Open19 Based Next Gen Distributed Mega-Switch

Rochan Sankar - Broadcom
Liz Hardin – Molex
Rob Stone – Broadcom/Cobo
Yuval Bachar - LinkedIn
Tomahawk3 Switching Engine
Role of the Network Switching Element in Scale-Out Cloud

- Fully connect N x M compute and storage endpoints at massive scale
- Provide high-bandwidth, highly available, multi-tiered interconnect fabric
- Rapidly scale over generations on performance, cost, power efficiency
- Get out of the way
Introducing StrataXGS® Tomahawk 3

Monster Bandwidth
Record 12.8Tbps Line Rate Switching in a Single Chip

Bleeding Edge Speeds & Feeds
First to 32x400GbE / 64x200GbE / 128x100GbE Ports

Architected for Hyperscale
Serves Exploding Next-Gen AI & Storage Workloads

Step Function in Opex/Capex Savings
40% Lower Power and 75% Lower Cost per 100Gbps

High Velocity to Deployment
Leads the 400GbE Build-out, 14 months post 6.4T

32 ports 400GbE
64 ports 200GbE
128 ports 100GbE
Critical Enablers for Fast Scaling to 12.8Tbps

- 40% Reduced Power/Port
- Scalable, High-Performance Packet Processing & Buffer Architecture
- Robust 50G PAM-4 Serdes IP
- Ultra-efficient design in 16nm

Next Performance Node Drivers
- Deep Learning Clusters
- NVMe over Fabrics
- Evolved DC Pod Architecture

Hyperscale Network Evolution
Tomahawk Accelerates 200/400GbE Deployment Cycle

20X Bandwidth Increase over 7 Years
Outpacing Moore’s Law

12.8T 16nm
32 port 400 GE

6.4T 16nm
64 port 100 GE

3.2T 28nm
32 port 100 GE

1.28T 40nm
32 port 40 GE

640G 40nm
64 port 10 GE

2010 2012 2014 2016 2017
Scaling Up Performance with Chip Architecture (Beat Moore’s Law)

Scalable MMU
- Scalable internal wiring
- Enqueue / Dequeue scales to higher port speeds
- Buffer Sharing
- Burst Absorption
- Congestion Control

Scalable Pipeline
- Lightweight for scaling-up pipeline count w/ limited freq
- Ruthlessly efficient memory & flop count
- Decoupled datapath
- Table resource sharing

I/O Integration
- Design control of Serdes IP for channel performance
- Advance implementation on new process nodes
- 25G NRZ → 50G PAM-4 → 100G-PAM4 signaling
- Low Power PMA, PMD, PCS
Tomahawk 3 Collapses End-to-End Application Latency

Lower Switch Radix Creates Multiple Hops

3 Intra-Cluster Switch Hops ~ 1 μs
Using ‘lowest latency’ 64x100GbE

Tomahawk 3 Flattens Pods & Slashes Nominal Latency

Single Switch Hop ~ 400 ns
60% Reduction in intra-cluster network latency
Open19 Optical Data Cable
Box Overview

- Backshell and cable ‘internal’ portion updated with passive fiber solution
  - MTP fiber shown
- Compatible with existing panel opening and snap function
  - Facilitates simple upgrade of rack
Within box fiber management:

- Routed Ribbon FlexPlane
  - Height reduction improves fiber management
  - Routing allows in ribbon shuffling Routed Ribbons form connections
  - Pre-routing eliminates the need to add slack management areas on system

- Supports any MT type ferrule and connector, MPO/MTP, MXC, HBMT etc
- Panel mount at rear of panel will be compliant with Open19 panel
Cobo On board Optics
Specification Targets

• 12.8 Terabit 1 RU switch
• 400G+ optical modules
• Two module widths
  • 8 lane for 400G modules
  • 16 lane for 800G (e.g. dual 400G Ethernet) modules

