400G and beyond

CCR Workshop - June 05, 2019

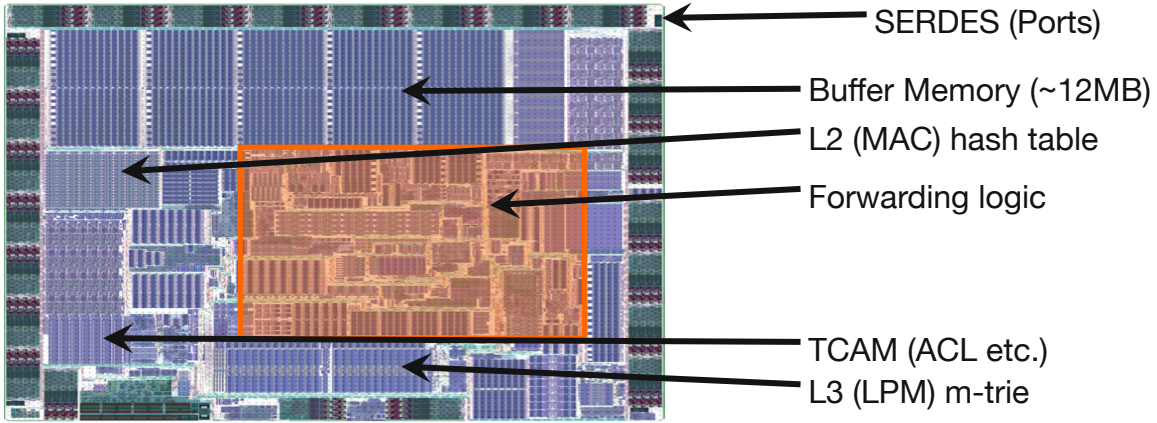
Davide Bassani <davide@arista.com>

What is happening on Silicon side?

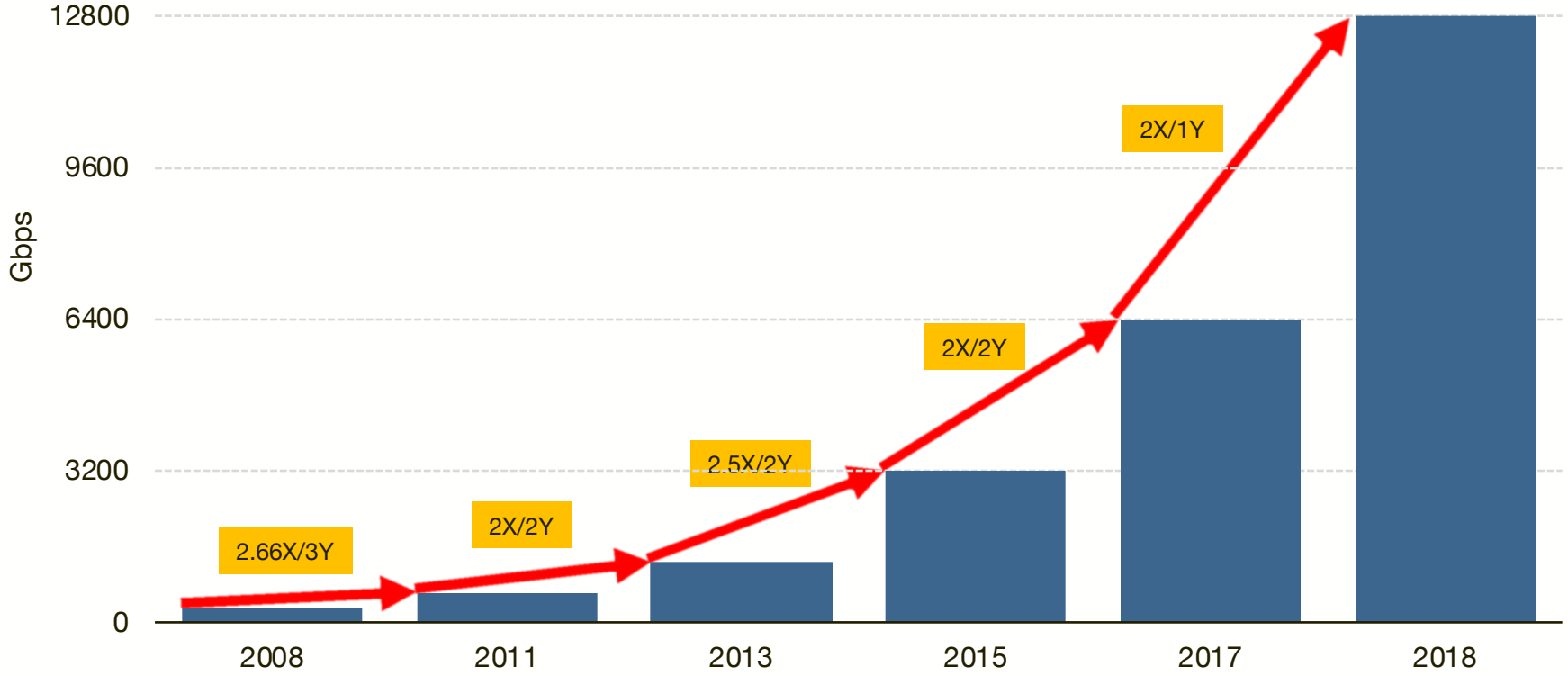
Merchant Silicon – Design trade-offs

Designing merchant silicon – multi-dimension challenge, silicon finite size

- Number of Pins and their speed (10G/25G) – defining the throughput requirements
- Pins for physical interfaces, fabric links, access to off-chip resources?
- Layer 2/3 forwarding logic and tunnel support (GRE, MPLS, VXLAN etc)
- Tables sizes, MAC table (Exact match table), ARP tables (LEM), LPM for Layer 3 tables



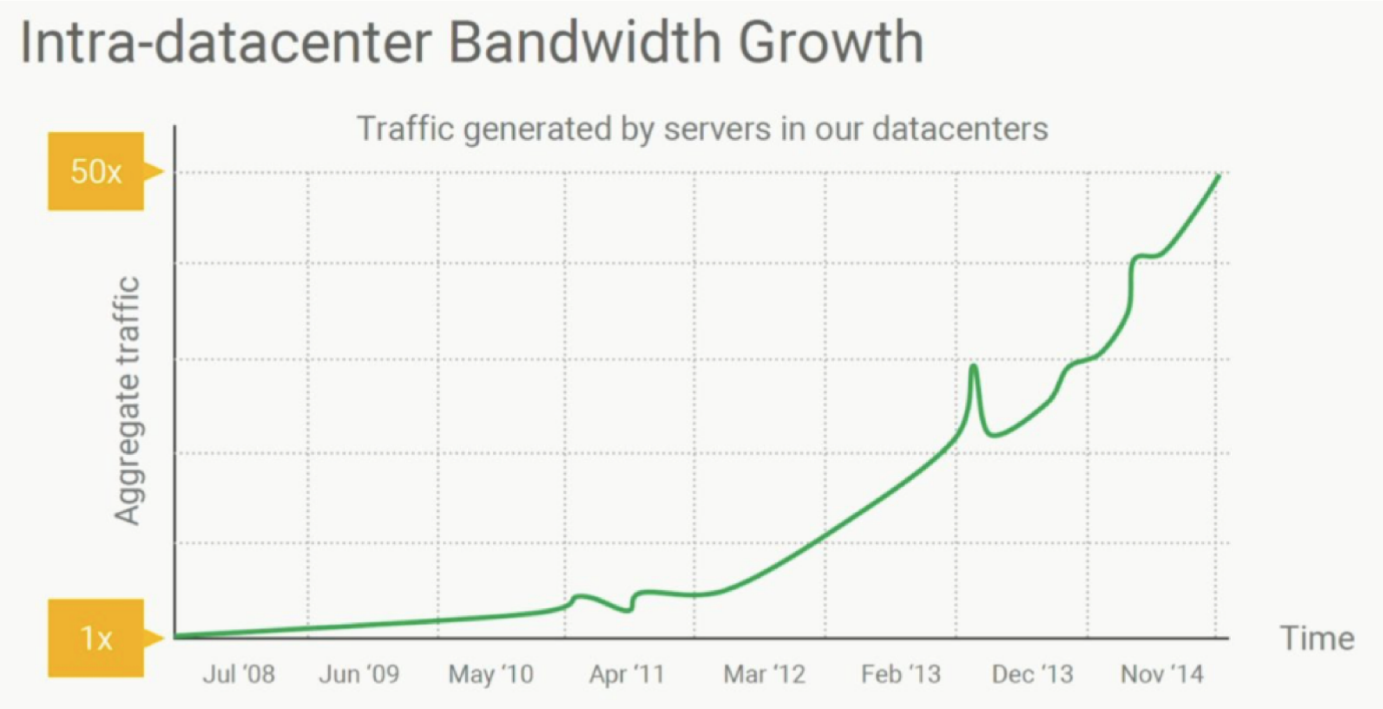
Merchant Silicon Switch Bandwidth Growth



ARISTA

OK for Silicon
But what about optics?

