Open. Together.

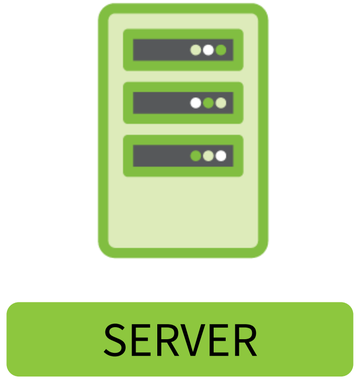
OCP
REGIONAL
SUMMIT

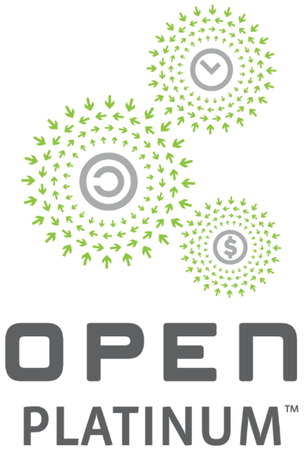
OAI Overview:

*An Open Accelerator Infrastructure Project for OCP Accelerator Module (**OAM**)*

Siamak Tavallaei
Principal Architect,
Microsoft, Azure
OCP Server Project co-Lead

Sept 27, 2019





Preface

Recognizing the need for a standard module form factor to accommodate accelerators from different suppliers, we developed the OCP Accelerator Module (OAM) spec and contributed it to OCP in March 2019 (Facebook, Microsoft, and Baidu). After presenting the OAM spec as a group effort at 2019 OCP Global Summit, we formed a subgroup in April and encouraged other OCP members to join a team effort to build a modularly interoperable infrastructure around OAM. Many companies have joined.

Open Accelerator Infrastructure (OAI) subgroup operates under OCP Server Project.

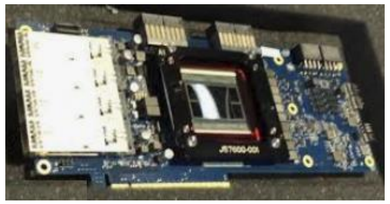
Under a joint development agreement (OAI JDA), the scope of work at OAI subgroup for the following 9 schedules is to define the physical and logical aspects such as electrical, mechanical, thermal, management, hardware security, and physical serviceability to produce solutions compatible with existing/traditional operation systems and frameworks to run heterogeneous accelerator applications. The OAI-JDA group will contribute the resulting specification to OCP at multiple revision levels (e.g., 0.3, 0.5, 0.7, 0.9, and 1.0)

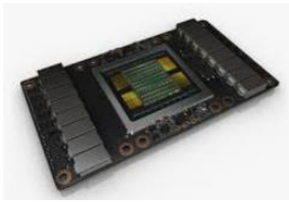
1. Open Accelerator Infrastructure (**OAI**)
2. OCP Accelerator Module (OAI-**OAM**)
3. OAI Universal Baseboard (OAI-**UBB**)
4. OAI Host Interface (OAI-**HIB**)
5. OAI Power Distribution (OAI-**PDB**)
6. OAI Expansion Beyond UBB (OAI-**Expansion**)
7. OAI Security, Control, and Management (OAI-**SCM**)
8. OAI-**Tray**
9. OAI-**Chassis** (This chapter will address **air-cooled** and **liquid-cooled** aspects as well.)

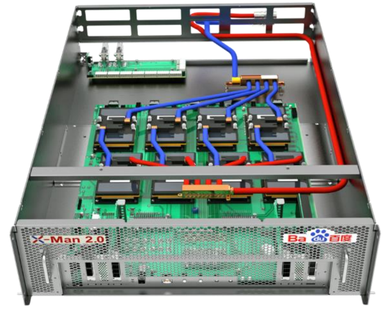
The research and development in
Artificial Intelligence (AI),
Machine Learning (ML), Deep Learning (DL), and
High-Performance Computing (HPC)
are driving rapid evolution in
new types of hardware accelerators

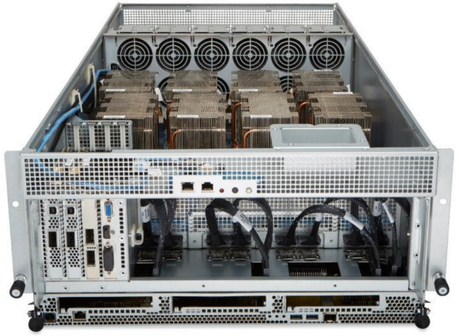
ASIC | FPGA | GPU | IPU | NNP | NPU | TPU | xPU…

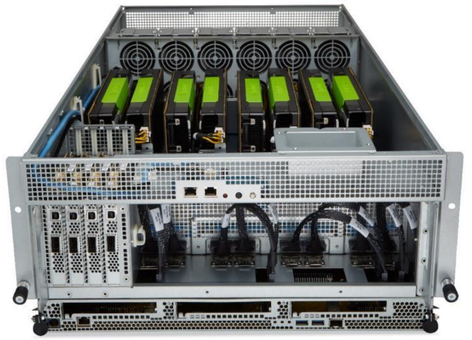
Diverse Module and System Form Factors











Different Implementations

Targeting Similar Requirements!

We need an

Open Accelerator Infrastructure

for these

Complex and Expensive Systems

Increase Interoperability

Accelerate Innovation

Via

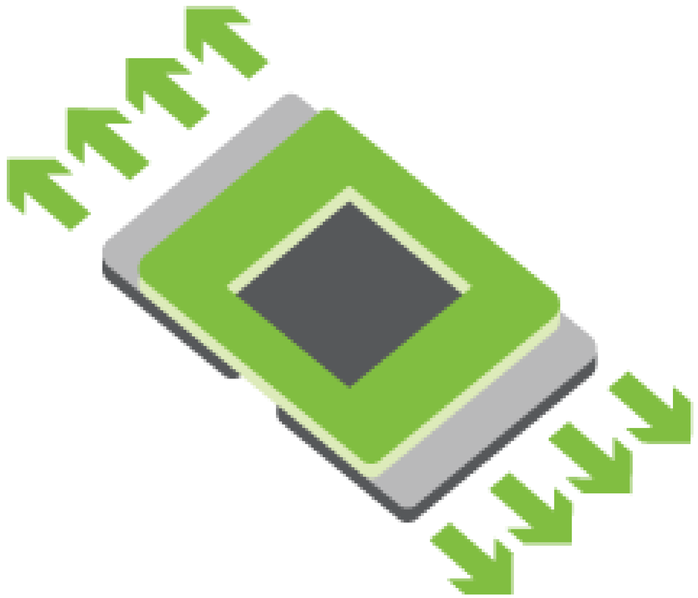
Modular Building Block Architecture (MBA)

We started with an *OCP Accelerator Module*

To go beyond what's possible with PCIe CEM form factor

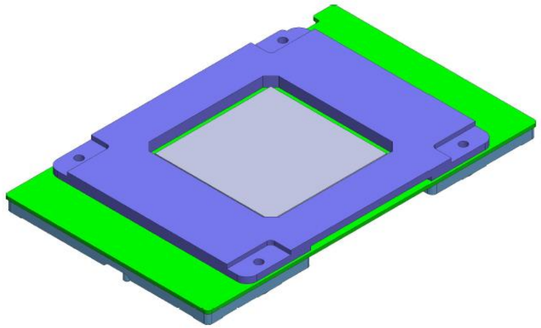
- High-density connectors → increase # of input/output Links
- Low signal insertion loss → high-speed interconnect
- Enough space for Accelerators and local logic & power
- Flexible heatsink design for air- and liquid-cooling
- Flexible inter-Module interconnect topologies

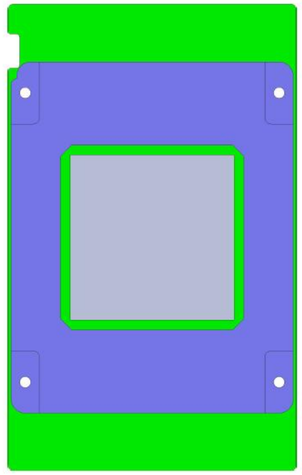
Current Work: **OAM** Spec (1.0)

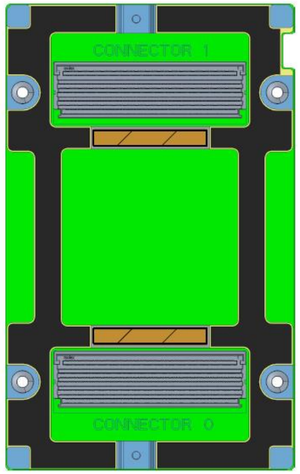


OCP Accelerator Module Spec

- 102mm x 165mm Module Size
- With two high-speed Mirror Mezz connectors (MPN: 2093111115)
- 12V and 48V input DC Power
- Up to 350w (12V) and up to 700w (48V) TDP
  - Up to 440W (air-cooled) and 700W (liquid-cooled)
- Support single or multiple ASIC(s) per Module
- Up to eight x16 Links (Host + inter-module Links)
  - Support one or two x16 High speed link(s) to Host
  - Up to seven x16 high speed interconnect links
- System management and debug interfaces

