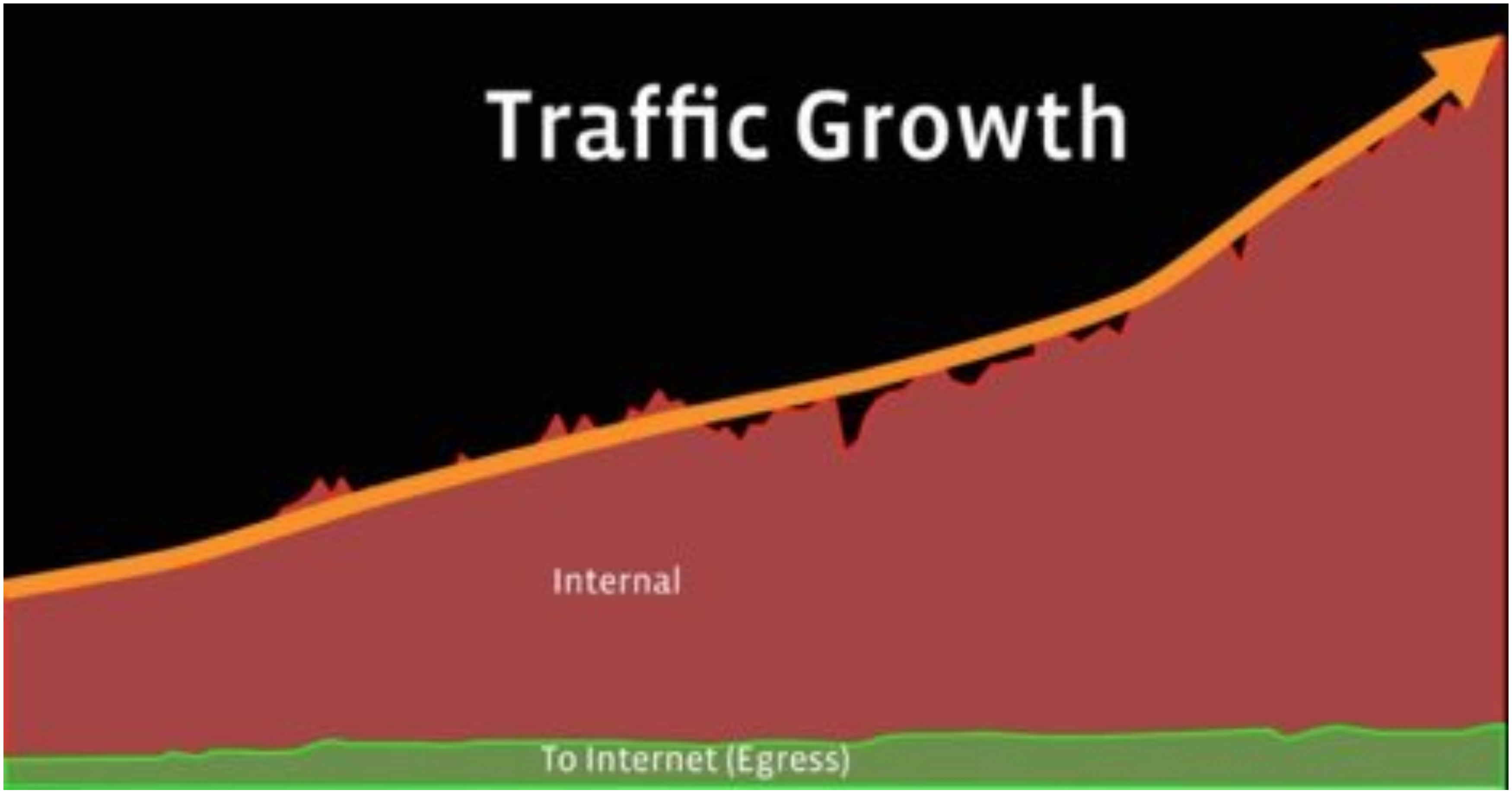
40G -> 100G -> 400G

10X speedup in 5 years

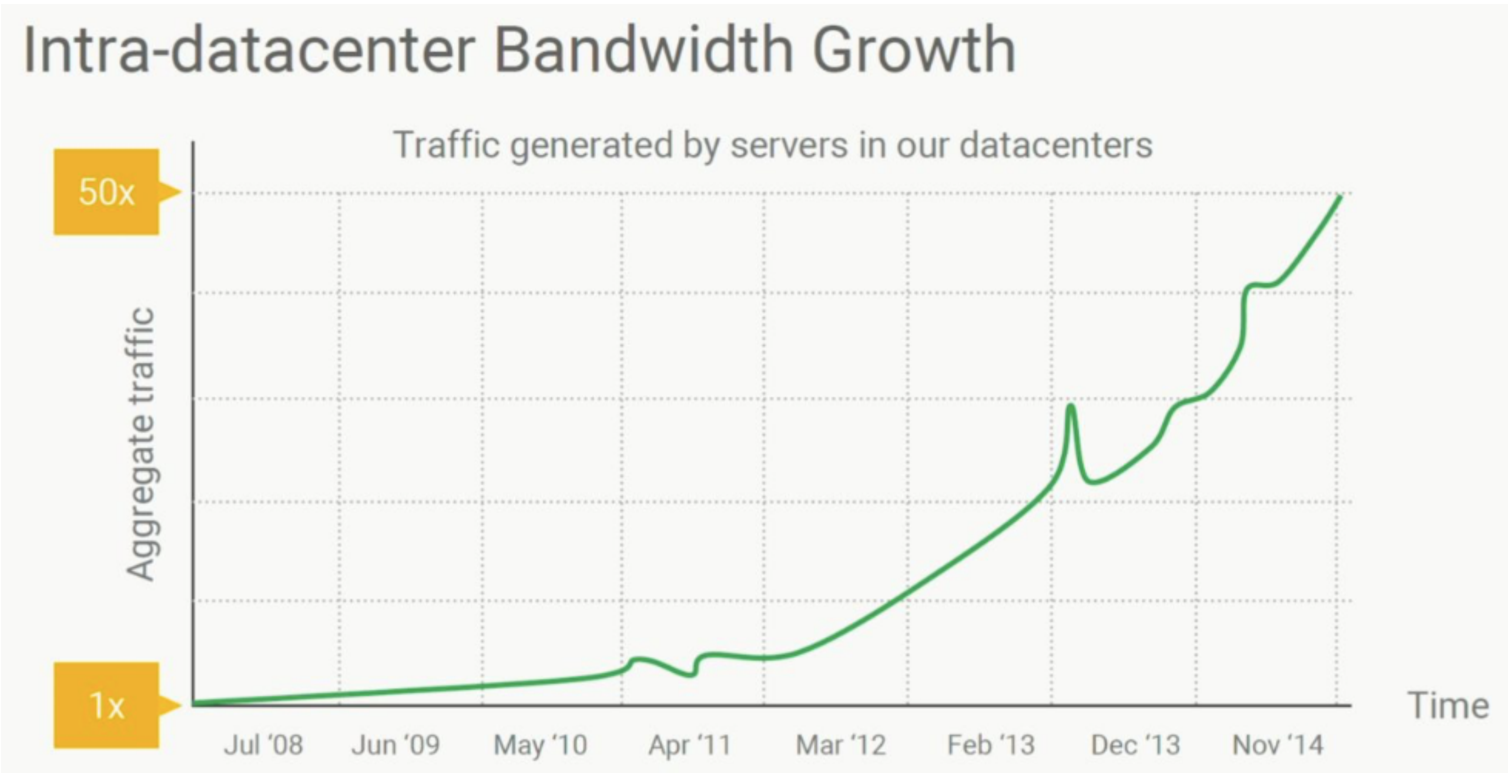
Consistent Performance

No more clusters

Growth in Cloud Network Bandwidth at Facebook



Cloud Network Bandwidth Demand Doubling/Year



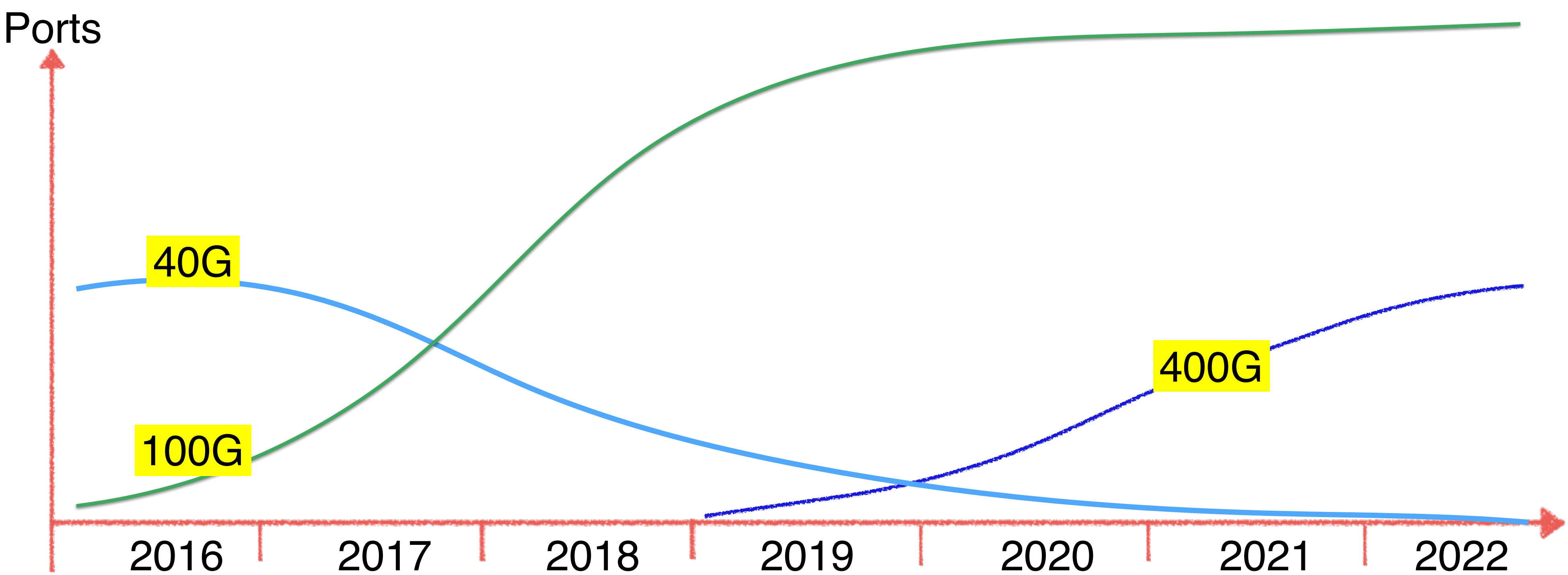
Source: Urs Hoelzle, Google

Driven by Video, AI and ML

The Easiest Way to go Faster is to go Faster

Ethernet Speed Transitions are the easiest way to scale the throughput of data center networks, in particular hyper-scale cloud networks

40G - 100G - 400G Switch Port Transition



100G has passed 40G Ethernet in Ports end of 2017
400G Volume ramps in 2020, passes 100G bandwidth in 2022

Source: Dell'Oro Group Jan 2018 Ethernet Switching Forecast