Cloud Network Bandwidth Demand Doubling/Year

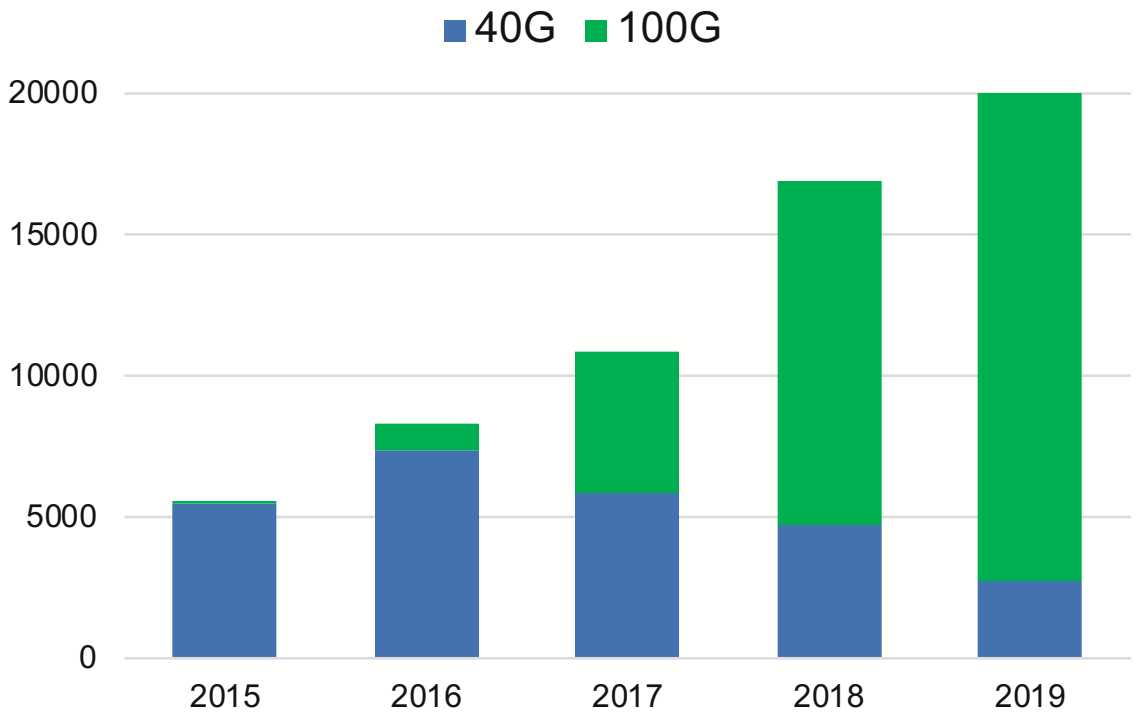


Source: Urs Hoelzle, Google, OFC 2017

Driven by Flash IO, Serverless Compute, AI and ML

ARISTA

40G to 100G Ethernet Transition

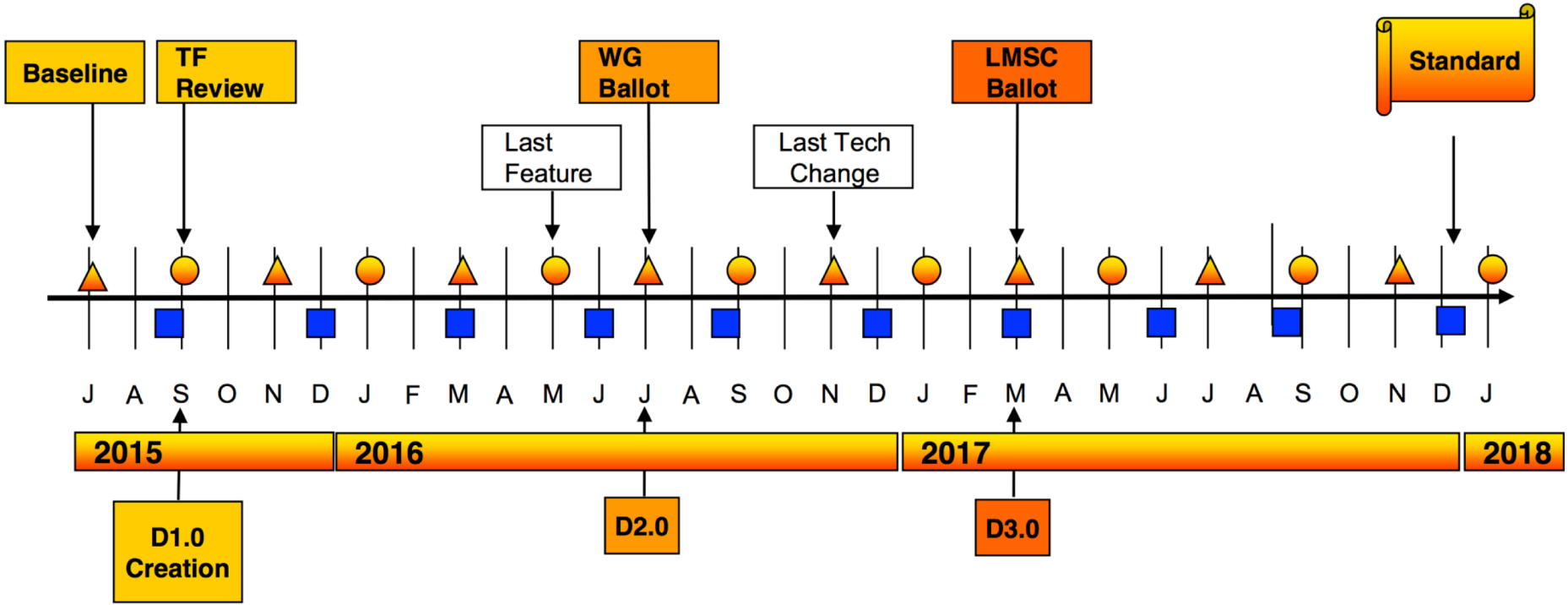


100G went from <10% to >50% in one year

Source: Dell'Oro Market Research, Ethernet Switch Update, July 2018

ARISTA

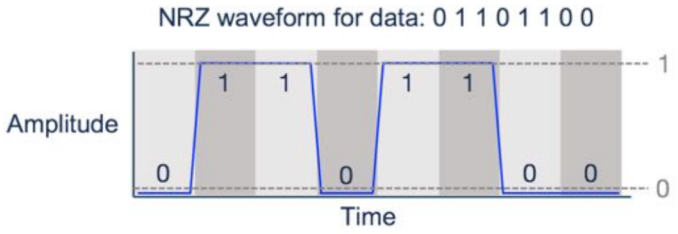
IEEE 802.3bs (400G Ethernet) Timeline

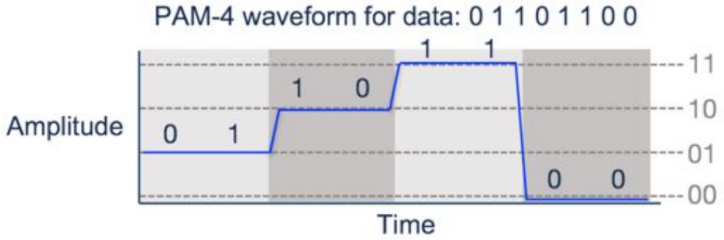


ARISTA

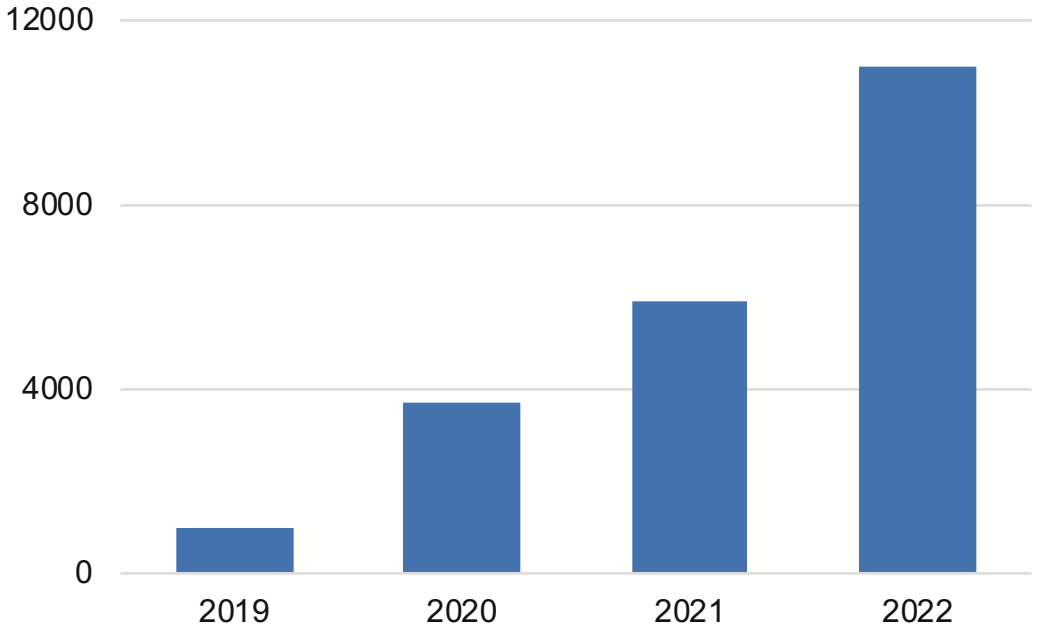
400G Overview

- 400G standard defined by1EEE 802.3bs
- Ratified December 2017
- Defines various physical layer specifications, for 200G and 400G.