Thermal environment
• Replaceable, interchangeable, interoperable

Optical connectivity only

Specification Release 1.0
Published
8 & 16 Lane On-Board Optical Modules

Key features:
✓ Single footprint
✓ Three power classes
✓ Module width:
  20 mm (8 – lane)
  36 mm (16 lane)
✓ Module lengths:
  ✓ Class A: 30 mm
  ✓ Class B: 40 mm
  ✓ Class C: 60 mm
Operational Benefits

- COBO based Switches arrive at data center ready to deploy
  - Factory built and tested
  - Faster time to go live
- Improved thermal and signal integrity performance
  - Reduced failure rate
  - Lower power consumption
- Decreased failures due to human interaction
- Support both front panel and blind mate backplane optical connectivity
Next Step - Enabling 1.6T Modules @ 100G / electrical lane

TE Connectivity COBO test board with Macom silicon at 100 Gbps

Note: limit lines are TE Connectivity hypothetical and do not reflect actual limits
# Application Space and Comparison

<table>
<thead>
<tr>
<th>Feature</th>
<th>QSFP</th>
<th>OSFP-DD</th>
<th>OSFP</th>
<th>COBO</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1 RU Front Panel I/O:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.2T</td>
<td>Yes (32 x 100)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6.4T</td>
<td>Yes (32 x 200)</td>
<td>Yes (32 x 200)</td>
<td>Yes (32 x 200)</td>
<td>-</td>
</tr>
<tr>
<td>12.8T</td>
<td>No</td>
<td>Yes (32 x 400)</td>
<td>Yes (32 x 400)</td>
<td>Yes (32 x 400)</td>
</tr>
<tr>
<td>25.6T</td>
<td>No</td>
<td>Yes (32 x 800)</td>
<td>Yes (32 x 800)</td>
<td>Yes (32 x 800 / 16 x 1600)</td>
</tr>
<tr>
<td>&gt; 25.6T</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes (32 x 1600)</td>
</tr>
<tr>
<td><strong>Field Serviceable</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td><strong>DAC Support</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td><strong>Optical Blind Mate Backplane</strong></td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Flexible Board Placement &amp; Heatsinking</strong></td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Integrated Mega-System
Platform Description

- None Blocking
- Front end and aggregating slices
- 1:2 ratio of aggregators vs. Front end
- Max Config 100T (1000x100G) – 12.8T chips sets
- All optical interconnect
- Dynamic or static optical shelf

- Max config half rack
- 16x2 @ 400G fibers per front slice
- 32x4 @400G fibers per aggregator slice
- 24x16x2 @ 400G at the optical shelf
Front Slice

12.8T Switching engine

16xQSFP-DD
64x100G

16xCOBO
16x400G

Open19 Optical

Open19 Power

Open19 Power
Aggregator Slice

12.8T Switching engine

32xCOBO
32x400G

Open19 Optical
Open19 Power
Open19 Optical
Open19 Power
## Configurations options

<table>
<thead>
<tr>
<th>Config</th>
<th>Front Slices</th>
<th>Agg slices</th>
<th>Supported ports 100G</th>
<th>Total RUs</th>
<th>Total system capacity</th>
<th>Optical connection per slice</th>
<th>Overall Optical cable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max</td>
<td>16</td>
<td>8</td>
<td>1000</td>
<td>26RU</td>
<td>100T</td>
<td>16@400G</td>
<td>192@400G</td>
</tr>
<tr>
<td>Mid level</td>
<td>12</td>
<td>6</td>
<td>768</td>
<td>22RU</td>
<td>76.8T</td>
<td>16@400G</td>
<td>192@400G</td>
</tr>
<tr>
<td>Entry level</td>
<td>8</td>
<td>4</td>
<td>512</td>
<td>14RU</td>
<td>51.2T</td>
<td>16@400G</td>
<td>192@400G</td>
</tr>
<tr>
<td>Legacy</td>
<td>6</td>
<td>3</td>
<td>384</td>
<td>12RU</td>
<td>38.4T</td>
<td>16@400G</td>
<td>192@400G</td>
</tr>
</tbody>
</table>
Thank You