Zaius & Barreleye G2
Accelerator Ecosystem around Google / Rackspace 48V OpenPOWER Platform

OpenCAPI / Nvlink / PCIe Gen4

Adi Gangidi
Sr. Systems Design Engineer
Rackspace
Agenda

- Introduction | Recap
- Development Update
- Technology Update
- Accelerators
  - PCIe Accelerators
  - NVLink Accelerators
  - OpenCAPI Accelerators
  - Enablement Frameworks
- Samples
Introduction
Zaius + Barreleye G2

• **2016**: Collaborative Effort Announced
  - OpenPOWER Foundation, Open Compute Project, OpenCAPI, IBM, Google & Rackspace

• **2017**: Donated EVT, DVT Specifications to OCP Server Workgroup
  - Open/Emerging Technologies

• **2018**: Development Update & IO/Accelerator Eco-system
  - Today’s Announcements

Zaius Block Diagram
2016 Announcement
ZAIUS Shelf & Sled
- Compact enclosure of Zaius MB
- 1.5OU height
- Shorter than HH cards
- Compatible with 48V open rack v2 with deployment shelf

ZAIUS Motherboard
- 2 x POWER9 LaGrange
- 48V input
- Front IO & service access
- 80 Lanes of PCIe Gen4
- 32 Lanes of OpenCAPI / NVLink 2.0
- Open Source BMC & Host Firmware

BARRELEYE G2 Chassis
- Full-depth 48V open rack v2
- 2 OU chassis supports FHFL cards
- Hot swap storage bay (24 Drives)
- Hot swap fans and VGA access
- Wattage plan to support accelerators via Power board
From Design... ...
to Test
To Test...

...through EVT & DVT
PVT Entry:
The Green Board
What’s New?

- 48V Motherboard
- Processor: Power9
- IO: PCIe Gen4 / OpenCAPI / NVLink 2.0

The Approach

- Solve for various workloads
- Provide future flexibility
- Decrease adoption time

### IO QUALIFICATION

<table>
<thead>
<tr>
<th>NETWORKING</th>
<th>RAID / HBA</th>
<th>GP GPU</th>
<th>OPENCAPI</th>
<th>STORAGE</th>
<th>MEMORY</th>
</tr>
</thead>
<tbody>
<tr>
<td>CX5 100GbE OCP</td>
<td>9460-16i Broadcom</td>
<td>PCIe Volta V100</td>
<td>ADM-9V3 Alpha Data</td>
<td>PM863A SATA</td>
<td>DDR4 2666 MHz</td>
</tr>
<tr>
<td>Mellanox</td>
<td>250S+ Nallatech</td>
<td>NVIDIA</td>
<td>Innova 2 Mellanox</td>
<td>Samsung</td>
<td>16GB 1R x4 Micron</td>
</tr>
<tr>
<td>CX4 25GbE OCP</td>
<td>PEX 8734 Switch Board</td>
<td>SXM2 Volta V100</td>
<td>Flash Storage Accelerator</td>
<td>Micron</td>
<td>DDR4 2666 MHz</td>
</tr>
<tr>
<td>Mellanox</td>
<td></td>
<td>NVIDIA</td>
<td>Molex</td>
<td>Samsung</td>
<td>8GB 1R x4 Micron</td>
</tr>
<tr>
<td>NetXtreme 25GbE</td>
<td></td>
<td></td>
<td></td>
<td>PM963 NVMe</td>
<td>DDR4 2400 MHz</td>
</tr>
<tr>
<td>Broadcom</td>
<td></td>
<td></td>
<td></td>
<td>Samsung</td>
<td>1R x4 Samsung</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
48V Rack
Power System

30 kW Power Shelf (Delta)

- 3 OU height, 4x 60A AC input
- 12x 3KW PSU Modules
- 6x ATS Modules (1 per 2 PSU)
- 10+2 Redundancy (If ATS Fails)
- Rack Management Controller
# 48V Open Rack V2

**Rack, Bus-bar & Clip**

<table>
<thead>
<tr>
<th><strong>OCP 48V Rack</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>+48VDC bus bar</td>
</tr>
<tr>
<td>(1 power zone)</td>
</tr>
<tr>
<td>30kW power shelf</td>
</tr>
<tr>
<td>(3 OU, 36kW without redundancy)</td>
</tr>
<tr>
<td>17 Barreleye G2 Servers</td>
</tr>
<tr>
<td>41 OU total</td>
</tr>
</tbody>
</table>