- The standard increases the channel capacity to 50G using PAM-4 modulation.
- PAM-4 utilises 2 bits per symbol for double the data rate.
- An increase from 25G NRZ with 100G





Expected 400G Ethernet Ramp



400G switch port can support 1x400G, 2x200G or 4x100G port modes

Source: Dell'Oro Market Research, Ethernet Switch Update, July 2018

Switch Silicon Speed transition

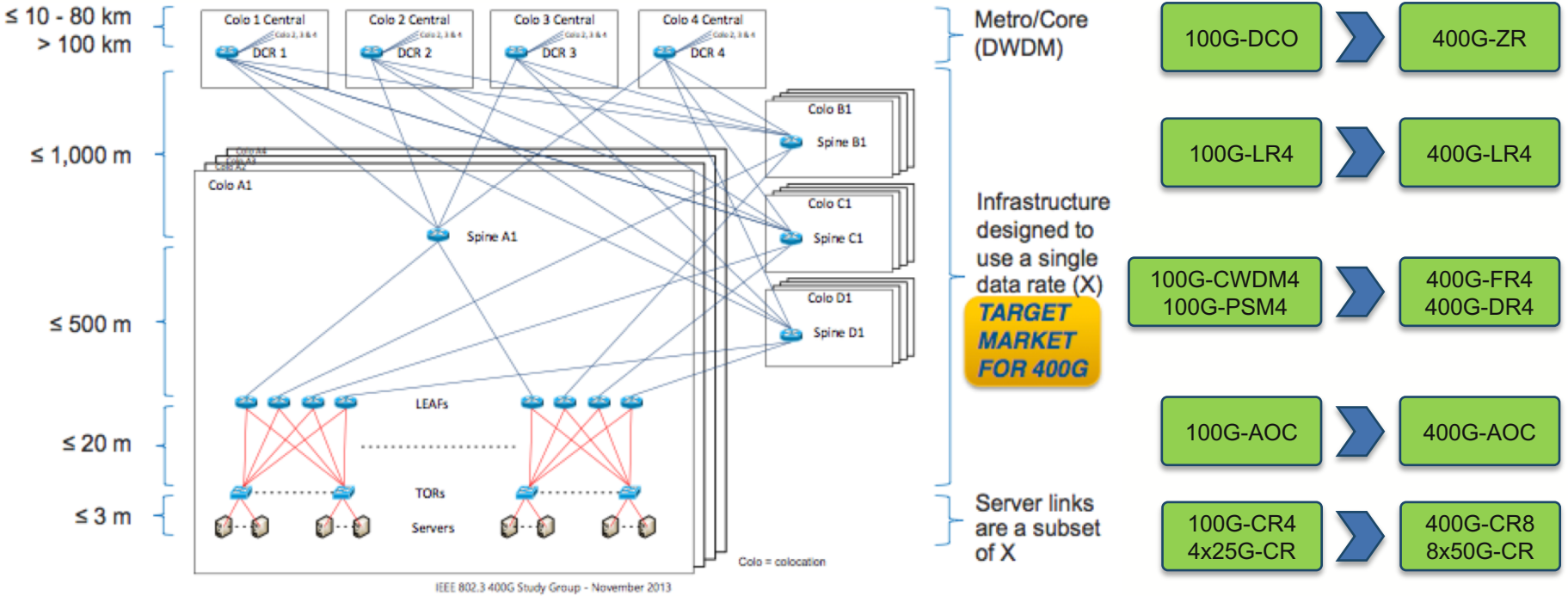
Lane Speed	10Gbps	25Gbps	50Gbps	100Gbps
1X	10G	25G	50G	100G
2X	-	50G	100G	200G
4X	40G	100G	200G	400G
8X	-	-	400G	800G
Availability and Ramp	2011/2012	2015/2016	2018/2019	2020/2021

Server Interface

Leaf-Spine Interface

ARISTA

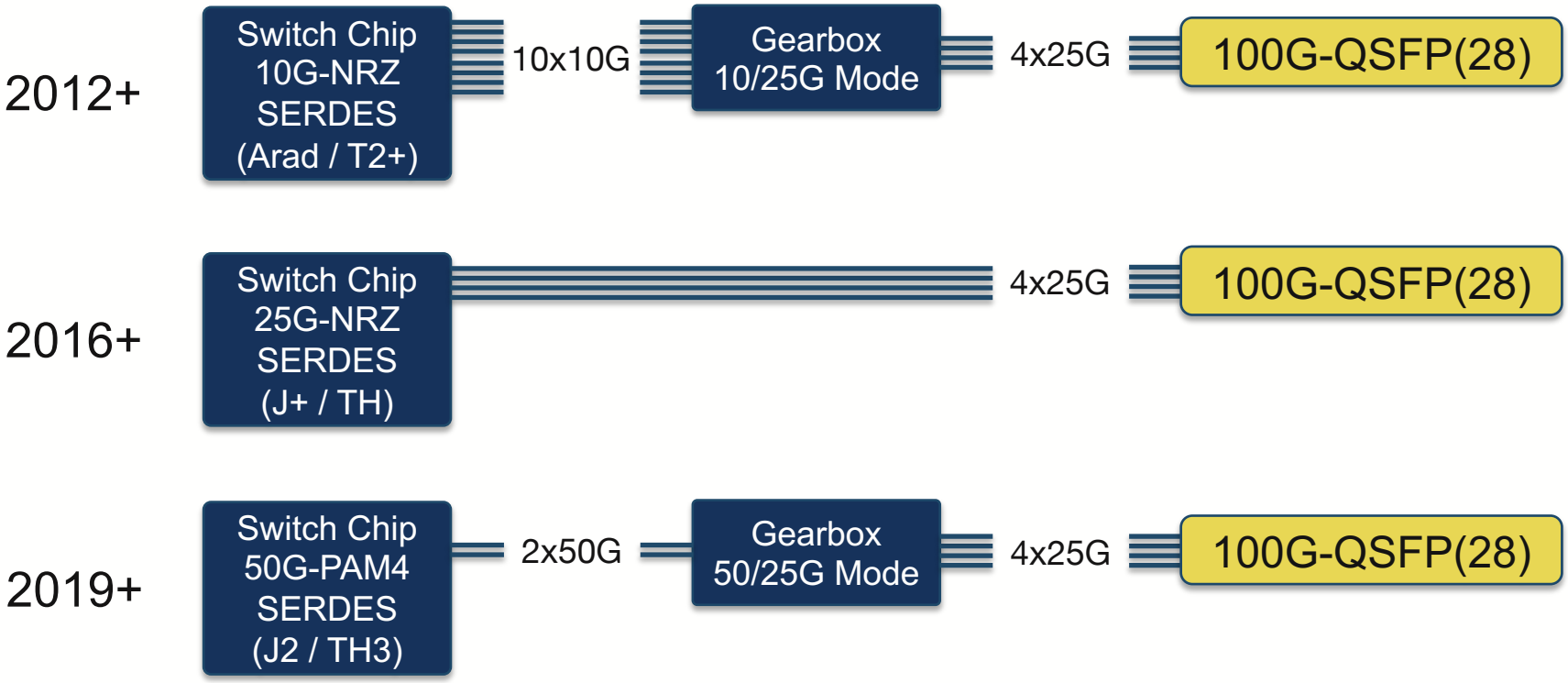
Transition of Cloud Networks from 100G to 400G



