# 400G and beyond

NLNOG, Sep 6, 2019 Ralf Korschner<ralf@arista.com>

# Drivers for 400G

## 1. Bandwidth demand of hyperscale cloud networks





Source: Urs Hoelzle, Google, OFC 2017

# 2. Lowest cost / bit

A customer perspective:

"It is all about cost, in particular, \$ per Gbps, there is no other religion"

# 3. Lowest power / bit

# **Datacenter Ethernet Transition Expectations**



Source: Dell'Oro Group Jan 2018 Ethernet Switching Forecast

# **Merchant Silicon**

#### Tomahawk 3: By the numbers

•12.8 Tb/s multilayer L3 Ethernet switching

•Configurable as 32 ports 400GbE, 64 ports 200GbE or 128 ports 100GbE

•256 dual-mode, high-reach SerDes, each of which supports 56G-PAM4 and 28G-NRZ over long-reach optics, Direct Attach Copper (DAC) and backplanes

•Low-latency StrataXGS pipeline architecture

•Delivers 40% lower power consumption per 100GbE switch port and up to 75% lower cost per 100GbE switch port

•Integrated 12.8Tb/s shared-buffer architecture offers 4X higher burst absorption and provides the highest performance and lowest end-to-end query completion times (QCT) for RDMA over Converged Ethernet (RoCEv2) based workloads

•Broadview<sup>™</sup> Gen 3 network instrumentation feature set and software suite, providing comprehensive visibility to network operators into packet flow behavior, traffic management state, and switch internal performance

•Comprehensively and configurably supports all packet processing and traffic management requirements for next-gen hyperscale network use cases: >2X IP route forwarding scale, 2X ECMP scale, Dynamic Load Balancing and Group Multipathing, In-Band Network Telemetry, Elephant Flow detection and re-prioritization



# **Merchant Silicon**



Carrier networks require a high level of flexibility to address emerging use cases, standards, and protocols. Jericho2 introduces the innovative Elastic Pipe<sup>™</sup> technology, an evolution of the StrataDNX field-proven programmable pipe architecture. The Elastic Pipe future-proofs Jericho2 Switch-Routers by enabling new services while avoiding costly hardware replacements. In addition, the innovative Modular Centralized Database capability of Jericho2's Elastic Pipe enables flexible deployments across different networks and in multiple locations within the same network.

#### **StrataDNX Family Key Features**

- <u>Jericho2 (BCM88690)</u>
  - Future-proofing via Elastic Pipe ™
    - >> Flexible binding of a centralized database to any stage of the pipe
    - >> Extending Jericho2 pipe via reserved general purpose stages
  - High bandwidth, low power, and in-package HBM packet memory offering up to 160X more traffic buffering compared with on-chip memory, eliminating packet drops
  - Up to 10Tb/s switching capacity per device
  - High-speed, high-density port interfaces up to 400GE leveraging best-in-class 50Gb/s
     PAM-4 SerDes
  - Carrier grade Hierarchical Traffic Manager
  - Large scale, hardware-accelerated instrumentation and telemetry

# **Merchant Silicon**



•Key attributes and benefits of the new StrataXGS Trident 4 series include: Enables the next major advancement in enterprise data center and campus networks, supporting up to 128x100GbE or 32x400GbE switching and routing with a compiler-programmable architecture

Provides the flexibility to implement both standard Ethernet switch/routing and advanced network functionality such as DDoS protection, application load balancing, and large-scale NAT
An extremely rich, enterprise-grade feature set with a high degree of feature concurrency

•Industry-leading packet buffer and database sizes

•Extensive programmable in-band telemetry including support for IFA 2.0 (In-band Flow Analyzer version 2)

•Four GHz-class processors on chip enabling powerful out-of-band (streaming) telemetry and a variety of Broadcom-provided embedded applications

•

# 400G Platforms Offer Cost Effective and High Density 100G

Upgrade Spine to 400G Systems



# Evolving Network Bandwidth from 100G to 400G



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

Increase Network Bandwidth 4x with no change in fiber plant

# NRZ and PAM4 Modulation

- PAM-4 and NRZ are different ways to modulate signals to transmit data.
- NRZ is dead simple. PAM-4 is a bit less simple.



# Silicon Serdes Speeds and Optical Transceivers



# 400G Optical Transceiver & Cable Types

Application	Reach	Media Type	Solution
TOR to Server	3m	Copper	400G-CR8 (DACs)
TOR to Leaf	30m	AOC	400G-AOC
Switch to switch inside the datacenter	100m	Parallel MMF	400G-SR8
	100m	MMF	400G-BIDI
	500m / 2km	Parallel SMF	400G-DR4 / XDR4
	2km	SMF	400G-FR4
	2km	Dual SMF	400G-2FR4
DCI, Metro & Long-Haul	120km / 1000km+	SMF	400G-ZR / ZR+

### Arista will support all available types

# Transceiver form factors

400G Form Factors

# CFP8



# CFP8 400G Optics

#### • CFP8

- Higher power budget, easier to cool
- 28G SERDES (16x28G) on the electrical side
- More options for high-power optics in the future
- Optical facing LR8 (8x56G) available as of today
- Outlook to the future
  - As AUI is 28G SERDES, optics might need to be replaced to support 800G
  - Only optics released are 8 lane on the optical side so far (no LR4/DR4+)
  - Low port density due to size of optic



**CFP8 Form Factor** 



https://multilaneinc.com/s-parameter-data-cfp8/

# **QSFP-DD 400G Optics**

#### QSFP-DD

- Reduced power budget as harder to cool
  - >> Options with 'extended optics' has been presented, but that means optic will stick further out of the port
- Electrical AUI can be either 8x56G SERDES or 4x112G SERDES
- QSFP-DD is backwards-compatible with QSFP28
- High Port Density 24 to 36x400G per 1U (depending on the ASIC)
  - $\gg$  Up to 28.8Tbps with Nx112G Serdes
- Outlook to the future
  - Challenges for 800G on QSFP-DD
    - >> Some constraints due to power/cooling
  - 76 pin AUI connector faces challenges around signal synchronicity at 112G-PAM4
  - Availability of 8 Lane and 4 Lane optical side (example LR4/LR8)

# QSFP-DD





https://multilaneinc.com/s-parameter-data-qsfp-dd/

# **OSFP 400G Optics**

#### OSFP

- Higher power budget, easier to cool
  - >> Demonstrated 20W power envelope
- More options for high-power optics (like ZR 120km 400g)
- Eight Lanes at 56 or 112Gbps
  - $\gg$  Supports 400G and 800G (2x400G)
  - >> 60 pin AUI connector optimized for high bandwidth SERDES
- High Port Density 24 to 36x400G per 1U (depending on the ASIC)
  - $\gg$  Up to 28.8Tbps with Nx112G Serdes
- Outlook to the future
  - Roadmap to 800G (2x400G)
    - >> Outlook to 2020/2021
  - Availability of 8 Lane and 4 Lane optical side (example LR4/LR8)



**OSFP Form Factor** 



https://multilaneinc.com/s-parameter-data-osfp/

# **OSFP 400G Optics**

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



# 400G Optical Modules for Multi-Mode Fiber



400G-SR8, 8 fiber pairs, 100m reach Breakout to 2 x 200G-SR4 or 2 x 100G-SR4 QSFPs

MPO-16 APC (Angled MMF) Requires new MPO-16 fiber connectivity



400G-BIDI, 4 fiber pairs, 100m reach Breakout to 4 x 100G-BIDI QSFP

# Optical Breakout of 400G-DR4 to 4 x 100G-DR



# Thank You For Your Attention