*48 V Bus Bar Clip*
Amphenol / TE
Increase chamfer to improve Lining with Bus-bar

Barreleye G2 48V Rack
Cloud Evolution
The Road to 10x

OpenPOWER
2014: Rackspace joins OpenPOWER
2015: 1st OpenPOWER based OCP design
2017: Zalus / BG2 OCP Contribution

Pilot
2018: Barreleye G2 Limited Access Beta

Open Compute
2011: Rackspace joined OCP

Accelerators
2017: Started qualifying Co-processors
TECHNOLOGY UPDATE
HOT-SWAP STORAGE, FANS, VGA

OPEN SOURCE: OPENBMC / OPENPOWER

POWER9

PCIE GEN 4 MEZZ 2.0 2X 100GBE

TRI-MODE (SAS – SATA–NVME), OPENCAP

OPENCAP / NVLINK 2.0

ACCELERATORS

48V POL ON MOTHERBOARD

EFFICIENCY

SERVICEABILITY

FIRMWARE

ALT. ARCH

RACK & POWER

STORAGE / MEMORY

NETWORKING

48V OPENRACK V2 WITH ATS
PCle G4/Mezz 2.0

Recap

- 6 PCle G4 slots
- 2 x8 slots
- 3 x16 slots
- 1 x16 / x8 OCP Mez 2.0 slot
- (x8 if using LOM+VGA mux card)

<table>
<thead>
<tr>
<th>SLOT TYPE</th>
<th>QTY</th>
<th>TOTAL THROUGHPUT (UNI-DIR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>x16 Gen4</td>
<td>3</td>
<td>31.5 Gbytes / sec (252 Gbps)</td>
</tr>
<tr>
<td>x8 Gen4</td>
<td>2</td>
<td>15.75 Gbytes / sec (126 Gbps)</td>
</tr>
<tr>
<td>x16 Mez A/B Gen4 or x8 Mez A Gen4</td>
<td>1</td>
<td>31.5 Gbytes / sec (252 Gbps) or 15.75 Gbytes / sec (126 Gbps)</td>
</tr>
</tbody>
</table>
PCIe G4/Mezz 2.0 Demo
Network & Storage Adapters

- GEN4 X16 NETWORK CARD
  Mellanox CX5 Mezz 2.0
- Gen3 Limitation: ~94 Gb/s

- GEN4 X8 STORAGE ADAPTER
  Eideticom NVM Express Offload
  Gen3 Limitation: ~6.8 GB/s

Result: 187.7 Gb/s
Result: 13.5 GB/s
PCIe G4 at Scale

Demo
PCIe G4 at Scale Demo

- **RECIPE**
  - 2 x Barreleye G2 servers
  - 8 “x16” PCIe Gen4 + 8 “x8” PCIe Gen4 NICs (Mellanox ConnectX5)
  - 100 G Switch

**Network Bandwidth**

1 Tbps

Server-Server
PCIe G4 Switch Ecosystem
Last Step for G4 Adoption
Barreleye G2 Storage
Tri-mode

- 24 Drives
- SAS / SATA / NVMe
- Hardware RAID on NVMe
- Tri-mode Hot-Swap

Interested?
Join me at 1:00pm at the OCP Storage Workshop for more in-depth information.
48 V Barreleye G2

Update

12V from Expander Board to Drive Plane

48V Input to MB

CPU VR
48V -> 0.7V – 0.8V

12V to Fans

12V to GPUs

48V OPEN RACK
BUS BAR & FAN ADAPTER
(Converts to 12V)

12V to Mezz / M.2 / SSD / BMC

DIMM VR
48V -> 1.2V

12 / 3.3 /5V
@ PCIe / Mezz / M.2 / SSD / BMC

12V from Expander Board to Drive Plane

48V – 12V
Power Brick

48V – 12V

* Zaius sled & shelf use only motherboard. Additional PCBAs not used.
** Barreleye G2 chassis uses Zaius motherboard and additional PCBAs shown here.
### OpenCAPI / NVLink Connector & Speeds