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

ARISTA

Adapting SERDES Speeds to 100G Interfaces



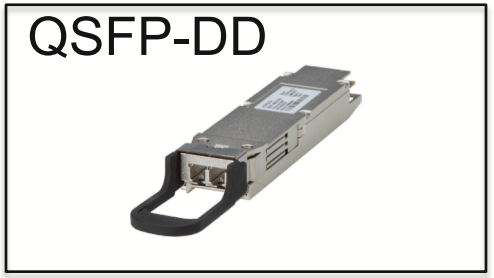
ARISTA

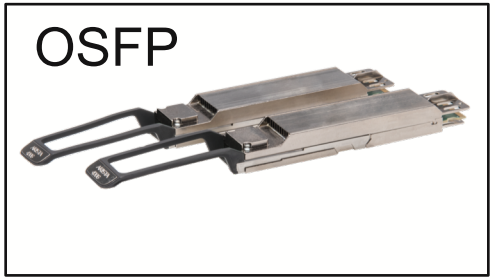
OSFP and QSFP-DD 400G Optics

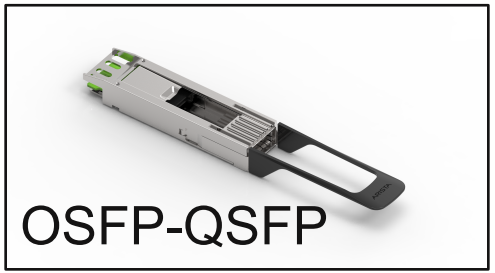
Arista will have OSFP and QSFP-DD products

- Two connector standards: OSFP and QSFP-DD
- We prefer OSFP as it is technically superior
  - Higher power budget, easier to cool
  - More choices earlier in OSFP (not dependent on 7nm gearboxes)
  - More options for high-power optics (like ZR 120km 400g)
  - Supports 100G electrical which is the ***most*** cost-effective
- QSFP-DD is backwards-compatible with QSFP-100

- Solution: OSFP-to-QSFP Adapter for 100G compatibility
  - Inserts into an OSFP slot
  - Lets you deploy a 400G switch and run it at 100G!
  - Mechanical adaptor - purely passive



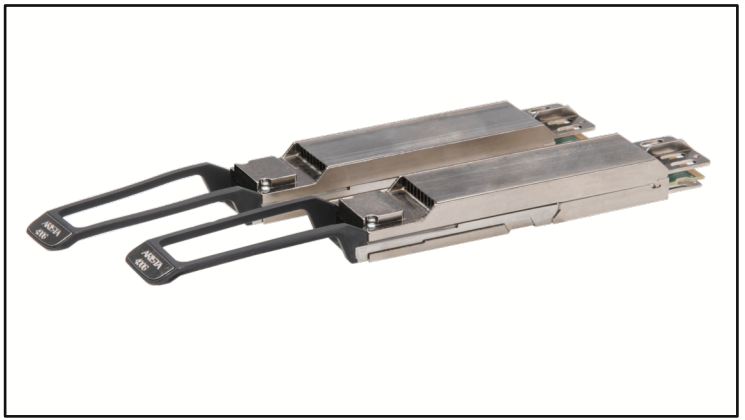


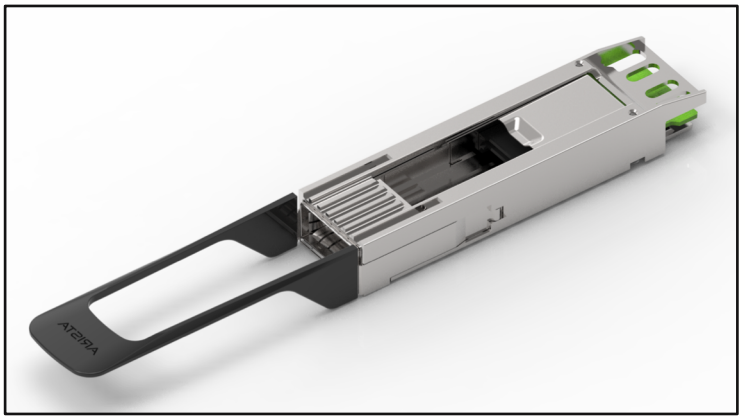


ARISTA

The OSFP (Octal Small Form Factor Pluggable)

- Eight Lanes at 56 or 112Gbps
  - Supports 400G and 800G (2x400G)
- High Port Density – 36 per 1U
  - 28.8Tbps with Nx100G Serdes
- High Thermal Capacity
  - Demonstrated 18W power envelope
- Supports full range of Optics
  - Data Center to Metro Reach
- Roadmap to 800G (2x400G)
  - Required by 2020
- Backward Compatible with QSFP
  - Simple OSFP-QSFP adaptor required





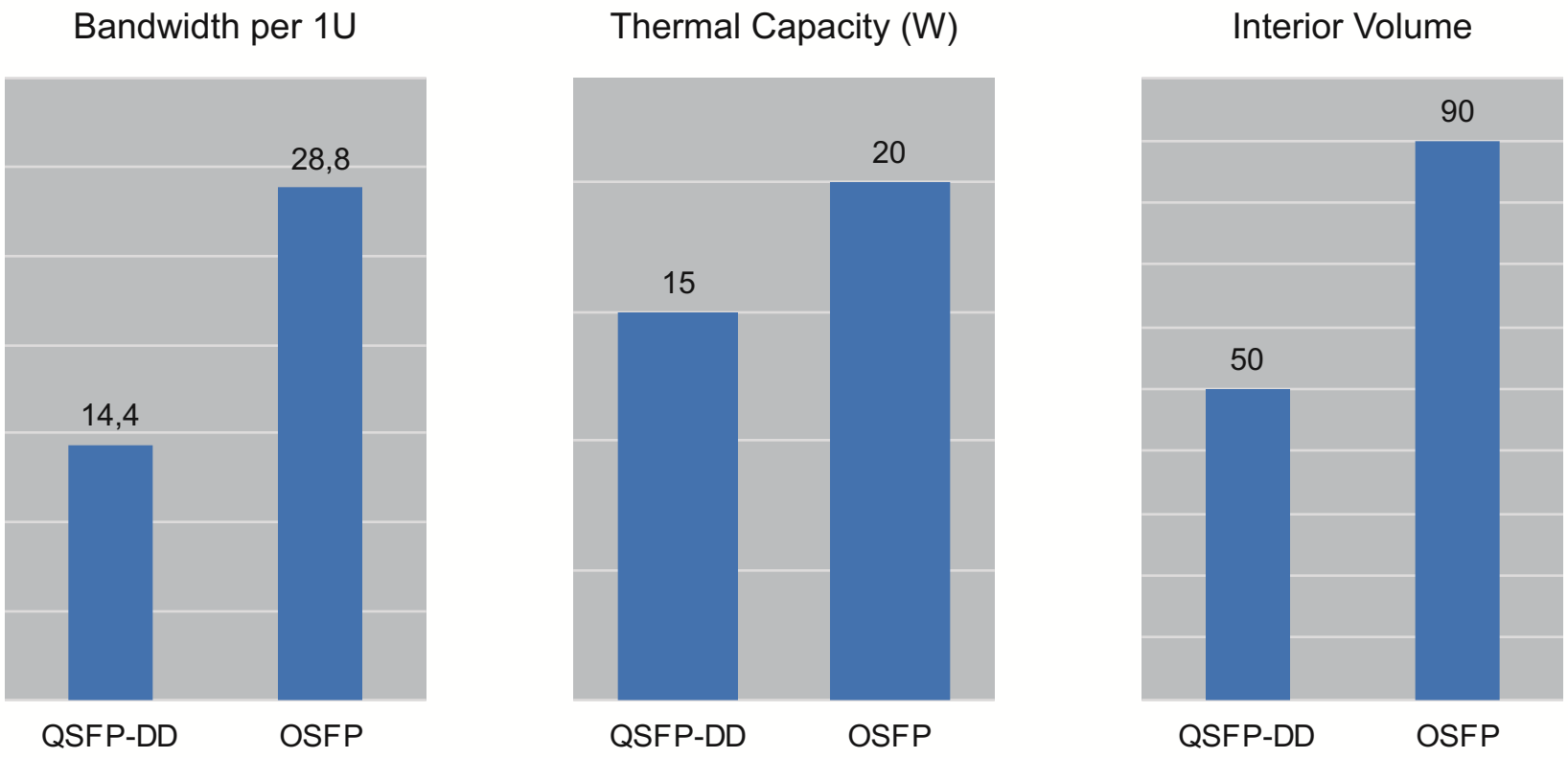
ARISTA

400G OSFP Optical PMDs

- Supports all 400G use cases up to Metro Reach and Beyond
  - No single 400G optics technology addresses all market requirements
- OSFP Supports Nx100G Dual 400G and 800G Optics
  - Electrical and thermal performance supports eight lanes of 100G
- Over time, most 400G ports will use 100G electrical lanes
  - 100G lane switch silicon will ramp starting in 2021