400G Timeline

First 400G Switch silicon and 400G optics in lab now

Typically one year from first silicon to production release, allowing for one silicon spin on switch chip and optics

Ramping 400G optics is required for volume deployment

Nobody wants a replay of the 100G-CWDM4 experience

Volume availability of 400G optics expected in 2H2019

400G Ports Market Forecast (Dell'Oro Market Research)

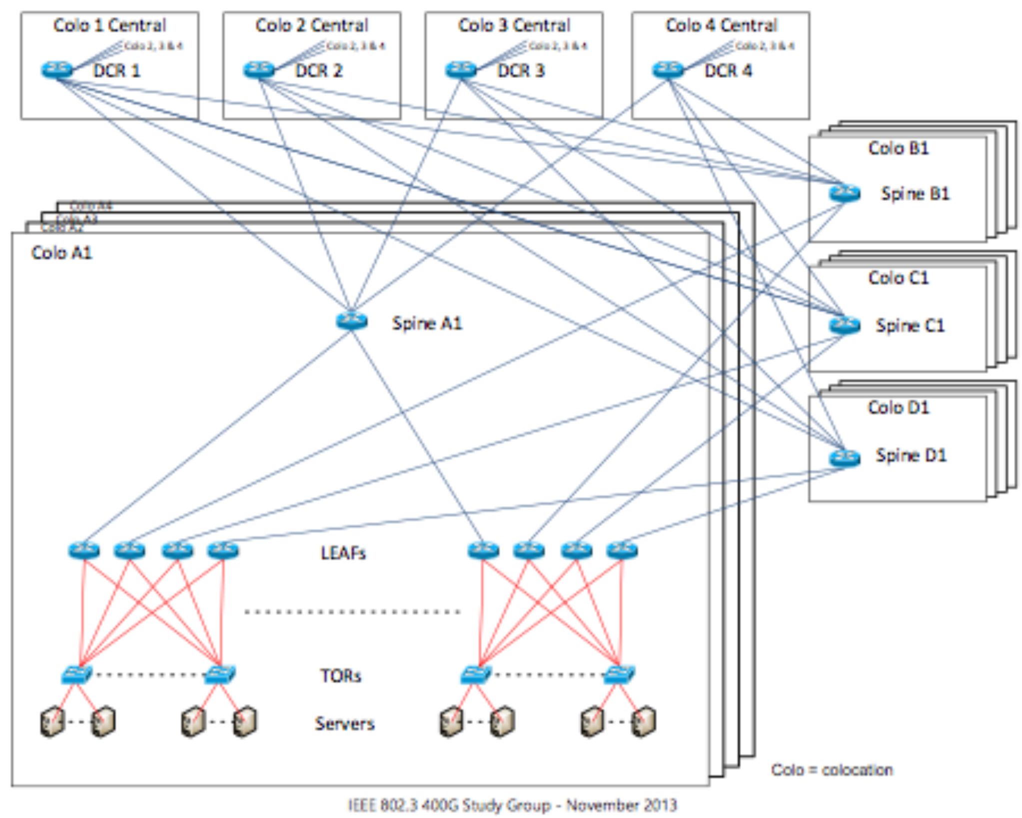
2019: 500K

2020: 3M

2021: 5M

400G In the Next-generation Cloud Network

Interface	Reach
400G-ZR	≤ 10 - 80 km > 100 km
400G-LR4	≤ 1,000 m
400G-FR4 400G-DR4	≤ 500 m
400G-AOC	≤ 20 m
400G-CR8 8x50G-CR	≤ 3 m



Metro/Core (DWDM)

Infrastructure designed to use a single data rate (X)