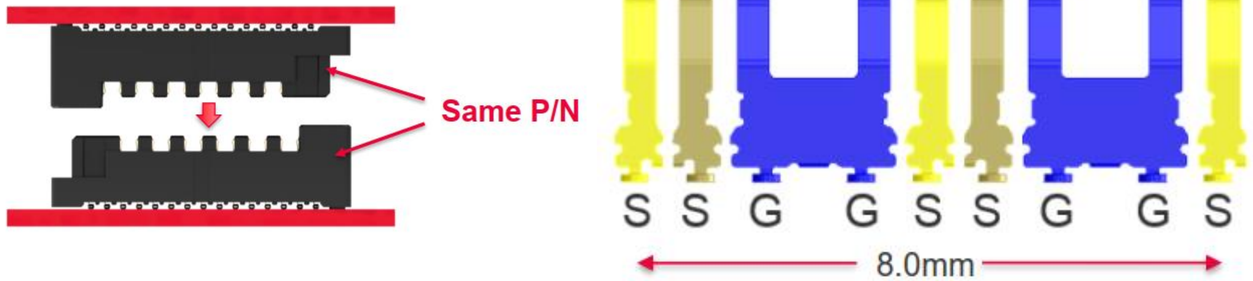




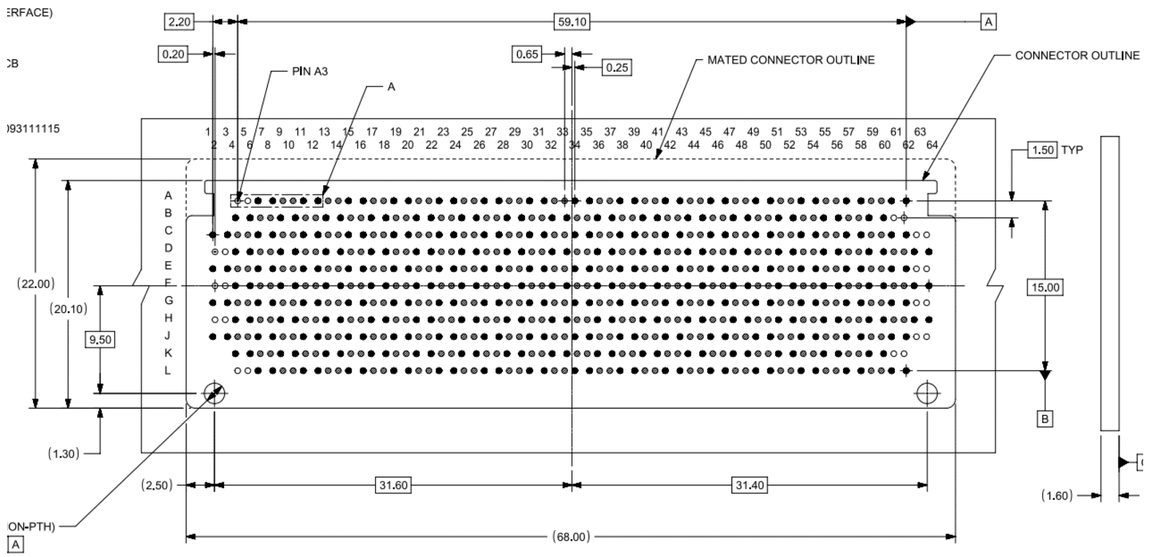


Molex Mirror Mezz Connector

- MPN: 209311-1115
- 68mm x 22mm after mating
- 172 differential pairs(161 non-orphan f
- 56Gbps or 112Gbps PAM4
- 1A/pin @1.5oz Copper after derating
- 90ohm+/-5%



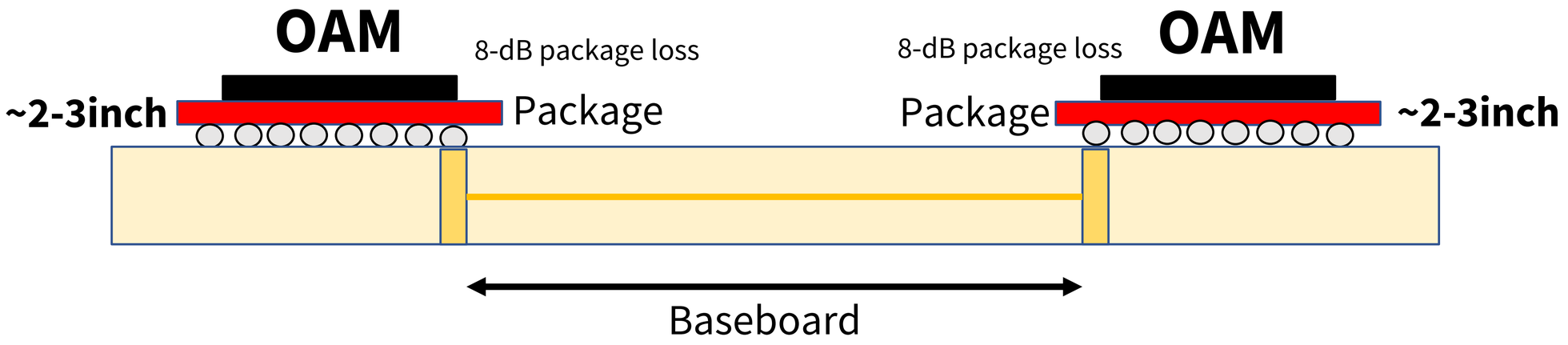
Images courtesy of Molex



Interconnect end-to-end Channel Loss

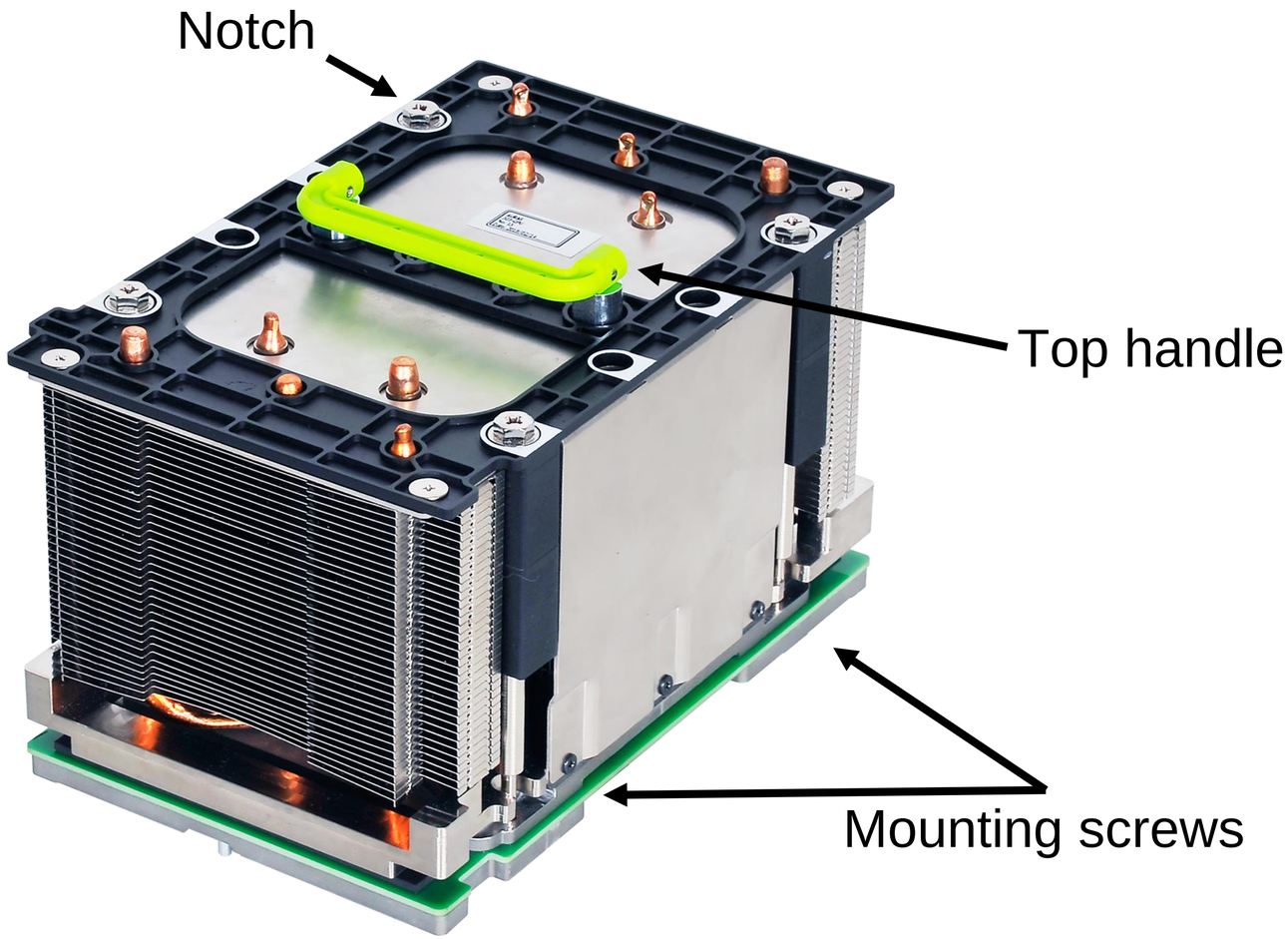
- The total insertion loss of the interconnect channel on the Module @28Gbps should be less than -8dB

- System baseboard IL budget = Die to Die IL from each OAM supplier – 16dB



HS Reference Design

- Heatsink reference design shown for 3U air-cooled system
- Top handle to accommodate handling for tight pitch and large weight (max 2kg)
- Long M3.5 mounting screw design for easy serviceability



Now, we are adding *Infrastructure Support*

- **OAM** is an Open Accelerator Module for multiple suppliers
- A multi-OAM, Universal Baseboard (**OAI-UBB**) for various Interconnect Topologies
- Host Interface Board (**OAI-HIB**)
- **OAI-Tray** for sliding a collection of OAMs (different UBBs)
- System Chassis, Power, and Cooling (different Trays)
- A Datacenter-ready, System- and Rack-level Security, Control, and Management (**OAI-SCM**) for all Chassis, Trays, UBBs, and OAMs as well as the Hosting Head Node

Open & Modular

in everyway!