OSFP: Single form factor for data center to metro reaches

OSFP Compared to QSFP-DD – Technically Superior



ARISTA

Pluggable Form Factors Comparison

	OSFP	QSFP-DD
36 ports per 1RU	Yes	Yes
20W Thermal Capacity for 400G-ZR+ & 800 G	Yes	No
Forward compatible with 800G systems	Yes	No
Backwards compatible with QSFP28	Yes, with adapter	Yes
Max Copper DAC length	3m	2.5m

400G Optics/Cables Portfolio

Copper Cables

- Up to 5 meters length
- OSFP-OSFP & OSFP-4x100G QSFP breakout options
- Allows for short reach connection

OSFP Form Factor

Active Optical Cables

- Up to 30 meters length
- Low cost Intra-rack connectivity
- Ease of fiber management

Plug and Play

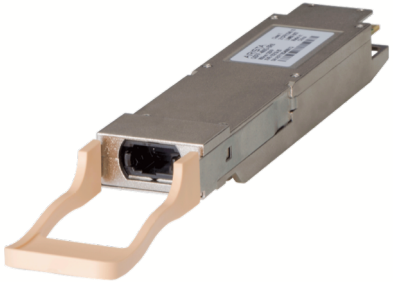
Optical transceivers

- 400G-SR8: up to 70m MMF
- 400G-DR4: up to 500m SMF
- 400G-LR8/FR4: 2km-10km SMF
- Compliant to industry standards

Standards Compliant

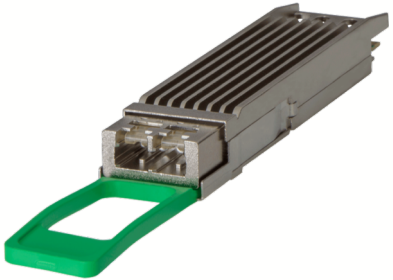
Broad Range of 400G OSFP and QSFP-DD Optics

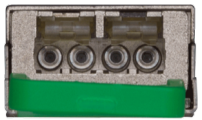
- 400G-SR8: up to 70m MMF
- Compliant to industry standards
- MTP-16 Connector



- 400G-DR4: up to 500m SMF
- Compliant to industry standards
- MTP12 Connector

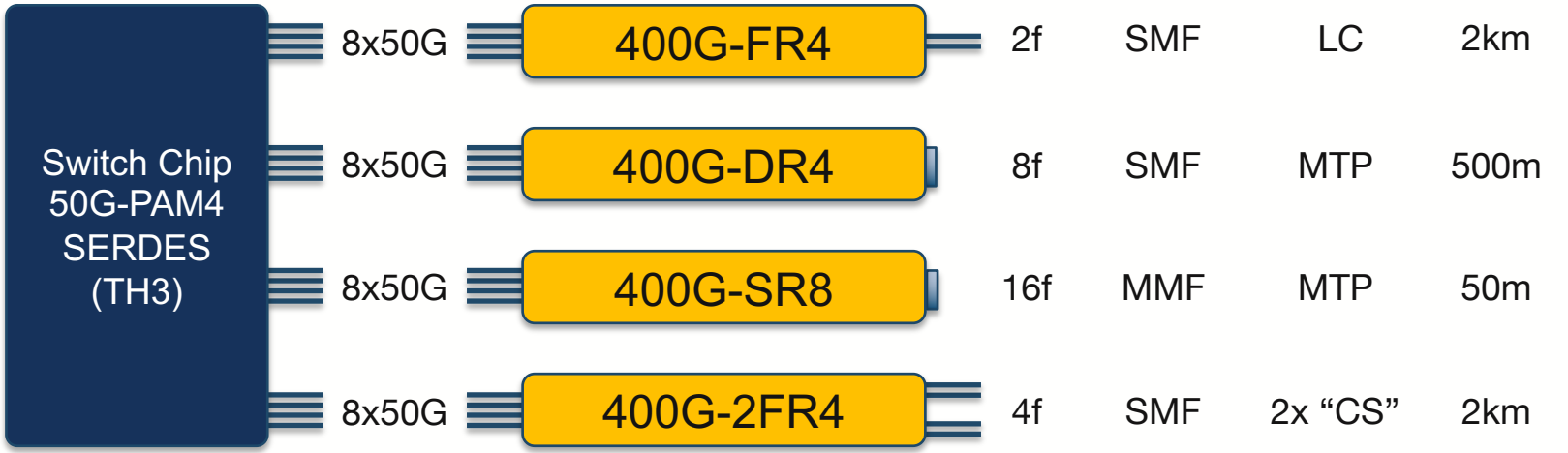
- 400G-2FR4: 2km-10km SMF
- Compliant to industry standards
- Dual SC (CS) Connector





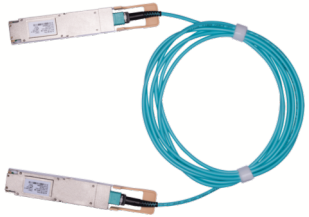
Standards Compliant → Plug and Play → OSFP Form Factor