TARGET MARKET FOR 400G

Server links are a subset of X

Source: Brad Booth and Tom Issenhuth Microsoft, IEEE 802.3bs 400G

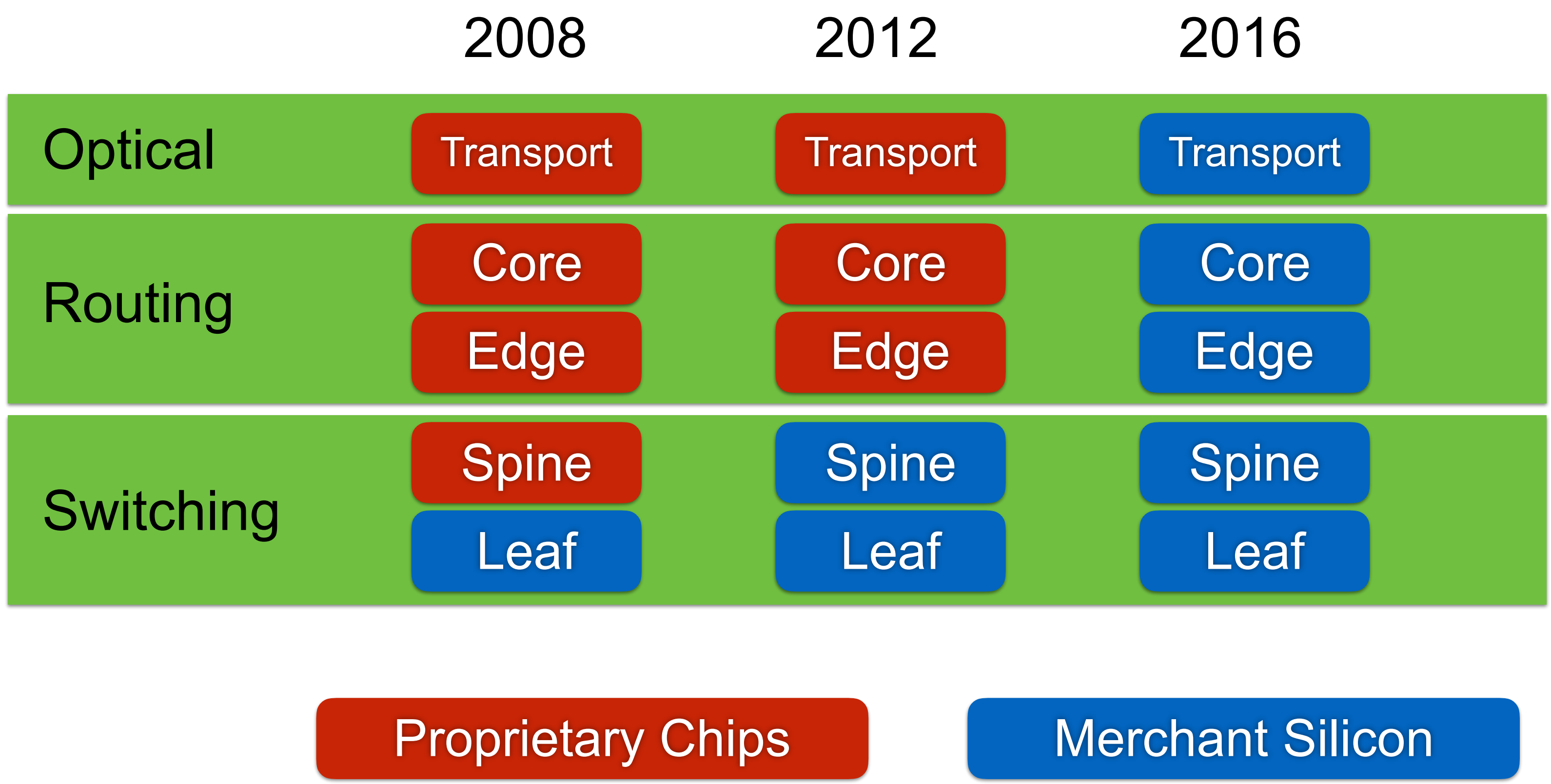
400G Use Cases

No Single 400G optics technology addresses all market requirements
In a hyper scale cloud data center, need at least the following:

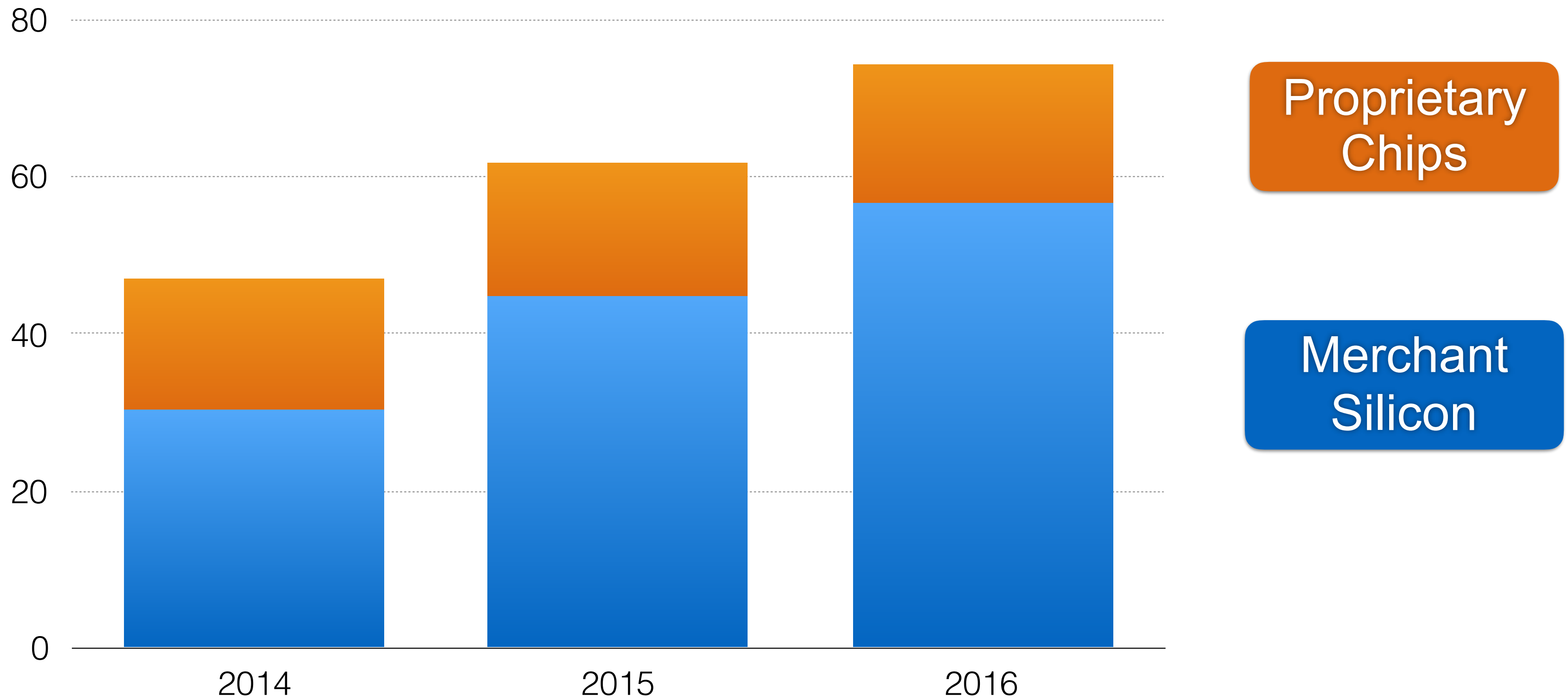
1. Copper cables for TOR-SERVER (3m max)
2. 400G-SR8 or AOC cables for TOR-LEAF (30m max)
3. 400G-DR4 or 400G-FR4 for LEAF-SPINE (500m - 2km)
4. 400G-LR8 or 400G-CWDM8 for Campus Reach (10km)
5. 400G-ZR for Metro Reach DCO (40km-100km)