Hierarchical Base Specification

Well-defined boundaries

Fostering Innovation while maintaining Interoperability

- Power and Cooling
- Mechanical
- Electrical
- Security & Management

- OAM
- UBB (Interconnect Topology)
- HIB
- PDB
- Tray, Chassis
- OAI-SCM
- Expansion

Designs and Products may be compliant to any or all specifications

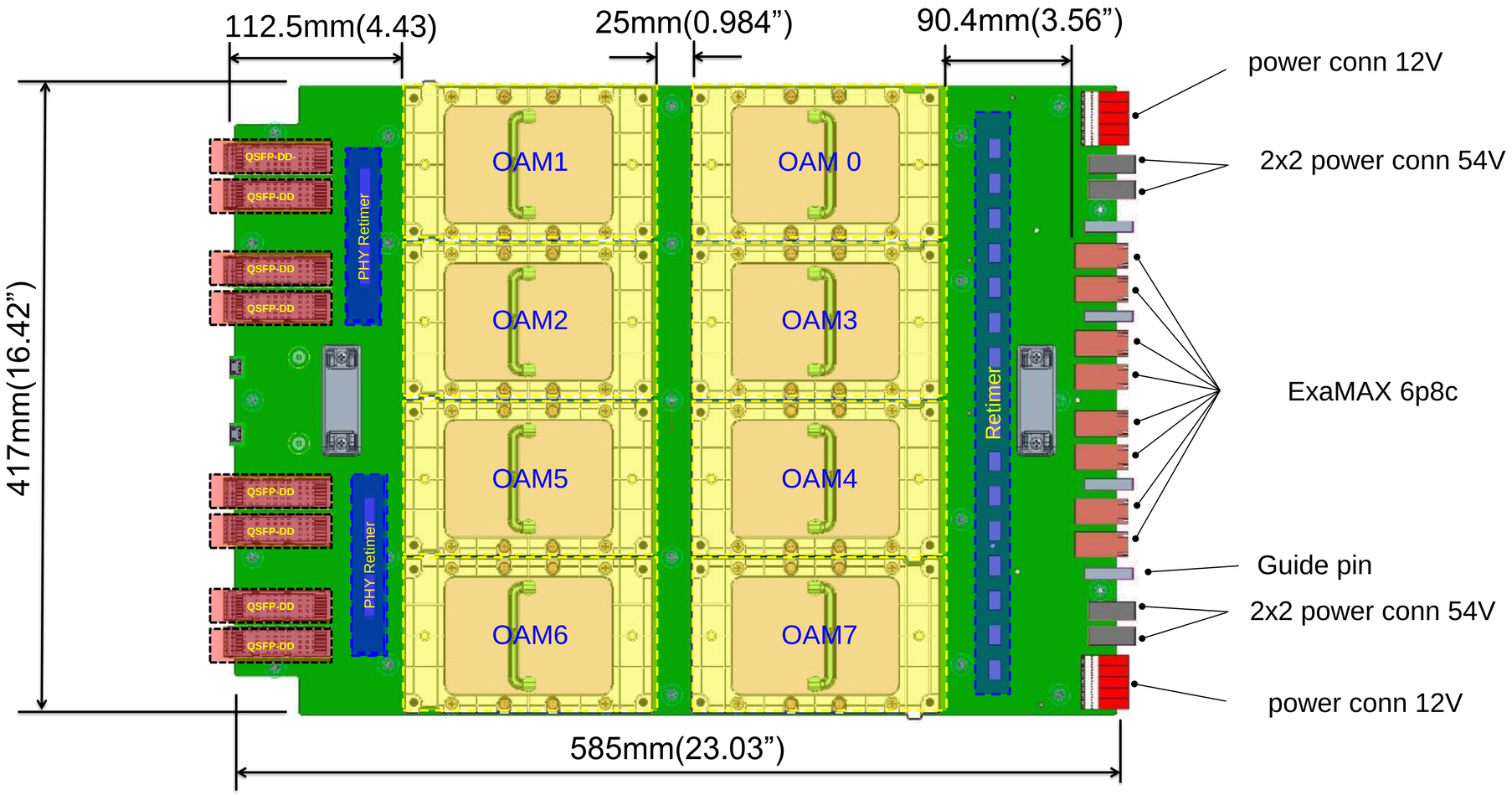
Well-defined boundaries (OAI)

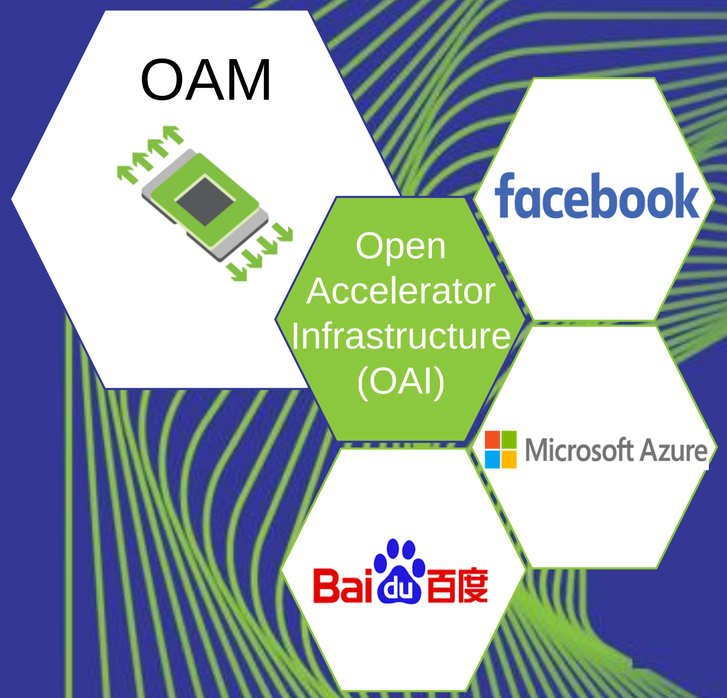
- Different manufacturers may offer **OAM**s with standard or propriety inter-OAM protocols
- **OAI-UBB** provides Host interface and native **Expansion** capabilities for eight OAMs
- **OAI-Tray** provides mechanical support to adapt various UBBs in 19” and 21” Chassis
- Modular power distribution allows 12V, 48V, and AC distribution to the Chassis
- **OAI-Chassis** supports Air- and Liquid-cooling in a modular way
- Rack-level Security and Baseboard Management (**OAI-SCM**)
- Each OAI Module is stateless; any FW or programmable code/logic is under RoT control
- Each OAI Module includes a FRU-ID to include vital product data (VPD)

Security, System Management, and Debugging

- RoT attestation
- Sensor reporting
- Error monitoring/Reporting
- Firmware Update
- Power-capping
- FRU Information
- IO Calibration
- JTAG/I^2C/UART interfaces for debugging

Universal Baseboard (UBB) with OAMs





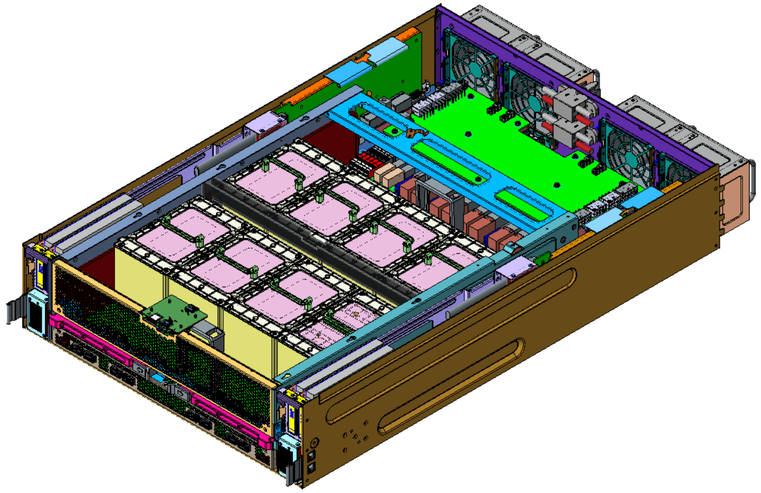
Framed an open-source infrastructure around OCP Accelerator Module

Invited others to join!