400G OSFP and Optics – Q1 2019

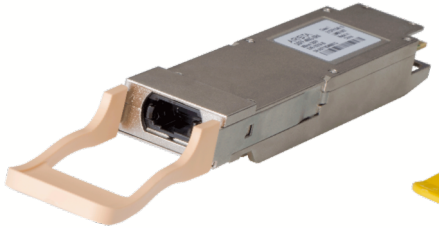


ARISTA

400G Optics breakout to 100G

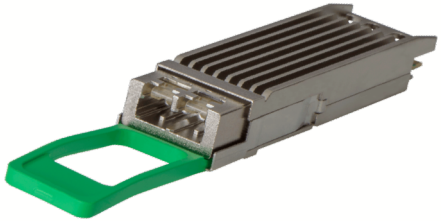


400G-AOC



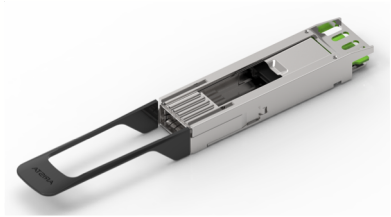
400G-SR8

400G-DR4

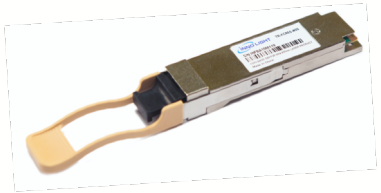


400G-FR4
400G-2FR4

OSFP → QSFP
Adapter



100G-SR4

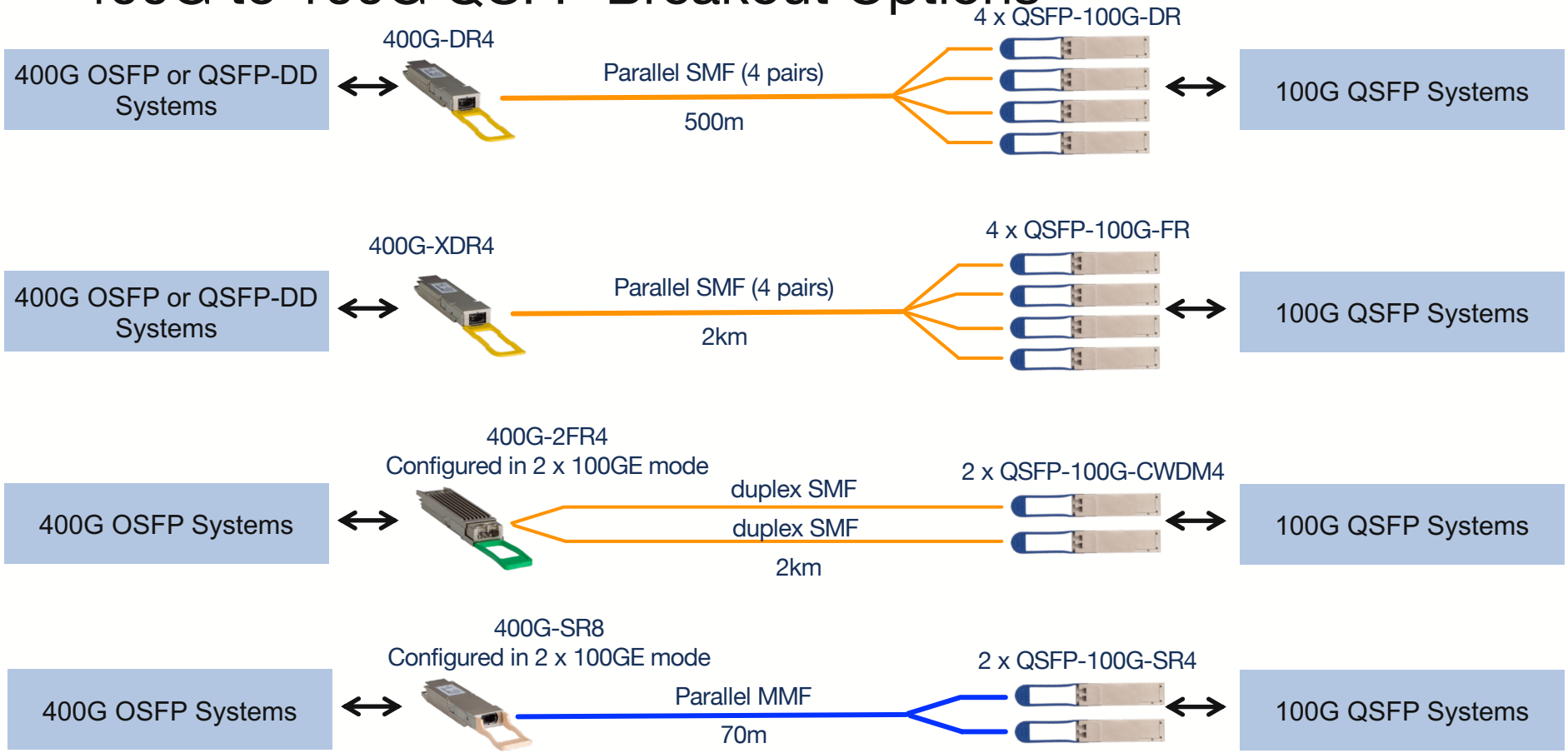


100G-DR1

100G-CWDM4

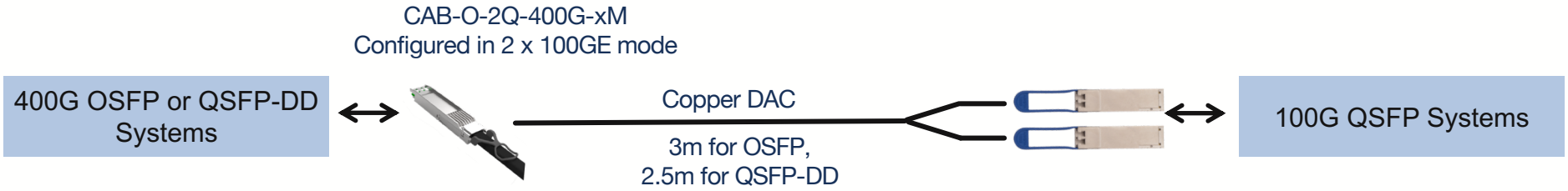
ARISTA

400G to 100G QSFP Breakout Options



ARISTA

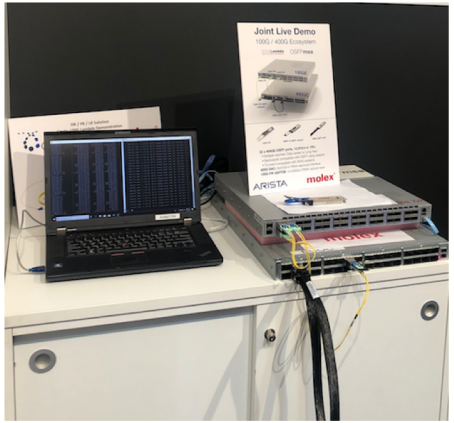
400G to 100G QSFP Breakout Options



- Additional 400G Optics are expected to enable breakout to 100G in 2H ’19

Arista 400G Demos at ECOC Sep 2018

- Demonstrated all flavors of 400G OSFP with live traffic



Arista 400G Demos at ECOC Sep 2018

Ethernet Alliance

Arista 400G	
400GBASE-CR8	OSFP
400GBASE-CR8	OSFP
400G AOC	OSFP
400G 2xFR4	OSFP
400GBASE-SR8	OSFP
400GBASE-CR8	OSFP
400G AOC	OSFP
400G 2xFR4	OSFP
400GBASE-SR8	OSFP
400GBASE-CR8	OSFP
400GBASE-CR8	OSFP
400GBASE-FR4	OSFP
400GBASE-DR4	OSFP

To Ixia

To Spirent

Breakout to 100GE switch

100G Lambda MSA

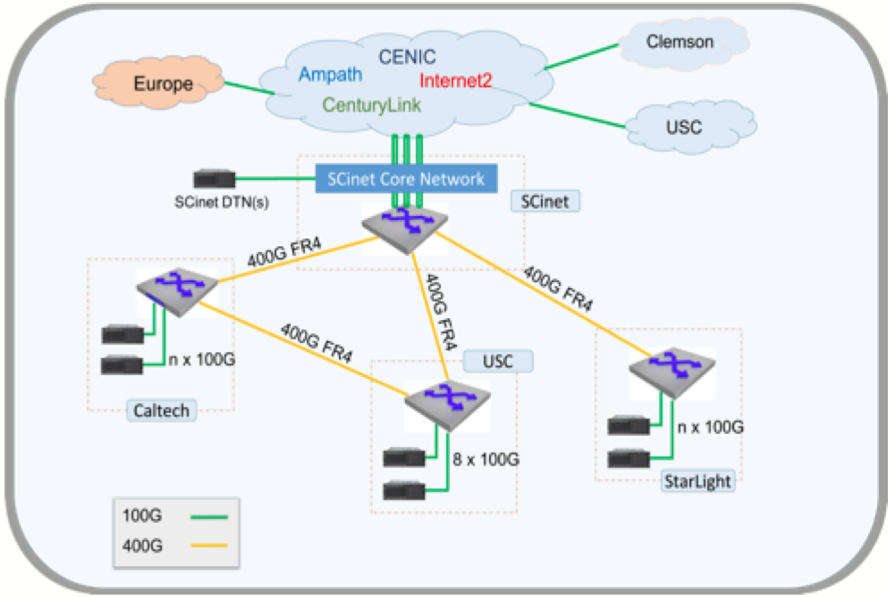
Why upgrade to 400G?

- Lower price performance per bandwidth
- Increased power efficiency per bandwidth
  - ~25% power efficiency per 100G
- Better port density
  - Up to 128x100G with TH3 (7006X4) in 1RU.
  - More interface options will be made available (mix of OSFP and QSFP-DD)
- Simplify the Network
  - Utilise existing optics and cabling you have invested in
  - Upgrade optics as required (400G, 100G breakout)
  - Long-haul

OK for optics
But what about switches?