Merchant Switch Silicon and Optics

The Expanding Merchant Silicon Roadmap



Merchant Silicon Driving Network Growth

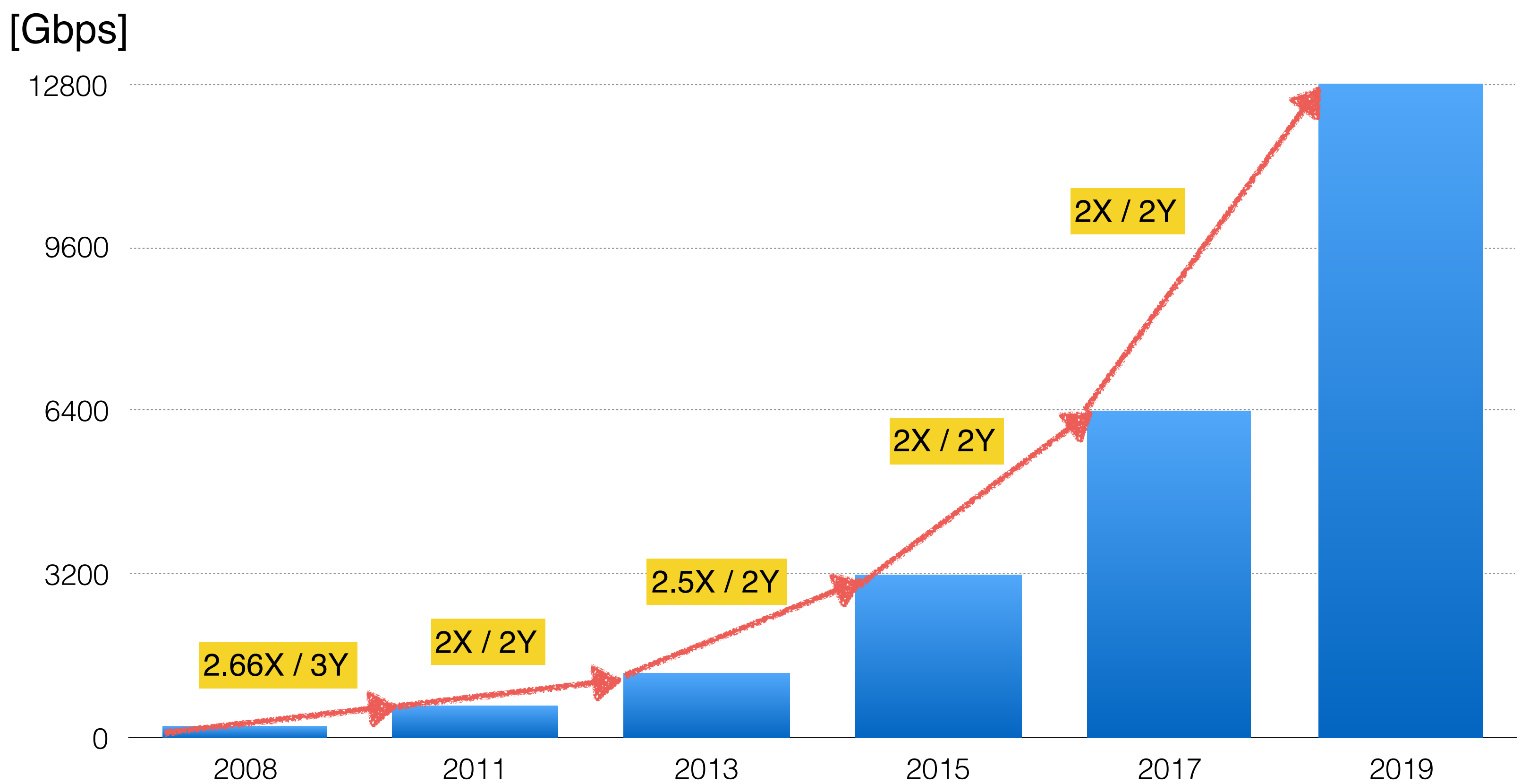


Source: The 650 Group, Jan 2017

Merchant Silicon Leading Industry in Performance

2008: First ultra-low latency 24-port 10G single chip
2010: First Large Buffer 10G Chip with VOQ Fabric
2011: First 64-port 10G single chip switch
2012: First 32-port 40G single chip
2013: First Large Buffer 40G Chip with VOQ Fabric
2015: First 32-port 100G single chip
2016: First Router 100G Chip with VOQ Fabric
2017: First 64-port 100G single chip
2018: First 32-port 400G single chip

Switch Silicon Bandwidth Growth



Switch Silicon Speed Transitions

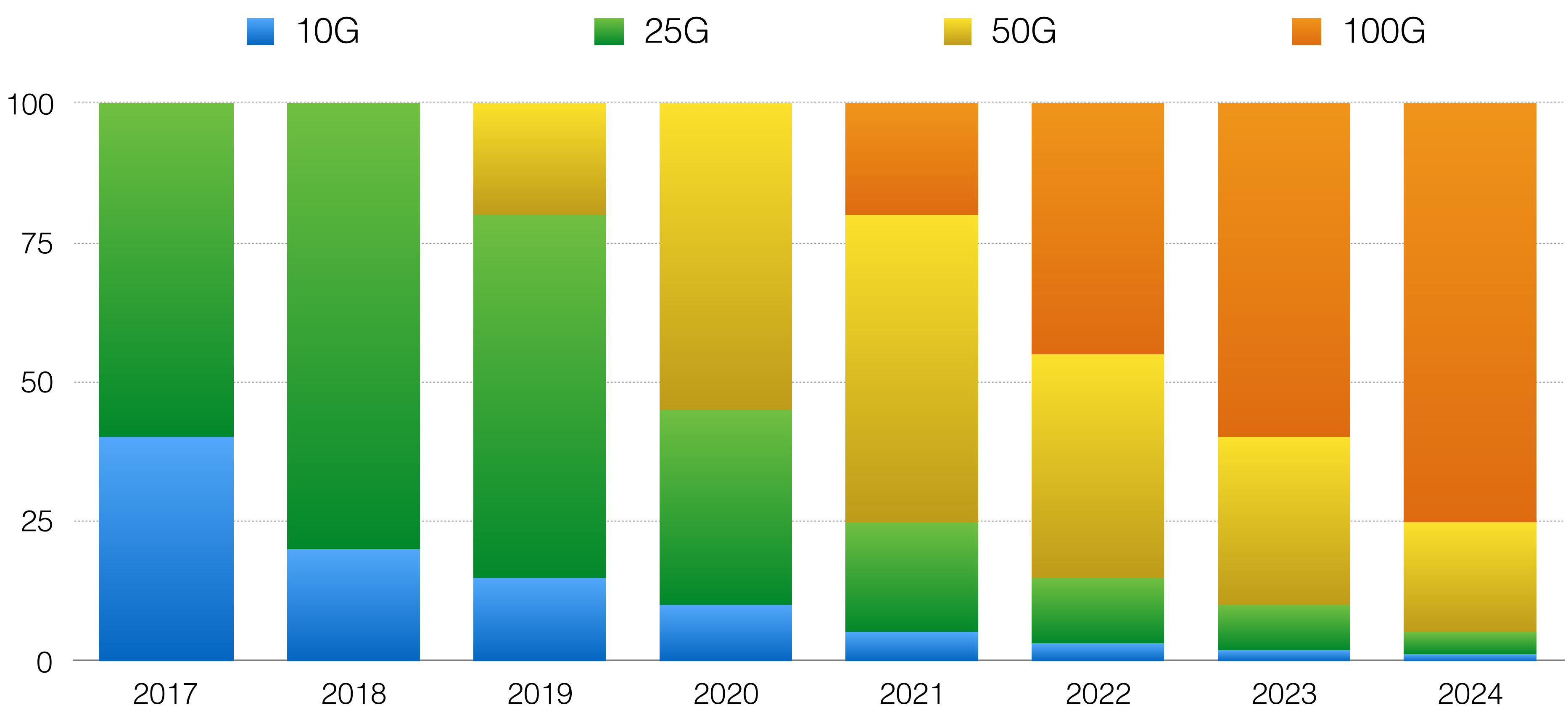
Lanes	10Gbps	25Gbps	50Gbps	100Gbps	
1X	10G	25G	50G	100G	Server Interface
2X	—	50G	100G	200G	
4X	40G	100G	200G	400G	Leaf-Spine Interface
8X	—	—	400G	800G	
First Product	2012	2016	2019	2021	