Formed the **OAI-JDA** Group
to Develop the Specification and Reference Designs

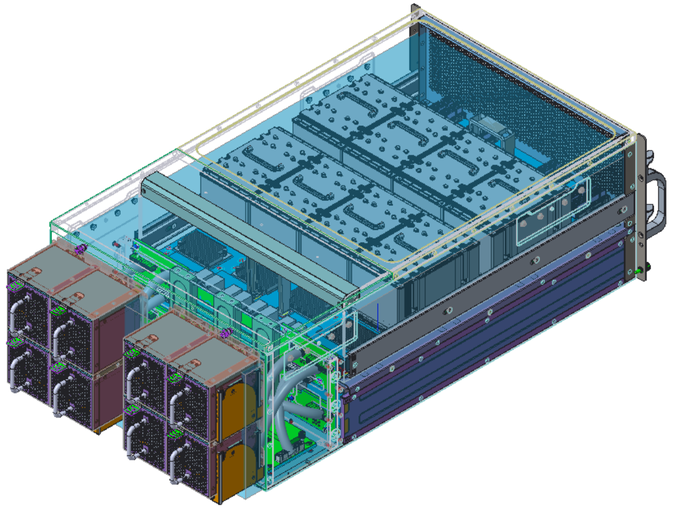
OAM Reference Designs

Inspur 21” Co-Planar system



- 21 inch 3OU, 34.6” (800mm) depth
- 8*OAMs
- UBB: Combined FC+ 6 port HCM Topology
- 4*PCIE Gen4 x16 Link to connect Hosts
- 4*PCIE Gen4 x16 Slots support 100G Infiniband or Ethernet for expansion

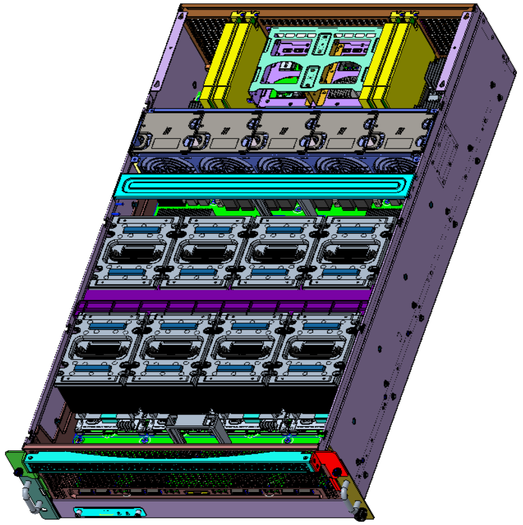
Hyve Design Solutions 19” Stacked System



- 19 inch 6RU, 30 inch (762mm) depth
- 8*OAMs
- UBB: Combined FC+ 6 port HCM Topology
- 4*PCIE Gen3x16 slots for host uplink
- 12*PCIE Gen3 x16 slots for flexible IO expansion

(PCIE interface will be revised to Gen4 in next release.)

ZT Systems 19” Co-Planar System



- 19 inch 4RU, 34.6” (880mm) depth
- 8*OAMs
- UBB: 8-port HCM topology
- 2*PCIE Gen4 x16 Uplinks for Multi-Host
- 4*PCIE Gen4 x16 Slots
- 4*2.5” NVME hot plug drives in front

OCP Server Project

<u>OCP Hierarchical Specifications:</u>
Base Specification
Designs Specification
Products

Adopters

<u>*Music:*</u>
Composer
Conductor
Musician

Audience

Call to Action

Get involved in the project:

OCP Server Project:	https://www.opencompute.org/projects/server
OAI subgroup:	https://www.opencompute.org/wiki/Server/OAI
OAI mailing list:	https://ocp-all.groups.io/g/OCP-OAI

Presenter

Siamak Tavallaei is a Principal Architect at Microsoft Azure, co-chair of OCP Server Project, and co-chair of CXL Technical Task Force. Collaborating with industry partners, he drives several initiatives in research, design, and deployment of hardware for Microsoft's cloud-scale services at Azure. He is interested in Big Compute, Big Data, and Artificial Intelligence solutions based on distributed, heterogeneous, accelerated, and energy-efficient computing. His current focus is the optimization of large-scale, mega-datacenters for general-purpose computing and accelerated, tightly-connected, problem-solving machines built on collaborative designs of secure, hardware, software, and management.

OCP REGIONAL SUMMIT

Open. Together.

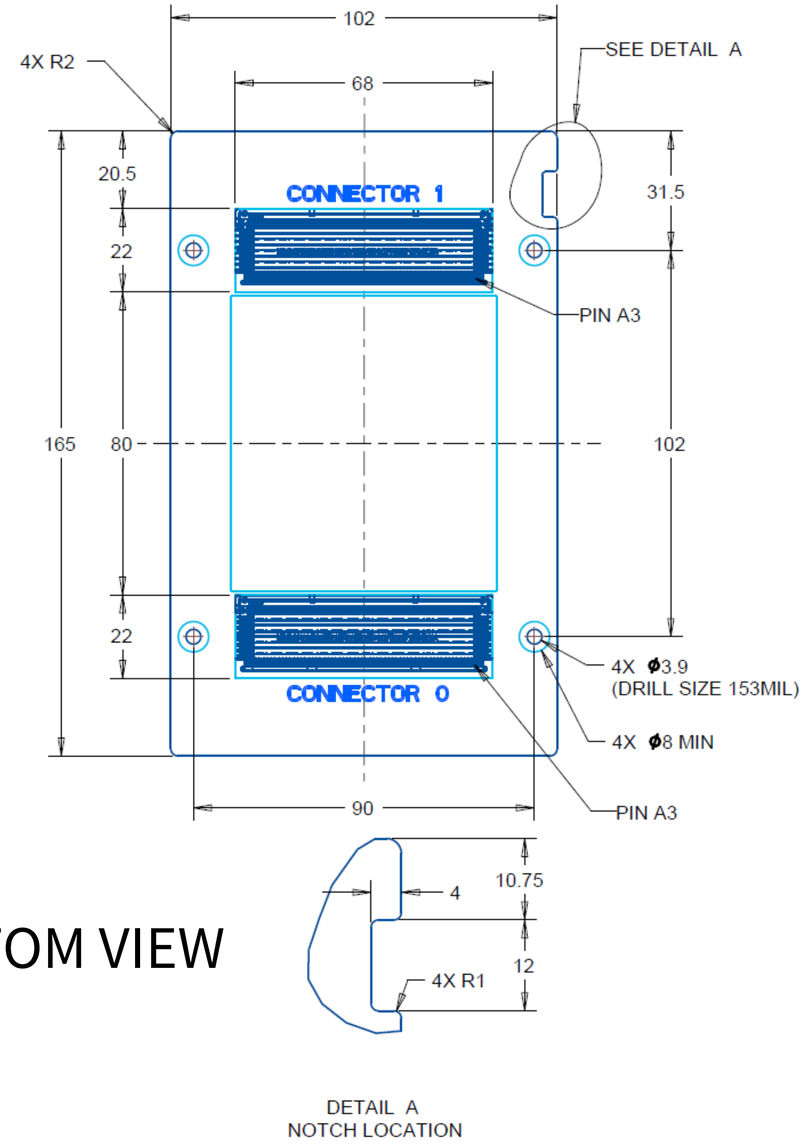
OCP Regional Summit

26–27, September, 2019

Backup

Mech Requirements – OAM PCB

- 102mm x 165mm footprint
- Connector pitch at 102mm
- M3.5 through holes with 8mm pad size
- Notch for alignment purposes



BOTTOM VIEW

Different Neural Networks

benefit from different

Interconnect Topologies

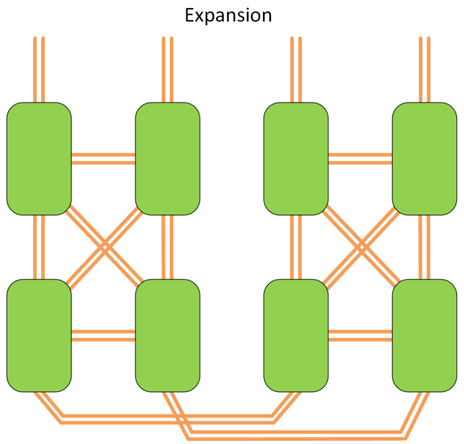
Universal Baseboard (UBB)