2 “Bricks” per CPU Socket

<table>
<thead>
<tr>
<th>DESCRIPTION</th>
<th>QTY</th>
<th>TOTAL THROUGHPUT (UNI-DIR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVLink 2.0 / OpenCAPI 3.0 Brick</td>
<td>1</td>
<td>25 Gbytes / sec (200 Gbps)</td>
</tr>
<tr>
<td>Bricks per Socket</td>
<td>2</td>
<td>50 Gbytes / sec (400 Gbps)</td>
</tr>
<tr>
<td>Bricks per Server</td>
<td>4</td>
<td>100 Gbytes / sec (800 Gbps)</td>
</tr>
</tbody>
</table>
Zaius / BG2 OpenCAPI & NVLink Interface

- 2 “Bricks” per CPU Socket
- 24G Slimline SAS Connectors
- (25GByte per Brick)
- Follows SFF-8654 Standard
OpenCAPI at Scale: Update

Demo with 4x OpenCAPI Cards

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Cycles</td>
<td>3,000,000</td>
</tr>
<tr>
<td>Good Resp Total</td>
<td>99.999%</td>
</tr>
<tr>
<td>Good Resp Store</td>
<td>99.999%</td>
</tr>
<tr>
<td>No cred cycles</td>
<td>0.001%</td>
</tr>
<tr>
<td>Metries - Loads</td>
<td>0.000%</td>
</tr>
<tr>
<td>Metries - Store</td>
<td>0.000%</td>
</tr>
<tr>
<td>Metries - Total</td>
<td>0.000%</td>
</tr>
</tbody>
</table>

**Image:**

A picture showing a server rack with OpenCAPI cards installed. The image also includes a data display showing various performance metrics such as cycles, loads, stores, and response times.
OpenCAPI at Scale: Update

Demo Details

• 4x OpenCAPI Cards
• Signal Integrity: 25.78125 Gbps (3x)
• Round Trip Latency < 80ns (~5x)
• Bandwidth Test: 88 GB/s (~700 Gbps)
• Coherent: No kernel Overhead
• Upstream Driver

OpenCAPI: Faster / Cooler PCIe
Generic PCIe Accelerators
Coherence tunneled via PCIe
Perpendicular Mount Accelerators

With Additional Coherent Attach

PERPENDICULAR MOUNT CARD WITH COHERENT ATTACH

Accelerator Card w/PCIe card edge & 25G Connector

PCIe slot

25G Cable

Co-Pro
Parallel Mount Accelerators
With Additional Coherent Attach

** Does not refer to OCP Mez

Top View of Parallel Mount card

PARALLEL MOUNT VIA MEZZ** WITH COHERENT ATTACH

Accelerator card (e.g. GPU, OpenCAPI, etc)
Supplier native connector
PCle to Native mez adapter
PCle slot

HBM 2
GPU or ASIC or FPGA
HBM 2

25G Cable
Accelerator Cards and Cable Mount Options*

- **Generic PCIe**: Accelerator Card w/PCIe card edge
- **Perpendicular Mount**: Accelerator Card w/PCIe card edge & 25G Connector, 25G Cable
- **Parallel Mount w/ Mez****: Supplier native connector, PCIe to Native mez adapter, 25G Cable

* Possible approach to co-processor card, mount, and cable designs.  
** Does not refer to OCP Mez standard.
VOLTA V100
NVIDIA
- 640 Tensor Cores, 250 W
- 16 GB HBM2
- Gen3 x16 Input
- FHFL card, fits with Riser
- 12 V power from fan board
- Deep learning

250S+ (Cabled)
Nallatech
- Gen4 x8 Host Connection
- FPGA Card with Cabled NVMe Storage
- Upto 12.8 TB
- Ideal for Coherent Memory Expansion

250S+ (On-Board)
Nallatech
- Gen4 x8 Host Connection
- 4x M.2 NVMe (x4) – On-Board
- Accelerated NoSQL Databases
FUTURE

Easy to Re-spin for changes
Leading to JBoG implementations
Solves the CPU-GPU Bottleneck

TESLA V100 SXM2
NVIDIA

300 W, 120 TFLOPS
Tensor Performance
16 GB HBM2, 640 Tensor Cores, 512 CUDA Cores
Ideal for GPU Analytics, Large Deep Learning Models (Kinectica, MapD, BlazingDB)

CPU-GPU NVLink 2.0
OpenPOWER

100 GB/s Bandwidth (Bidirectional)
3x Faster, Coherence
NVLink 2.0 on wire
2x SlimSAS Cables per GPU
OPENCAPI ACCELERATORS
ADM 9V3
ALPHA DATA