4 Years — 3 Years — 2 Years

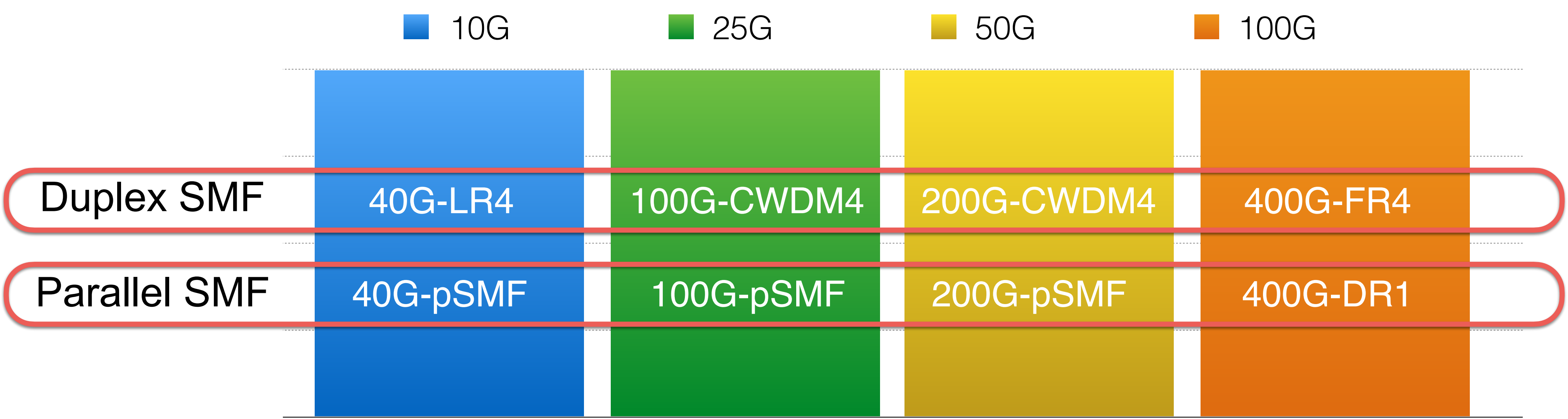
Why are SERDES transitions so Important?

1. They are the easiest way to scale switch performance
2. They drive Optics Standards and the Optics Ecosystem
3. Next Serdes Speed replaces previous one fairly quickly

SERDES Speed Transition Over the Years [% Mix]



Four-Lambda SMF Optics Transitions



The relentless march of Merchant Silicon drives rapid Transitions

The Three Most Important 400G Optics Modules for SMF

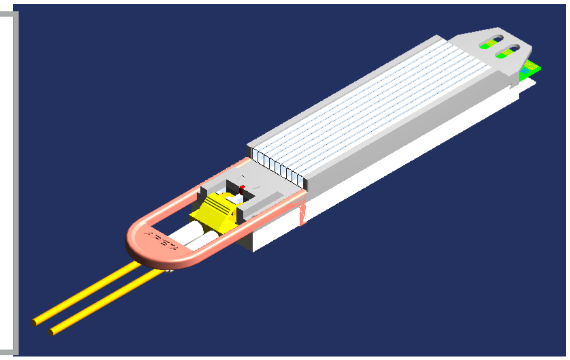
400G-DR4

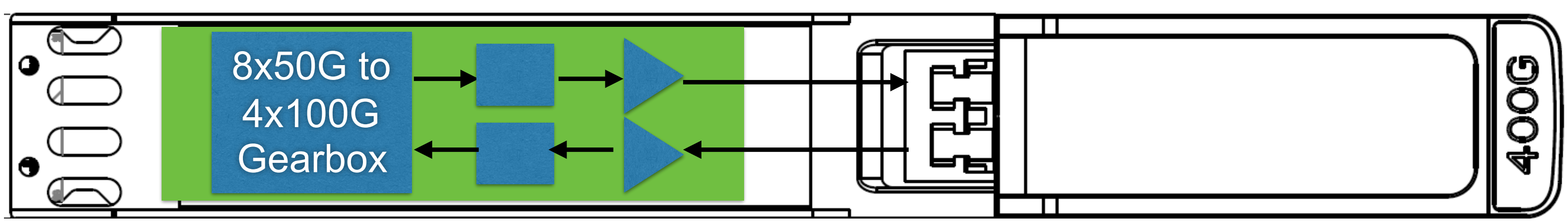
400G Over pSFM (8 Fibers)

500m Reach

MTP Parallel Fiber Connector

Estimated Power: 8W in 2020





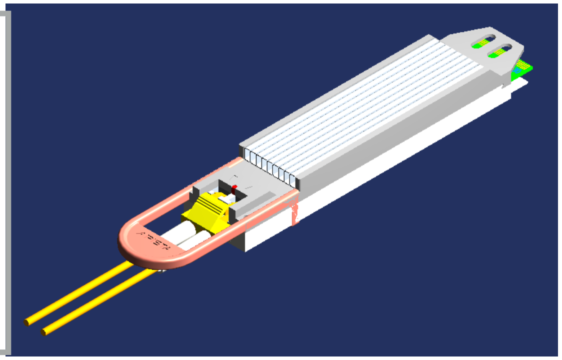
Works across same Fibre Plant as 100G-pSMF today
400G-DR4 can be split into four 100G-DR ports

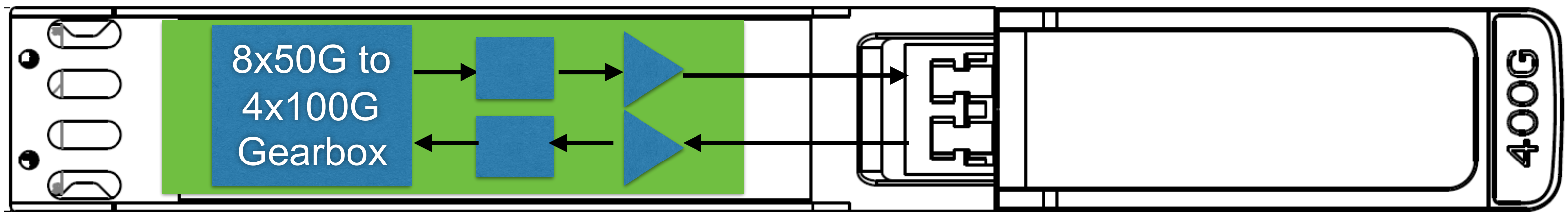
400G-FR4

400G Over Duplex Fiber

- 2km Reach (10km with LR4)
- Standard LC Fiber Connector

Estimated Power: 8W in 2020





Works across same fiber plant as 100G-CWDM4 today

ARISTA

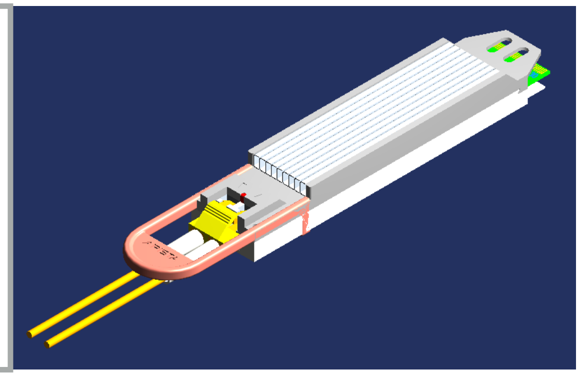
400G-ZR: 100km Reach DCO

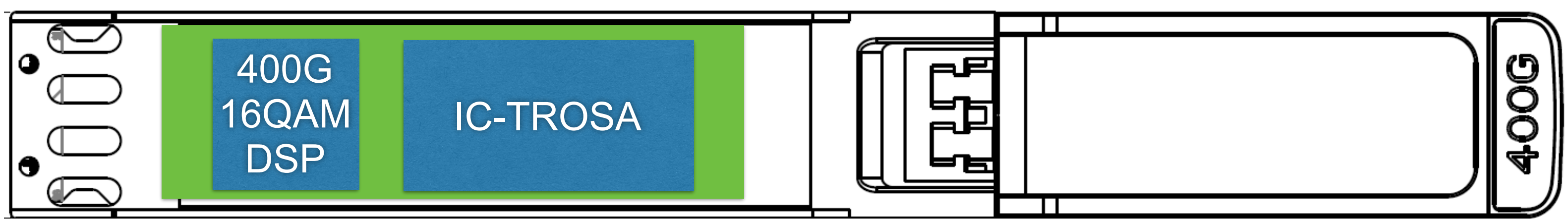
400G-16QAM DSP + Coherent Laser

20+ Terabits bandwidth per dark fiber

Pluggable Form Factor, 15W Power

Plugs into standard Switch Router Port





400G Coherent at the same port density as other Datacenter Optics

Three Key Optics Transition for 400G SMF

FIBER	100G	400G
500m pSMF (8F)	100G-pSMF	400G-DR4
2km SMF Duplex	100G-CWDM4	400G-FR4
100km Reach	100G-ColorZ	400G-ZR