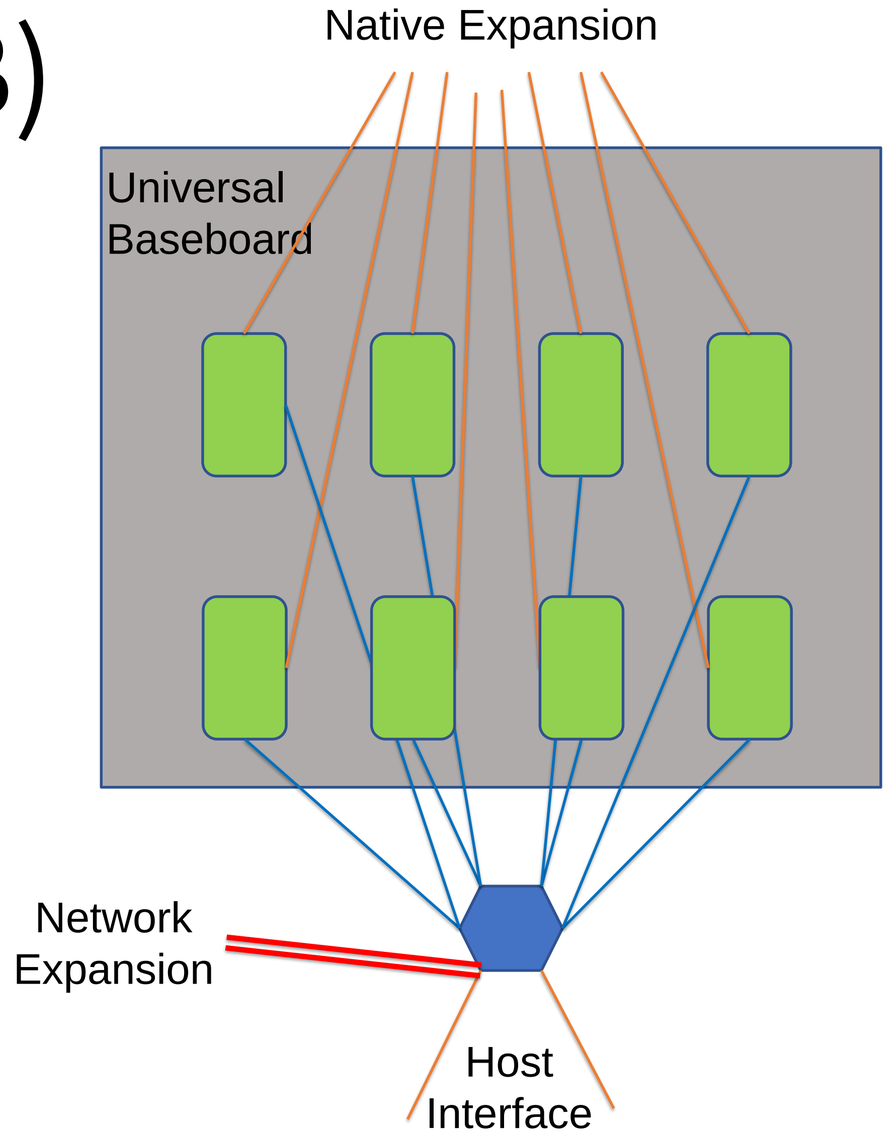
Consider a Grid of Planar OAM sites

Standard Volumetric

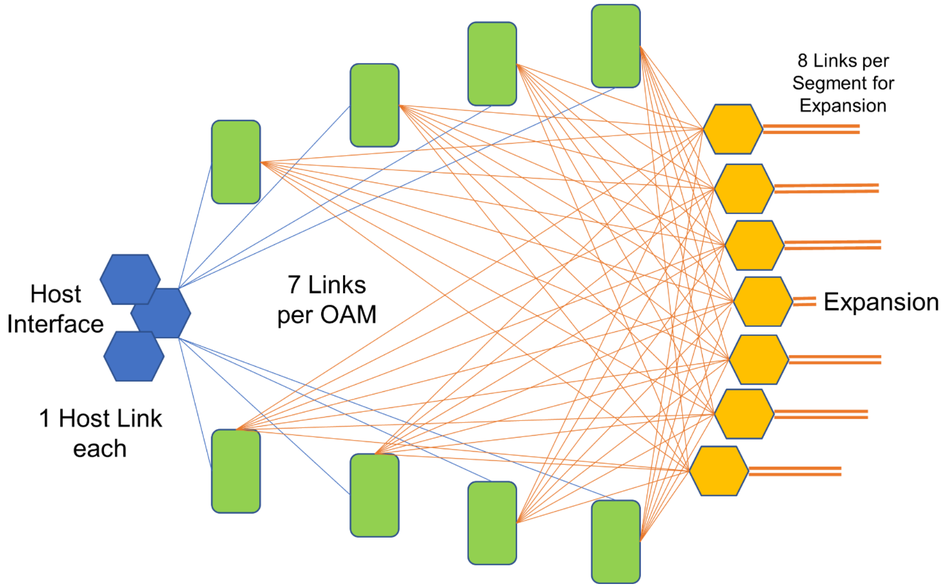
Protocol Agnostic Interconnects

Wires are Wires!

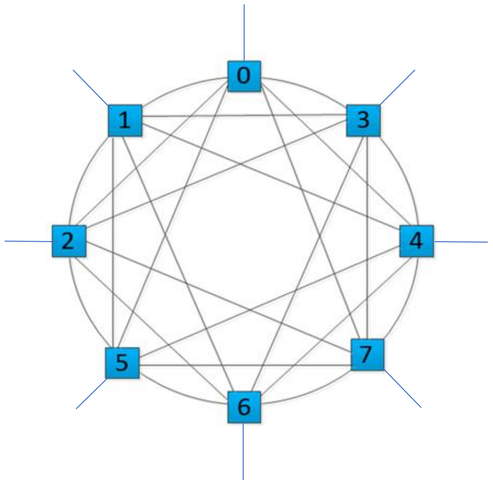




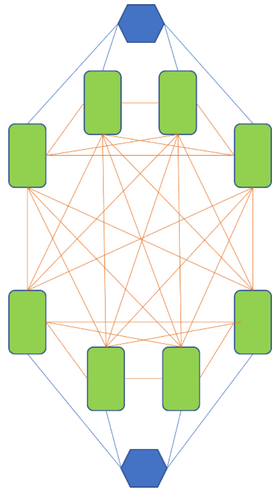
With different interconnect topologies



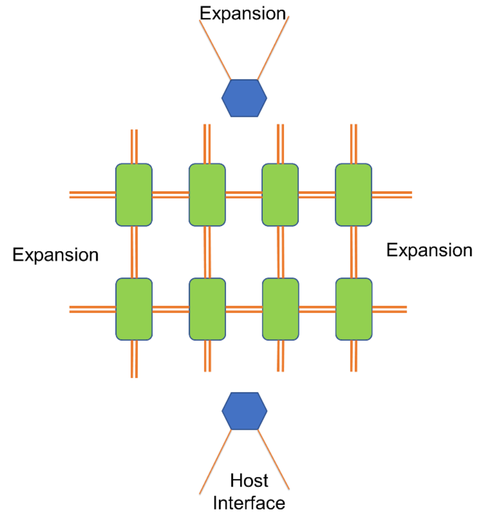
A Grid of interconnected OAMs,
Max Bisection BW
One Hop Away
Ready for Expansion



With six inter-OAM Links
and one Host Link



With seven inter-OAM Links
and one Host Link



Six inter-module Links may
create a 3D Mesh or Torus

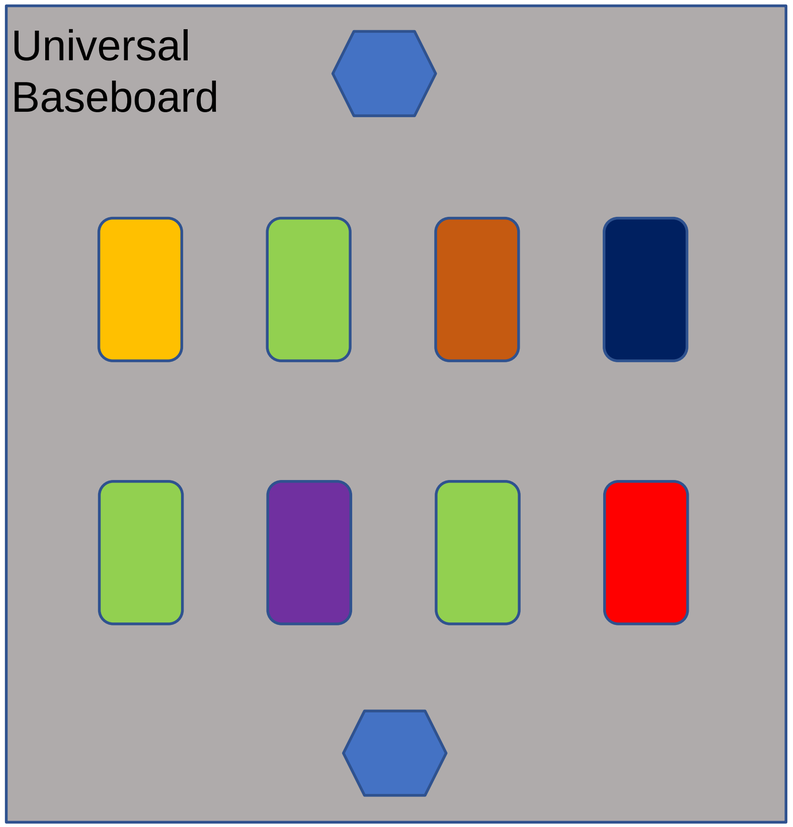
Heterogenous OAMs

These Modules need not be of the same type

Each one may be suited for a specific application/task

xPUs, FPGA, CPU, GPU, ASICs, SoCs, Memory, …

Chained, pipelined processing stages



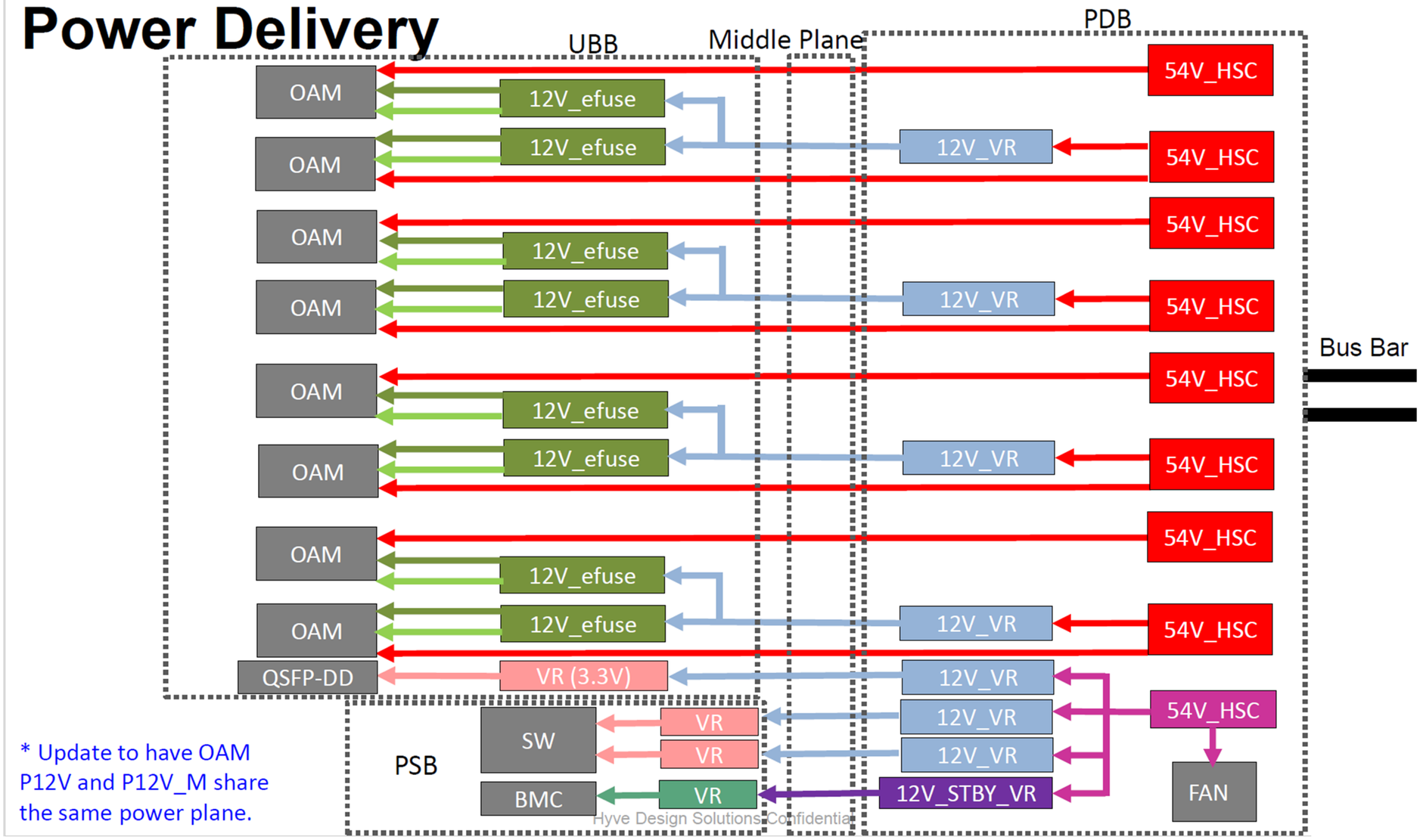
Well-defined boundaries (OAI-UBB)