TH3 SC18 **Nov 2018** Live Demo

- 7060X4-32 ("Blackhawk") live demo at Super Computing (SC18)
- Two 7060X4-32 (one in USC and another in CalTech booths) connecting to SciNet NOC
- SciNet NOC connects to external internet links to a University in Chile
- Showcasing 400G-400G, 100G-100G and 400G-100G breakouts with live traffic
- Wide range of 400G OSFP/DD optics as well as OSFP-QSFP28 adapter



ARISTA

Range of Systems for 100G/400G Scale-out Applications

7368X4

High Network Radix Modular System

- Choice of port module configurations
- Improved power efficiency per bandwidth
- Upgradeable to next generation
- 128x 100G QSFP or 32x 400G in 4RU

7060DX4-32

7060PX4-32

Efficient for density, performance and power

- Flexible 400G options with OSFP or QSFP-DD
- Ease migration to 400G with 128 x 100G mode
- Low Power, Latency and Scalable performance

Consistent Architecture with choices of industry standard interfaces

ARISTA

Broadcom Tomahawk 3 – 12.8T

3.2T
32 x 100G
128 x 25G Serdes

6.4T
64 x 100G
256 x 25G Serdes

12.8T
32 x 400G
256 x 50G Serdes

Key Drivers for Higher Performance

- Machine Learning Clusters
- NVMe over Fabrics
- Next Generation DC Pod Architectures

One Year step to 12.8Tbps

- 40% reduced Power / Port
- High Performance Packet Processor and Buffer Architecture
- Robust 50G PAM4 Serdes
- Ultra efficient design in 16nm

ARISTA

Arista 7060X4-32 Series 400G

400G High Performance Fixed System

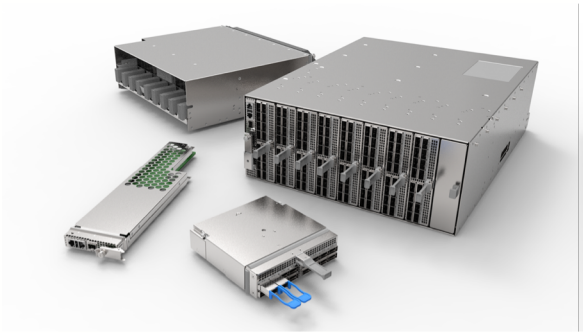
- High Performance 400G system with hyperscale features
  - High Performance with 12.8Tbps and 8Bpps
  - Latency - 800ns port to port with cut-through mode
  - Shared 64MB Smart-buffer and monitoring with LANZ
- Datacenter Optimized
  - Datacenter Spine and next gen Leaf
  - Under 17W per 400G port typical to lower TCO
  - Increased routing scale and robustness
  - Elephant Flow Detector to automatically manage large flows
- Hyperscale Cloud Networks Scalability
  - OSPF, BGP, Multicast & MLAG - 400K routes, 128-way ECMP
  - Dynamic Load Balancing & Dynamic Group Multipath
  - Optimized hashing and ALPM for large scale IPv4 and IPv6

32 Ports 400G – 12.8Tbps

Tomahawk3

7368X – Architected for Cloud Operations

- Switch Card – removes from rear without cable changes
- Management Module – removes from front
- Power Supplies – rear accessible and hot swap
- Fan Modules – individually removable and hot swap
- Choice of 100G and 400G Modules – mix and match



Arista 7368X4 Series 100G/400G

100/400G High Performance Semi-Fixed System

- High Performance 100G/400G system with hyperscale features
  - High Performance with 12.8Tbps and 8Bpps
  - Latency - 700ns port to port with cut-through mode
  - Shared 64MB Smart-buffer and monitoring with LANZ
- Datacenter Optimized
  - Datacenter Spine and next gen Leaf
  - Under 10W per 100G port typical to lower TCO
  - Increased routing scale and robustness
  - Elephant Flow Detector to automatically manage large flows
- Hyperscale Cloud Networks Scalability
  - OSPF, BGP, Multicast & MLAG - 400K routes, 128-way ECMP
  - Dynamic Load Balancing & Dynamic Group Multipath
  - Optimized hashing and ALPM for large scale IPv4 and IPv6

128 Ports 100G – 12.8Tbps

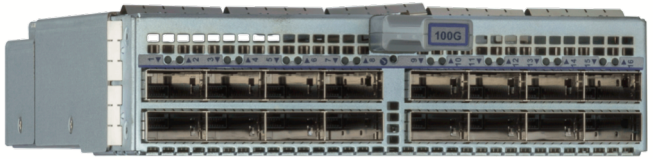
32 Ports 400G – 12.8Tbps

Tomahawk3

ARISTA

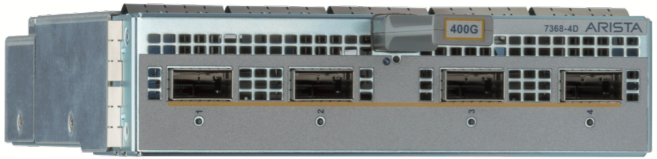
7368X – 100G and 400G Modules

- 16 x 100G Module
  - QSFP100 ports for range of cables and optics
  - Optional 200G Mode with Alternate ports
  - Hotswap with no power off
  - Integrated ejector and handle

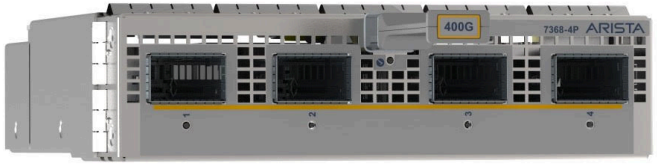


QSFP – 100G

- 4 x 400G Modules
  - QSFP-DD or OSFP
  - Widest range of cables and optics
  - Flexible 400G or 4x 100G (and 2x 200G) modes
  - Hotswap with no power off
  - Integrated ejector and handle

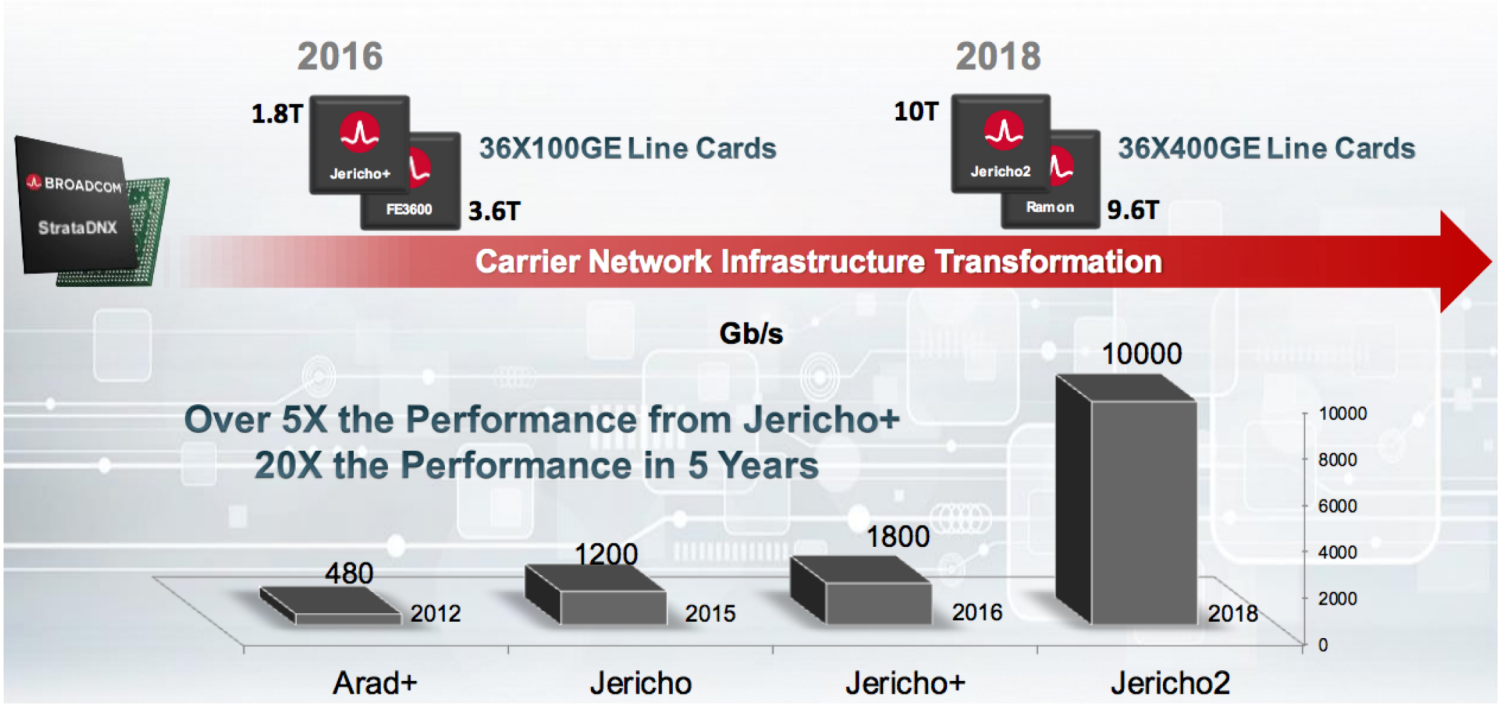


QSFP-DD – 400G



OSFP – 400G

Broadcom Jericho2 – 10Tbps

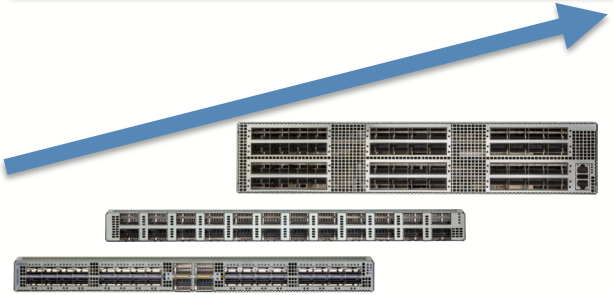


ARISTA

Next Generation R-Series Portfolio

7280R3 → 7500R3 → 7800R3

4Tbps – 38Tbps



76Tbps – 230Tbps

115Tbps – 460Tbps

7280R3 Series Fixed 100G/400G Switches

Wire Speed 100/400G with Deep Buffers

High Performance:

- Up to 48 x 400G wire speed ports in 2RU
- Non-blocking up to 19.2 Tbps and 8Bpps
- FlexRoute™ - 1.3M / 2.5 Million+ IPv4 & IPv6 Routes

R-Series Architecture:

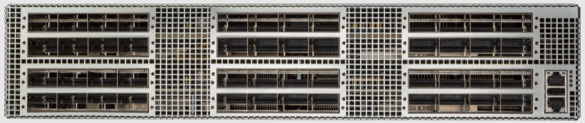
- VOQ architecture and deep buffers for lossless forwarding
- EOS for convergence and scale

Advanced Features:

- VXLAN Routing, Advanced Load Balancing
- Algorithmic ACLs, INT and Accelerated sFlow
- EVPN, MPLS, Segment Routing

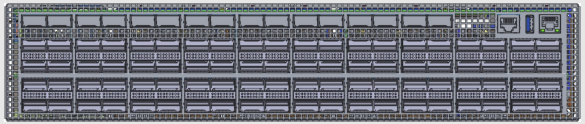
Cloud and Carrier Grade Networking:

- Dense 100G and 400G for SP, Cloud, Internet, HPC & CDN
- DC Optimized airflow and AC / DC power



7280PR3-48 - 48 x 400G

7280PR3-24 - 24 x 400G



7280CR3-96 - 96 x 100G

7280CR3-32P4 - 32 x 100G / 4 x 400G

ARISTA

7500R3 High Density 400G and 100G Spine Systems

High Performance 100G / 400G Spine:

- 230Tbps of throughput with choice of Chassis
- Consistent VOQ / Deep Buffers
- Backward compatible with 7500R and 7500R2

Chassis	400G OSFP	4 x 100G	100G QSFP
DCS-7512	288	1152	432
DCS-7508	192	768	288
DCS-7504	96	384	144

400G Spine:

- 24 x 400G OSFP linecards
- Supports range of optics and cables to ZR and ZR+
- Breakout to 4x100G and 2x200G

100G Spine:

- 36 ports of 100G with QSFP
- Supports copper cables, AOC, data center to DWDM optics
- 2.5M Route Scale Option

7800R3 Series Next Generation 100G/400G

Cloud and Carrier Grade Networking

High Performance for next 10 years:

- Up to 576 x 400G wire speed ports
- Non-blocking up to 460 Tbps and 96Bpps
- 14.4 Tbps / slot with 36 x 400G linecards
- Upgradable to 800G (28.8Tbps /slot) for higher density

R-Series Architecture:

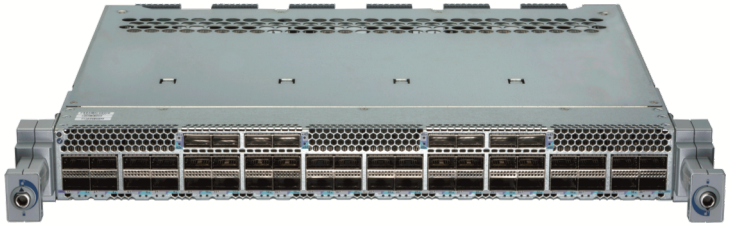
- VOQ architecture and deep buffers for lossless forwarding
- FlexRoute™ - 1.3M / 2.5 Million+ IPv4 & IPv6 Routes
- EOS for convergence and scale

Advanced Features:

- VXLAN Routing, Advanced Load Balancing
- Algorithmic ACLs, INT and Accelerated sFlow
- EVPN, MPLS, Segment Routing
- Dense 100G and 400G for SP, Cloud, Internet, HPC & CDN

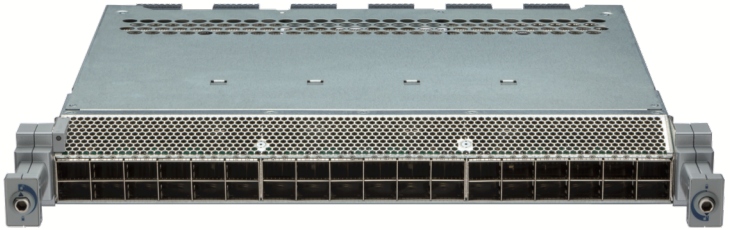
ARISTA

Highest Capacity 400G and 100G Spine System



100G Spine:

- 48 ports of 100G with QSFP
- Supports copper cables, data center to DWDM optics
- 2.5 Million Routes with features



400G Spine:

- 36 ports of 400G OSFP – 14.4Tbps
- 6 Billion Packets per second of L2 and L3
- Range of optics and cables - ZR and ZR+
- Flexible 4x100G and 2 x 200G Modes

High Performance Spine:

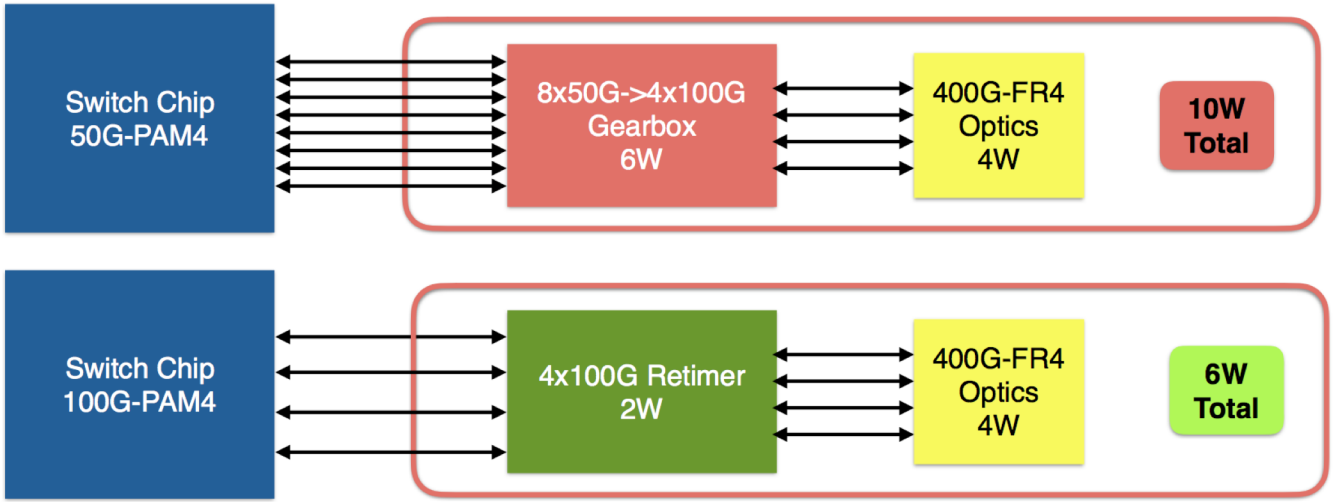
- Choice of Chassis (4/8/16 slot)
- Future higher density and 800G

Chassis	Throughput	400G OSFP	4 x 100G	100G QSFP
DCS-7816	460Tbps	576	2304	768
DCS-7808	230Tbps	288	1152	384
DCS-7804	115Tbps	144	576	192

What’s next?

What's Next?

- Next generation of ASICs will support 100G SERDES
- This will enable further reductions in optic costs and power.



Moving toward 800G Ethernet technology

Type	Distance	Cable	Lanes	Power
400G-DR4	500m	8 SMF	4	10W
400G-FR4	2km	2 SMF	4	10W
Dual 400G-DR4	2km	16 SMF	8	12W
Dual 400G-FR4	500m	4 SMF	8	12W
800G-FR8	2km	2 SMF	8	12W
800G-LR8	10km	2 SMF	8	12W

QSFP-100G-SR4 is 3.5W – 400G: >28% reduced power, 800G: >57% reduced power

Looking forward…

- 112G Electrical Performance needed
- Thermal Performance should be kept an eye on
  - Required for 800G and Dual 400G Optics (20W)
- Optical Interoperability with 100G Lambda Optics
  - Interoperability for 400G-DR4/FR4/LR4/etc

Thank You

www.arista.com