Three Key Benefits of making these Optics Transitions:

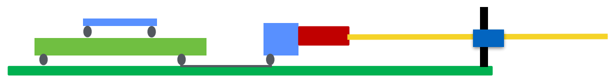
1. **4X Bandwidth** without Change to Fiber Infrastructure
2. **Forwards Compatible** with 100G Lane Switch Chips
3. **High Volume** drives best availability and economics

Co-packaged Optics

ARISTA

Placement of Optics

- Pluggable optics



- Move optics on-board (COBO)

- Optics Co-packaged with Switch Chip

Co-packaged optics enable much lower-power electrical I/O with a potential 30% power reduction at the system level

Co-Packaged Optics Switch

Packaging Study

(not an actual product)

51.2 Tbps in 1U

128 400G ports

Four Optical Tiles

128 lanes each

Four Laser Sources

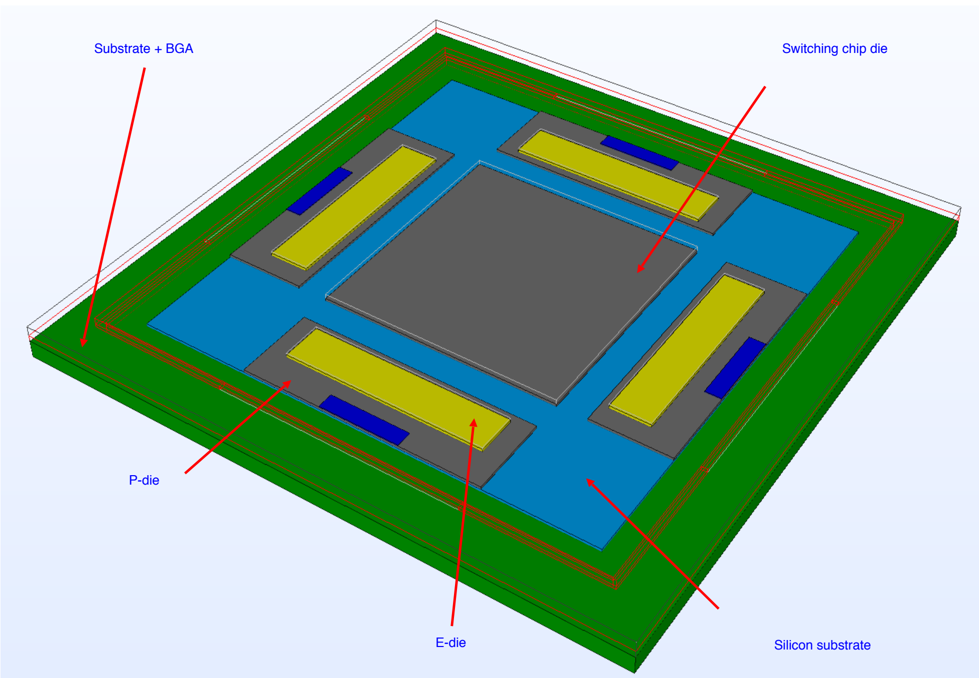
driving 128 lanes each

Double Density

compared to pluggable

Image Courtesy of Luxtera

Co-Packaged Optics Benefits



Lower Power / Higher Density

Eliminate high-power SERDES I/O

Cost Advantages

Sub-linear scaling of cost/channel

Greater Reliability

Separating out the laser sources

Co-Packaged Optics Challenges

Technical Challenges

Picking the best low-power electrical Interface

Multi-vendor Standardization

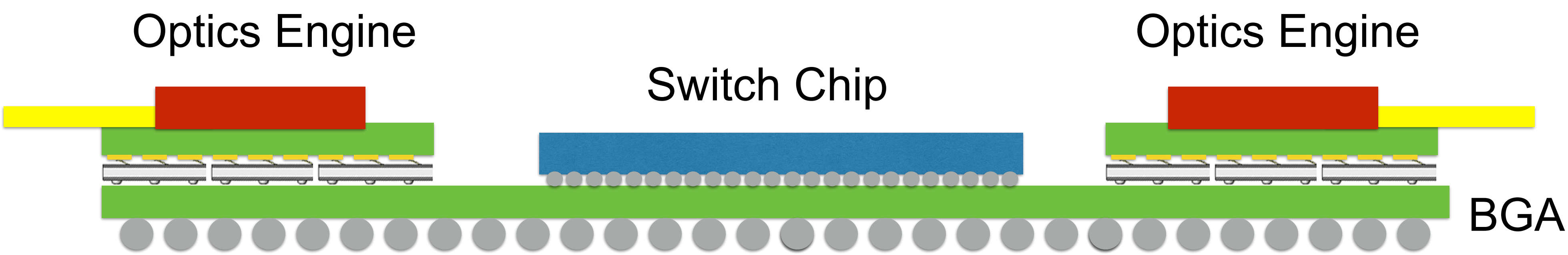
Need to enable multiple vendors to work together

Supply Chain (Switch Chip, Optics, CM)

Who owns the yield at each manufacturing stage

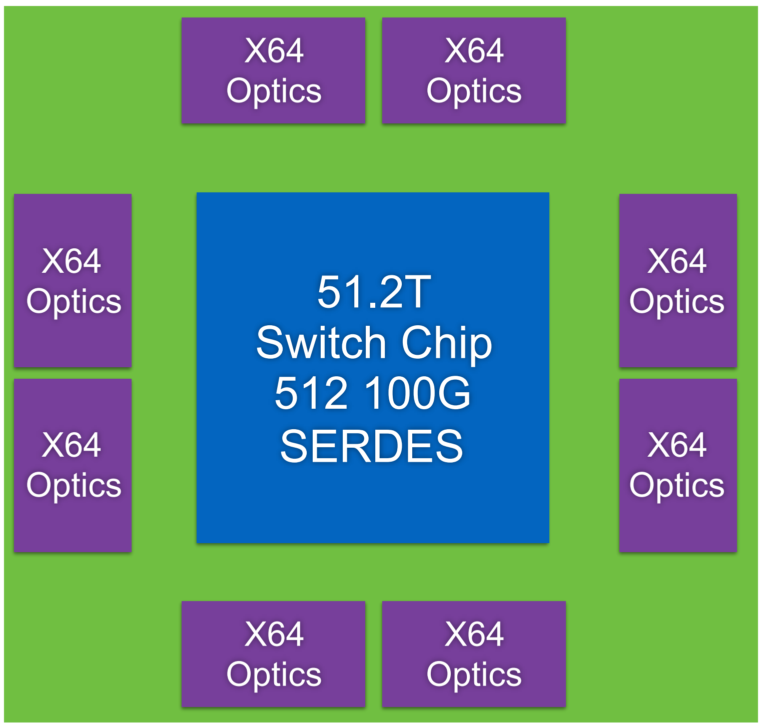
ARISTA

Solution: Electrical Interposer Connector for Optics



BGA/LCA Array Connector, 0.25mm thick

Interposer Solves the Co-Packaging Problem



Makes Product Manufacturable

High yield merge of fully tested Optics and fully tested switch chip at the CM

Enables Repairability

Failed Optics can be replaced

In manufacturing or even in the field

Supports Configurability

Different Optics can be Configured

For example: 400G-DR4, FR4, LR4, etc

Co-Packaged Optics Summary

Workable Solution Must Solve all Problems

Manufacturability, Serviceability, Configurability

Standardized Electrical Connector is Key

Easiest solution to the above challenges

Need Multi-Vendor Standardization

Define electrical interface and physical form factor

This is a multi-year project, let’s start now

Optics and Standards

Standards Drive New Optics Schedules

Need Standards to drive Volumes

Without Volume, Economics don't work

Silicon and Optics Developments take a long time

Typical 2-3 years from start of product development

Standards are gating the Speed of Progress

Can't start product development without a standard

2-3Y

Specification Complete
- Design
- Verification
- Shuttle Tapeout
- Fab Time
- Silicon Validation
- Second Tapeout
- Silicon Validation
- System Validation
- Production Ramp