- The Universal Baseboard (OAI-UBB) supports eight OAMs from different manufacturers
- Various UBBs support different interconnect topologies
- UBB provides independent power islands to OAMs
- UBB provides Host Interface for eight OAMs via eight x16 Links
- UBB provides Expansion Capabilities via eight x8 QSFP-DDs
- UBB managed by JTAG and I^2C
- To aid interoperability with other Modules, on its Host and Expansion interfaces, UBB provides signal isolation via Re-timers so that each connector of the same type (Host and Expansion) “sees” the same transmission line channel

Power Distribution Board (OAI-**PDB**)

Isolated Power Islands

Power Delivery



Well-defined boundaries (OAI-PDB)

- The Power Distribution Board (OAI-PDB) provides isolated power domains to UBB
- Hot-swap Controllers (HSC) and Fuses help monitor power consumption, voltage/current monitoring, and isolate power to reduce fault cases and the impact of faults

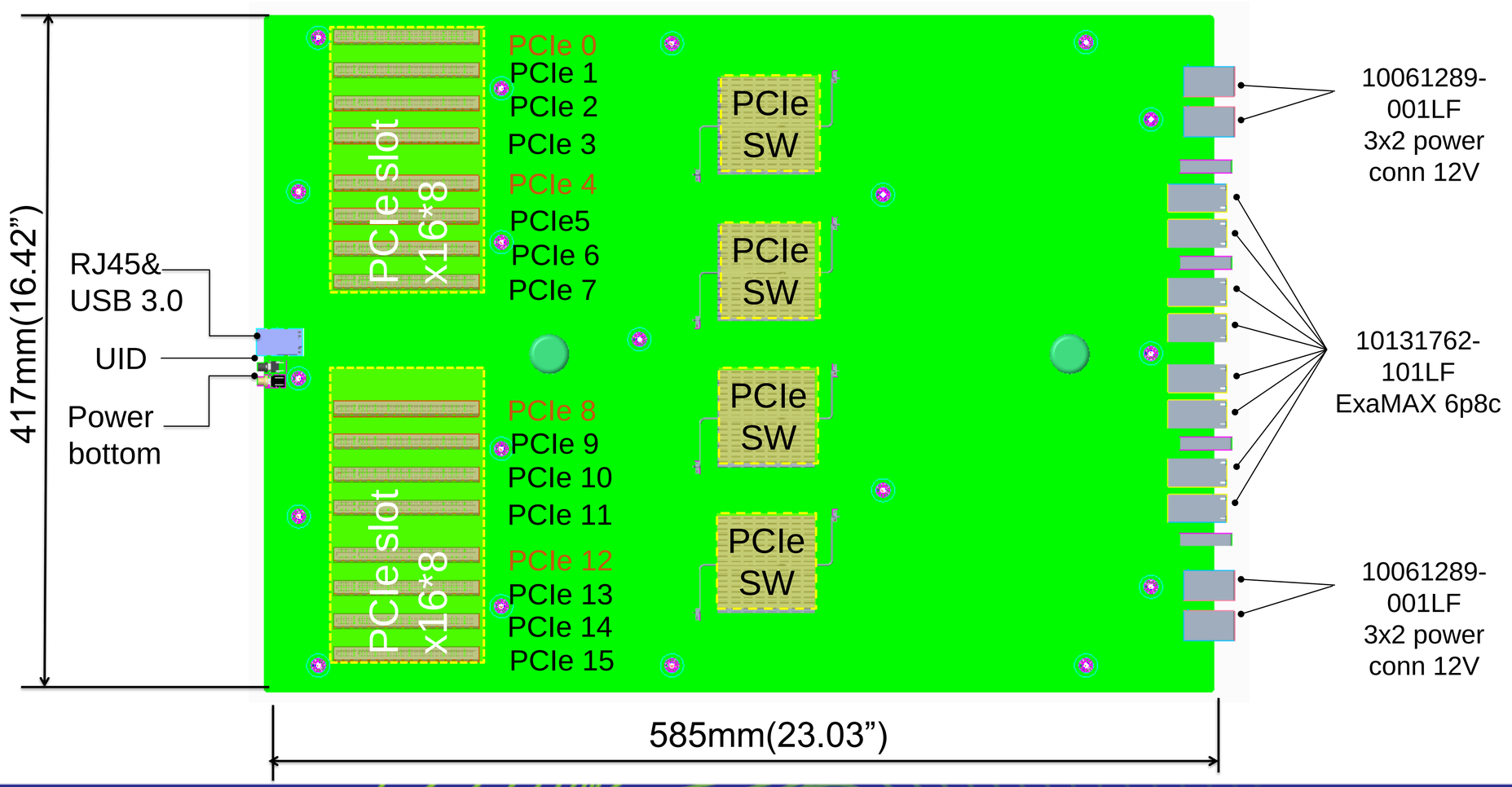
Host Interface Board
(OAI-**HIB**)

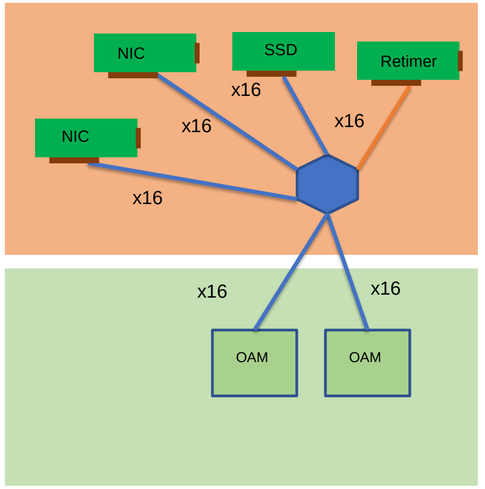
Different Host Nodes

provide different interfaces

PCIe, Infinity Fabric, CXL, etc.

HIB (Host Interface Board example)