- x8 25G OpenCAPI
- Development Platform
- Offers Board Support Package
- HPC / Market Analysis,
- Smart NIC
- Inference

Innova 2 Flex
MELLANOX

- x8 25G OpenCAPI
- CX-5 + FPGA goodness
- Network Acceleration (NFV, Packet Classification)
- Security and Storage Acceleration

Flash Storage Accelerator
MOLEX

- x8 25 G OpenCAPI
- Coherently Attached with FPGA & ARM cores
- 8 NVMe (x4) via OpenCAPI
- Low latency Storage (Up to 16TB)
- Streaming (CDN), NoSQL, HPC Simulations, Hyperconverged Architecture (Like Microsoft Catapult)
Accelerator Composition

<table>
<thead>
<tr>
<th></th>
<th>Nallatech 250 S+</th>
<th>Alpha Data (ADM-9V3)</th>
<th>Mellanox Innova 2 Flex</th>
<th>Molex FSA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Input</strong></td>
<td>PCIe Gen4 x8</td>
<td>x8 25G</td>
<td>x8 25G</td>
<td>x8 25 G</td>
</tr>
<tr>
<td><strong>FPGA / Processor</strong></td>
<td>Xilinx Kintex Ultrascale+</td>
<td>Xilinx Virtex Ultrascale +</td>
<td>Xilinx Kintex Ultrascale + &amp; Connect-X-5</td>
<td>Xilinx Zync Ultrascale+ (with ARM A53 4x 1.5GHz)</td>
</tr>
<tr>
<td><strong>Number of LUTs</strong></td>
<td>523K</td>
<td>394 K</td>
<td>522 K</td>
<td>523 K</td>
</tr>
<tr>
<td><strong>Storage</strong></td>
<td>12.8 TB</td>
<td>-</td>
<td>-</td>
<td>16 TB</td>
</tr>
<tr>
<td><strong>Memory (DDR4)</strong></td>
<td>4 GB</td>
<td>32 GB</td>
<td>8 GB</td>
<td>64 GB</td>
</tr>
<tr>
<td><strong>Max Cards in G2 Chassis</strong></td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td><strong>Network</strong></td>
<td>-</td>
<td>2 x 100GbE</td>
<td>2x 25GbE</td>
<td>2x 100 GbE</td>
</tr>
<tr>
<td><strong>Power and Initialization</strong></td>
<td>PCIe Gen4 x8</td>
<td>PCIe Gen4 x8</td>
<td>PCIe Gen4 x8</td>
<td>48V Power / Fan Board</td>
</tr>
</tbody>
</table>

PCIe Card Edge (Optional)

Coherent Attach (Optional)
ACCELERATOR ENABLEMENT
SNAP
OpenPOWER
Development framework
Program in high-level languages

PowerAI
IBM
Deep learning made easy
Take advantage of CPU-GPU NVLink 2.0

CAPI Flash
IBM
API driven, In memory expansion
Redis, Cassandra, Neo4J

Solution Providers
Ex: Eideticom, Burlywood
Make end-to-end solutions or License IP
Others available via card vendors
Solving Bottlenecks . . .

. . . Easing Adoption

TOGETHER
Get on Board
Learn More

• Engineering Presentation, 2017 OCP Summit
  https://www.youtube.com/watch?v=2XJQvGX9yVE

• OCP Community Preview
  http://files.opencompute.org/oc/public.php?service=files&t=4302505c9d59160ec03aaf4ea67b146c&download

• OCP Server Mailing List
  http://lists.opencompute.org/mailman/listinfo/opencompute-server

• OCP Server Mail Archives
  http://lists.opencompute.org/pipermail/opencompute-server/
Participate

- Interested in consuming this hardware?
- Want to make the platform available to a wider audience?
- Are you an OpenCAPI & NVLink solution developer?
- Inquire about development samples

Reach out to:
adi.gangidi@rackspace.com
Design Package

- Available on OCP Server Wiki and GitHub
- Enhancements Coming Throughout 2018
- GitHub
  https://github.com/opencomputeproject/zaius-barreleye-g2 (don’t forget to install Git LFS)
- OCP
  http://www.opencompute.org/wiki/Server/Working#Open_Rack
Thank You