Product FCS

Time needed to develop new Optics Modules is 2-3 Years

ARISTA

IEEE 802.3 LAN Standards Group

Table 1

Standard	Year	Description
802.3	1983	10BASE5 10 Mbit/s (1.25 MB/s) over thick coax. Same as Ethernet II (above) except Type field is replaced by Length, and an 802.2 LLC header follows the 802.3 header. Based on the CSMA/CD Process.
802.3a	1985	10BASE2 10 Mbit/s (1.25 MB/s) over thin Coax (a.k.a. thinnet or cheapernet)
802.3b	1985	10BROAD36
802.3c	1985	10 Mbit/s (1.25 MB/s) repeater specs
802.3e	1987	1BASE5 or StarLAN
802.3d	1987	Fiber-optic inter-repeater link
802.3i	1990	10BASE-T 10 Mbit/s (1.25 MB/s) over twisted pair
802.3j	1993	10BASE-F 10 Mbit/s (1.25 MB/s) over Fiber-Optic
802.3u	1995	100BASE-TX, 100BASE-T4, 100BASE-FX Fast Ethernet at 100 Mbit/s (12.5 MB/s) with autonegotiation
802.3x	1997	Full Duplex and flow control; also incorporates DIX framing, so there's no longer a DIX/802.3 split
802.3z	1998	1000BASE-X Gbit/s Ethernet over Fiber-Optic at 1 Gbit/s (125 MB/s)
802.3y	1998	100BASE-T2 100 Mbit/s (12.5 MB/s) over low quality twisted pair
802.3-1998	1998	A revision of base standard incorporating the above amendments and errata
802.3ac	1998	Max frame size extended to 1522 bytes (to allow "Q-tag") The Q-tag includes 802.1Q VLAN information and 802.1p priority information.
802.3ab	1999	1000BASE-T Gbit/s Ethernet over twisted pair at 1 Gbit/s (125 MB/s)
802.3ad	2000	Link aggregation for parallel links, since moved to IEEE 802.1AX
802.3ae	2002	10 Gigabit Ethernet over fiber; 10GBASE-SR, 10GBASE-LR, 10GBASE-ER, 10GBASE-SW, 10GBASE-LW, 10GBASE-EW
802.3-2002	2002	A revision of base standard incorporating the three prior amendments and errata
802.3af	2003	Power over Ethernet (15.4 W)
802.3ak	2004	10GBASE-CX4 10 Gbit/s (1,250 MB/s) Ethernet over twinaxial cables
802.3ah	2004	Ethernet in the First Mile
802.3-2005	2005	A revision of base standard incorporating the four prior amendments and errata.
802.3aq	2006	10GBASE-LRM 10 Gbit/s (1,250 MB/s) Ethernet over multimode fiber
802.3an	2006	10GBASE-T 10 Gbit/s (1,250 MB/s) Ethernet over unshielded twisted pair (UTP)
802.3as	2006	Frame expansion
802.3au	2006	Isolation requirements for Power over Ethernet (802.3-2005/Cor 1)
802.3ap	2007	Backplane Ethernet (1 and 10 Gbit/s (125 and 1,250 MB/s) over printed circuit boards)
802.3aw	2007	Fixed an equation in the publication of 10GBASE-T (released as 802.3-2005/Cor 2)
802.3-2008	2008	A revision of base standard incorporating the 802.3an/ap/aq/as amendments, two corrigenda and errata. Link aggregation was moved to 802.1AX.
802.3av	2009	10 Gbit/s EPON

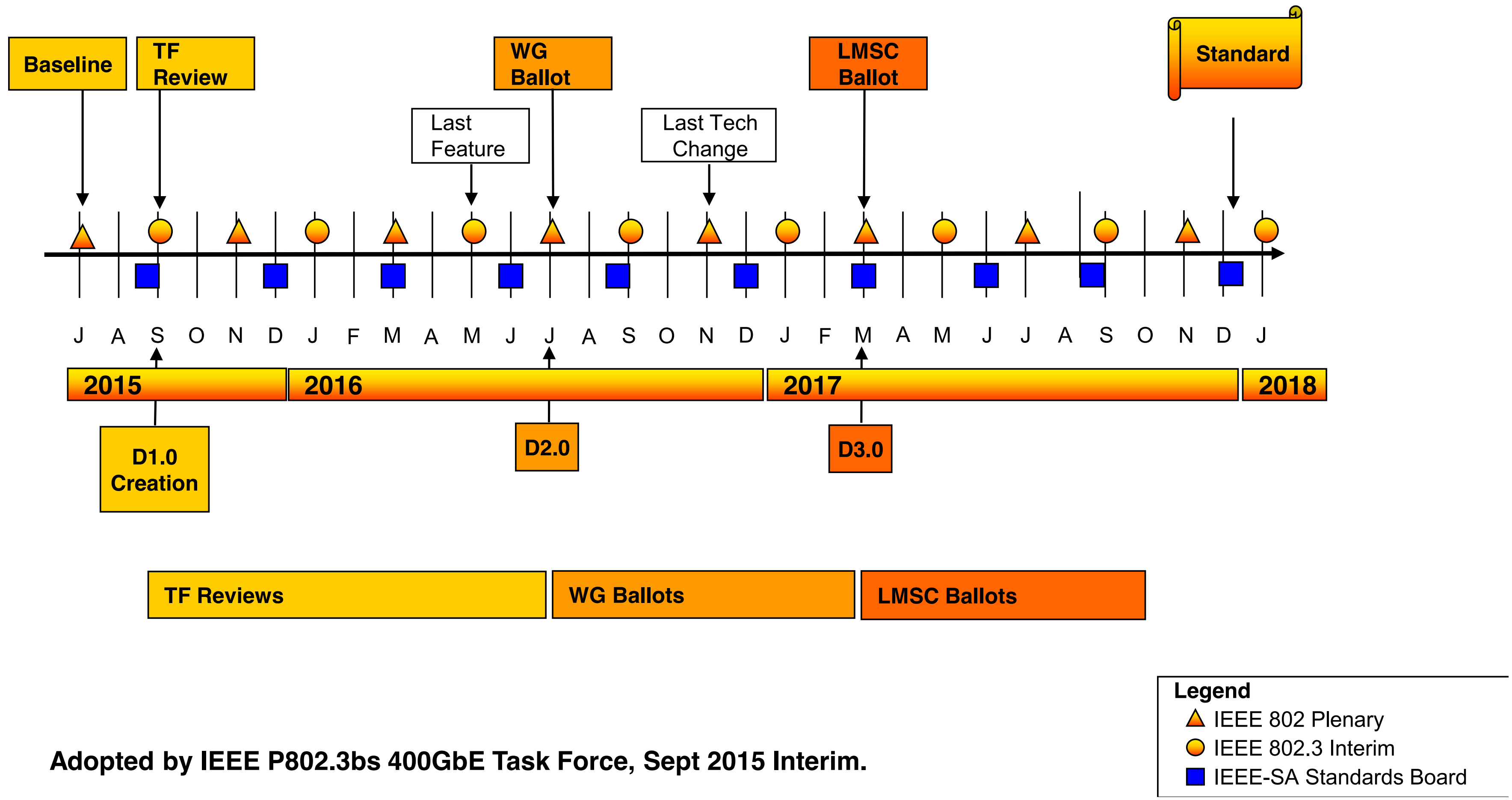
IEEE

IEEE 802.3 LAN Standards Group (cont)