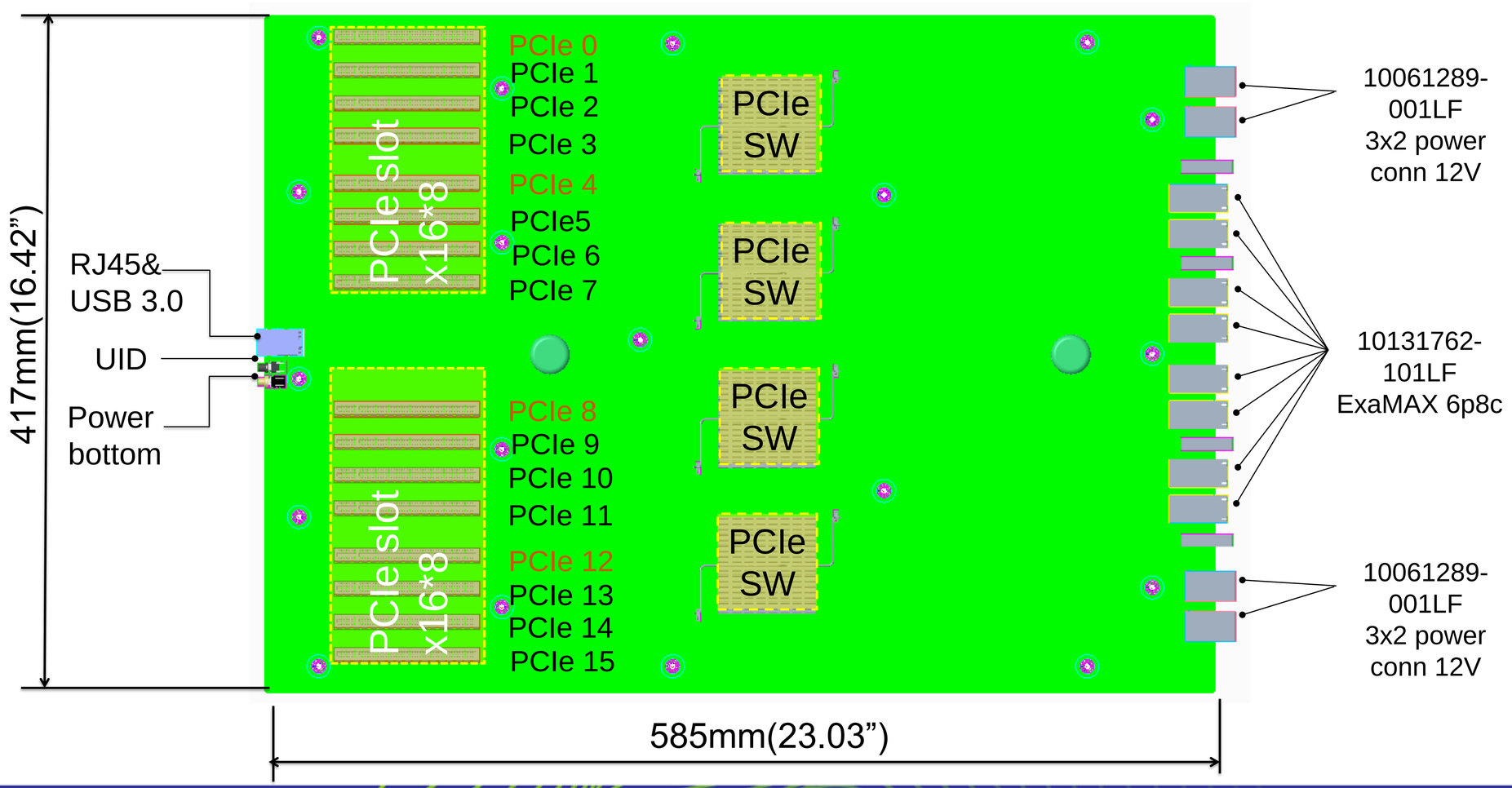


Well-defined boundaries (OAI-HIB)

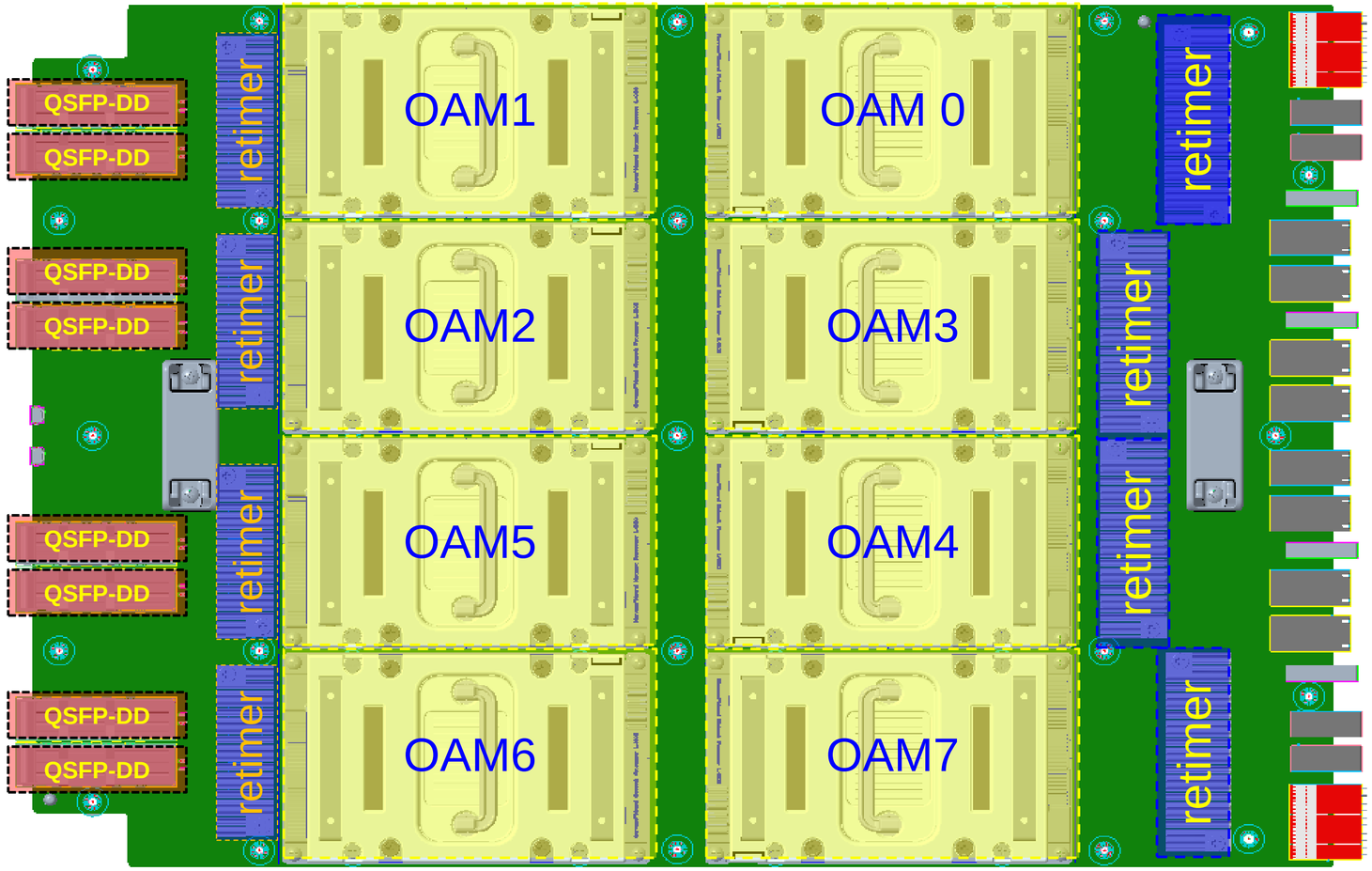
- The Host Interface Board (OAI-HIB) provides eight x16 high-speed Links such as PCIe Gen-4 to UBB
- HIB provides Clock, Reset, and PowerGood to UBB
- HIB provides security (root of trust– RoT), management, and control interface to UBB (OAI-SCM)

OAI
Expansion

OAI (Networked Expansion via Host Interface Board and PCIe Cards)