Table 1

Standard	Year	Description
802.3ba	2010	40 Gbit/s and 100 Gbit/s Ethernet. 40 Gbit/s over 1 m backplane, 10 m Cu cable assembly (4×25 Gbit or 10×10 Gbit lanes) and 100 m of MMF and 100 Gbit/s up to 10 m of Cu cable assembly, 100 m of MMF or 40 km of SMF respectively
802.3az	2010	Energy-efficient Ethernet
802.3bd	2010	Priority-based Flow Control. An amendment by the IEEE 802.1 Data Center Bridging Task Group (802.1Qbb) to develop an amendment to IEEE Std 802.3 to add a MAC Control Frame to support IEEE 802.1Qbb Priority-based Flow Control.
802.3.1	2011	MIB definitions for Ethernet. It consolidates the Ethernet related MIBs present in Annex 30A&B, various IETF RFCs, and 802.1AB annex F into one master document with a machine readable extract. (workgroup name was P802.3be)
802.3bg	2011	Provide a 40 Gbit/s PMD which is optically compatible with existing carrier SMF 40 Gbit/s client interfaces (OTU3/STM-256/OC-768/40G POS).
802.3bf	2011	Provide an accurate indication of the transmission and reception initiation times of certain packets as required to support IEEE P802.1AS.
802.3-2012	2012	A revision of base standard incorporating the 802.3at/av/az/ba/bc/bd/bf/bg amendments, a corrigenda and errata.
802.3bk	2013	This amendment to IEEE Std 802.3 defines the physical layer specifications and management parameters for EPON operation on point-to-multipoint passive optical networks supporting extended power budget classes of PX30, PX40, PRX40, and PR40 PMDs.
802.3bj	2014 (June)	Define a 4-lane 100 Gbit/s backplane PHY for operation over links consistent with copper traces on "improved FR-4" (as defined by IEEE P802.3ap or better materials to be defined by the Task Force) with lengths up to at least 1 m and a 4-lane 100 Gbit/s PHY for operation over links consistent with copper twinaxial cables with lengths up to at least 5 m.
802.3bw	2015[4]	100BASE-T1 – 100 Mbit/s Ethernet over a single twisted pair for automotive applications
802.3bm	2015	100G/40G Ethernet for optical fiber
802.3-2015	2015	802.3bx – a new consolidated revision of the 802.3 standard including amendments 802.3bk/bj/bm
802.3bp	2016 (June)[2]	1000BASE-T1 – Gigabit Ethernet over a single twisted pair, automotive & industrial environments
802.3bn	2016	10G-EPON and 10GPASS-XR, passive optical networks over coax
802.3bz	2016 (Sep.)[7]	2.5GBASE-T and 5GBASE-T – 2.5 Gigabit and 5 Gigabit Ethernet over Cat-5/Cat-6 twisted pair
802.3bq	2016 (June)[3]	25G/40GBASE-T for 4-pair balanced twisted-pair cabling with 2 connectors over 30 m distances
802.3by	2016 (June)[5]	Optical fiber, twinax and backplane 25 Gigabit Ethernet[6]
802.3bu	2016	Power over Data Lines (PoDL) for single twisted-pair Ethernet (100BASE-T1)
802.3br	2016	Specification and Management Parameters for Interspersing Express Traffic
802.3bs	2017 (Dec.)	200GbE (200 Gbit/s) over single-mode fiber and 400GbE (400 Gbit/s) over optical physical media
802.3cc	2017 (Dec)	25 Gbit/s over Single Mode Fiber
802.3bv	2017	Gigabit Ethernet over plastic optical fiber (POF)
802.3ce	2017 (March)	Multilane Timestamping
802.3cb	2018 (TBD)	2.5 Gb/s and 5 Gb/s Operation over Backplane
802.3cd	2018 (TBD)	Media Access Control Parameters for 50 Gbit/s, 100 Gbit/s, and 200 Gbit/s Operation
802.3bt	2018 (TBD)	Power over Ethernet enhancements up to 100 W using all 4 pairs balanced twisted-pair cabling
802.3cf	2018 (TBD)	YANG Data Model Definitions
802.3cg	2019 (TBD)	10 Mb/s Single Twisted Pair Ethernet
802.3ca	2019 (TBD)	100G-EPON – 25 Gbit/s, 50 Gbit/s, and 100 Gbit/s over Ethernet Passive Optical Networks

Sep 2015 Timeline for IEEE 802.3bs (400GigE)



Adopted by IEEE P802.3bs 400GbE Task Force, Sept 2015 Interim.

History of IEEE 802.3 Ethernet Standards

Ethernet Speed	PAR	Standard Ratified	Time (Years)
10 Mbps	1981	1983	2
100 Mbps	1992	1995	3
1 Gbps	1995	1998	3
10 Gbps	1999	2002	3
40/100 Gbps	2007	2010	3
400 Gbps	2014	2017	3

Problem: New Optics can’t wait for three years of standards process

Problems with IEEE 100G Optics Standards

IEEE 802.3ba (100G Ethernet) standardized two 100G optics:

- 100G-LR4 (10km reach duplex fiber) and 100G-SR10 (100m reach 10x10)
- Neither addressed the large cloud network market potential

IEEE 802.3bm (lower cost 100G optics standards) tried to correct this

- Proposed 4x25G 500m reach duplex SMF (100G-CWDM4) and parallel SMF
- After 2 years of meetings, neither proposal was accepted as an IEEE standard

IEEE Voting rules prevented standardization of the most common 100G Optics in use today

Similar Situation with 400G Optics

802.3bs Standard	Description	Reach	Comments
400G-SR16	16x25G lambda, 32-MMF	100m	Nobody will use this
400G-FR8/LR8	8x50G lambda, duplex SMF	2/10km	Limited Market Potential
400G-DR4	4x100G lamda, 8-SMF	500m	High-volume for pSMF

IEEE 802.3bs did not standardize the highest volume 400G optics for cloud, including 400G-FR4 and 400G-LR4

ARISTA

100GLambda
MULTI-SOURCE AGREEMENT

Home About Us News Promoters FAQs Contact Us

Source: www.100glambda.com

100G Lambda MSA SMF Optics Standards

Speed/Fiber	500m	2km	10km
100G Duplex Fiber	100G-DR	100G-FR	100G-LR
400G Parallel Fiber	400G-DR4	400G-DR4	TBD
400G Duplex Fiber	400G-FR4	400G-FR4	TBD

IEEE 802.3 Specs

100G Lambda MSA

Future Work

Timeline from announcement of 100G Lambda MSA to release of first set of specifications was four months (9/12/2017 to 1/9/2018)

How do Optics MSAs work?

- The outcome of any standards group activity can be predicted by (1) the group constituency and (2) its voting rules
- With MSAs, members have a shared goal to get a spec done. There are typically weekly meetings with active participation
- As a result, time lines become compressed. Most MSAs complete their specification work in a couple of months, not years.
- MSAs are driven by members that have shared goals There are no dissenting parties blocking progress

Optics MSA and Related Standard Efforts

400G Optics

- 4x100G-LAMBDA
- 400G-ZR
- 400G-CWDM8
- 400G-SR8
- 400G-SR4.2

100G Optics

- 100G-LAMBDA
- 100G-CWDM4
- 100G-PSM4

Form Factors

- OSFP
- QSFP-DD
- uQSFP
- D-SFP
- SFP-DD

Need Standards for everything not included in 802.3
One cannot build new products without a standard

Next-gen Optics Standards Summary

Standards for Next-gen 400G Optics are needed *now*

400G switch silicon is in the lab, products will ship in volume in 2019

MSAs are taking the initiative to create these standards

This is working well, specifically with the 100G Lambda MSA

Traditional Standards Bodies have not worked well for optics

Multi-year processes are simply too slow to make good choices

OCP can play a major role promoting and advocating optics standards that are good for cloud networks