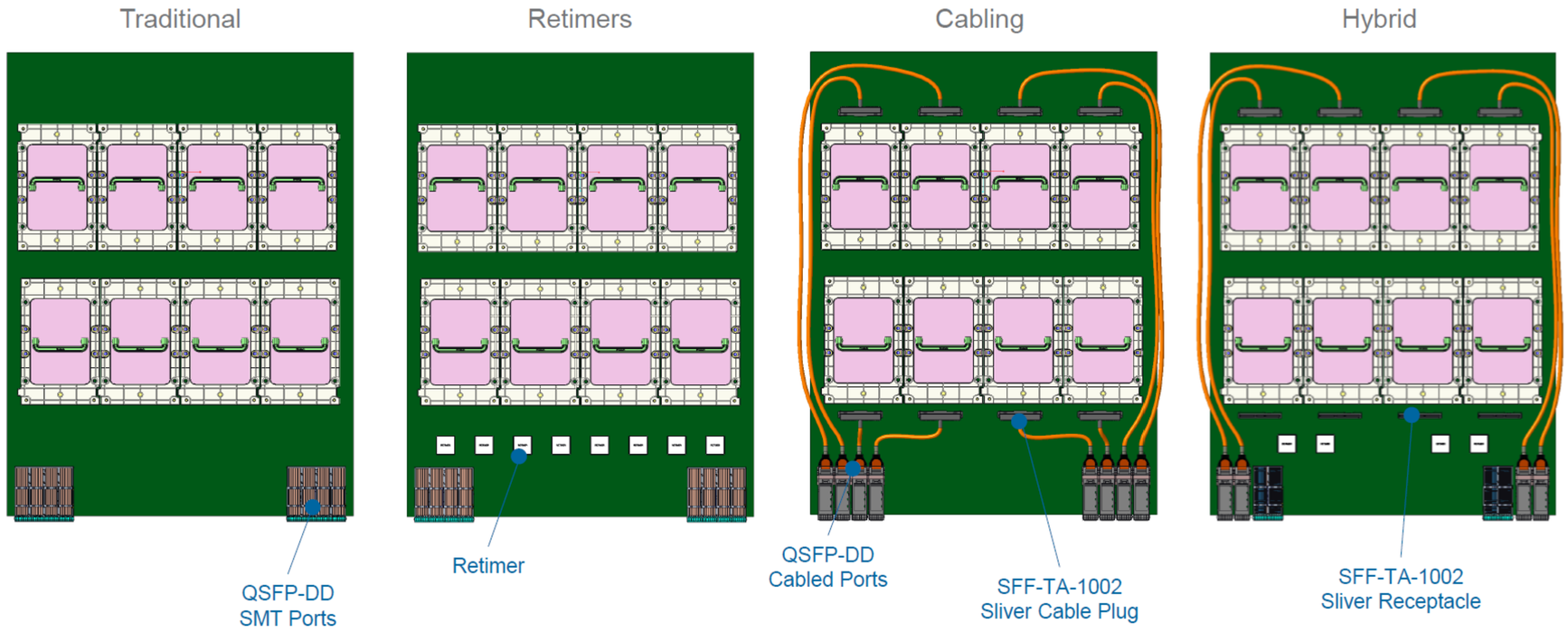


OAI Native Expansion via eight QSFP-DD Connectors



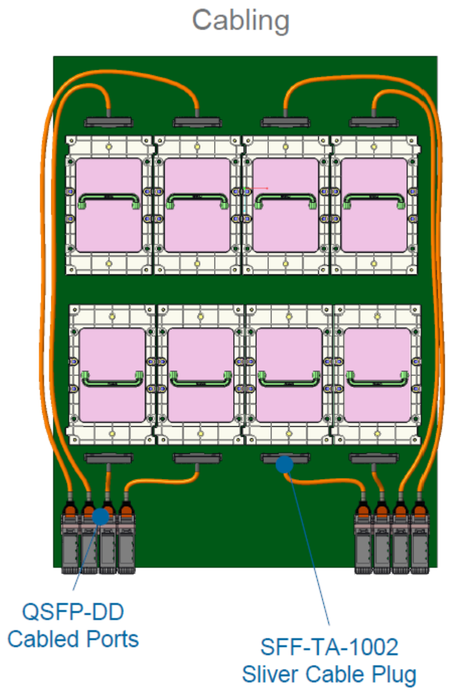
Hyve updated dwg

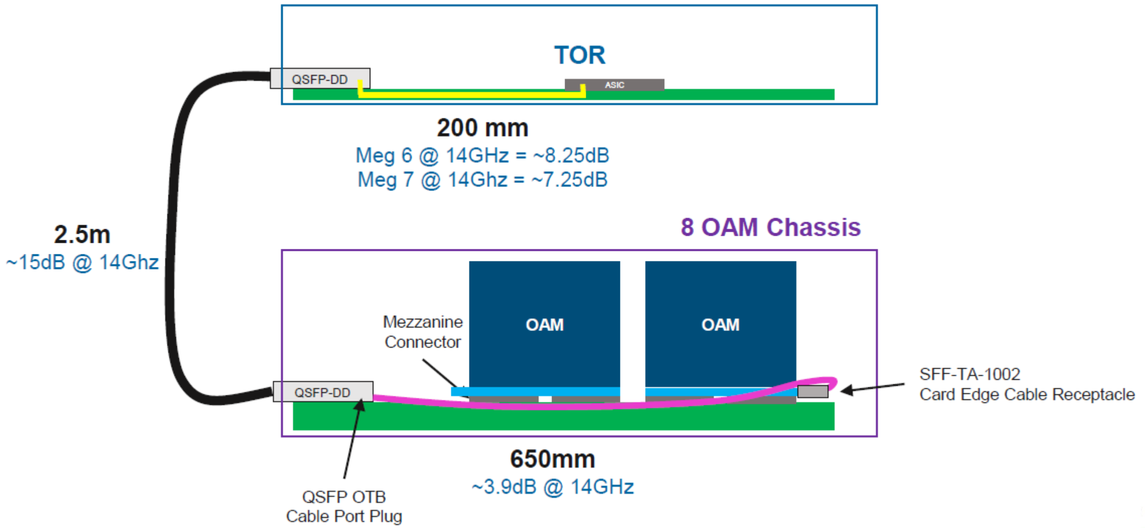
Reducing the Channel Loss for OAM Modules

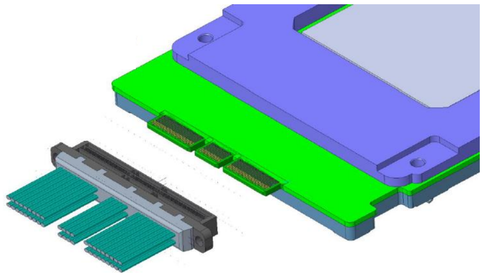


- Channel analysis to be completed to understand performance differences between traditional trace routing and cabling when linking the port to module

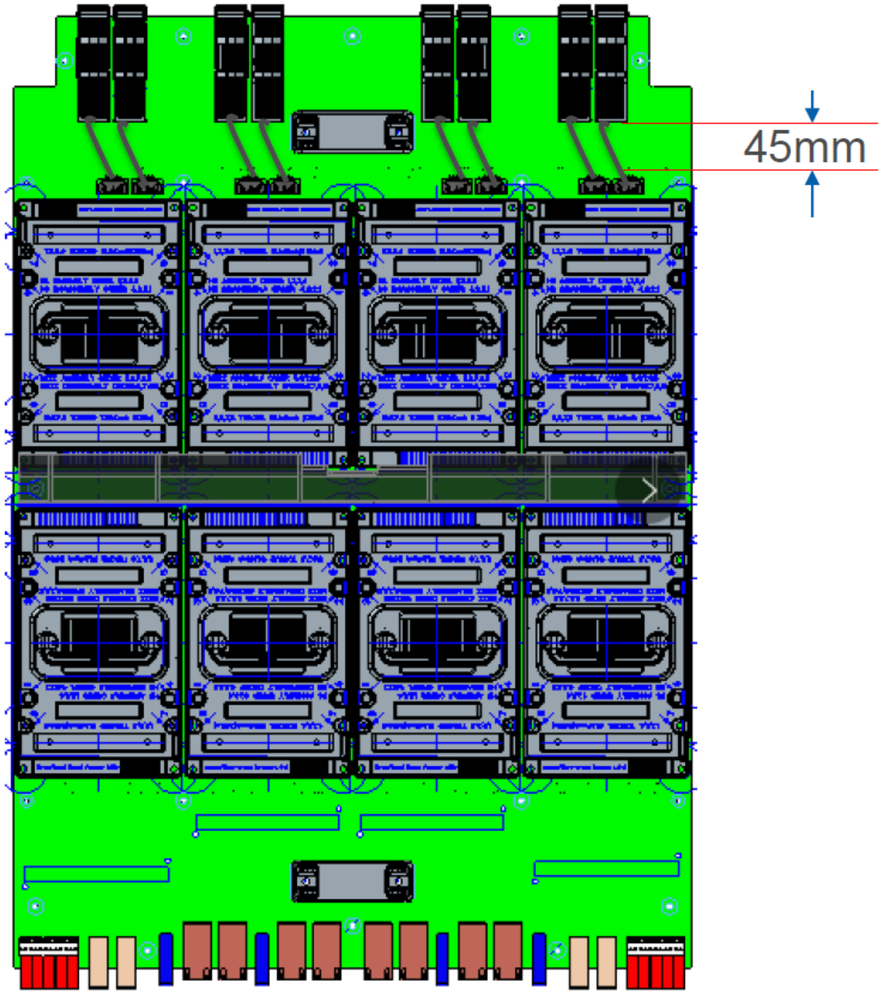
UBB Cabling from OAM directly to QSFP-DDs



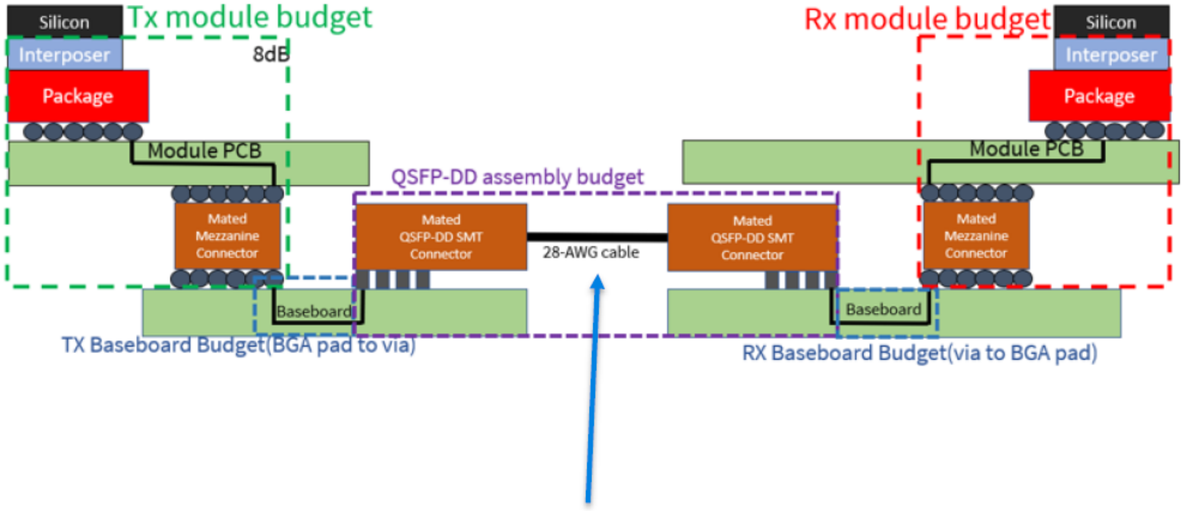




QSFP-DD BiPASS cable – 50.8 mm



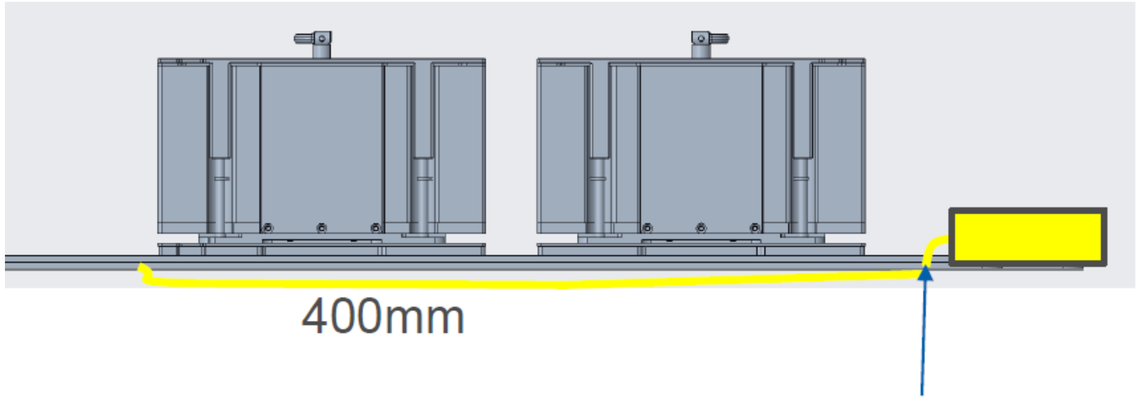
1. To increase about 2dB to expansion cable
2. Need vendor confirmed if the length is doable
   - Standard BiPASS cable minimum length is 115mm.

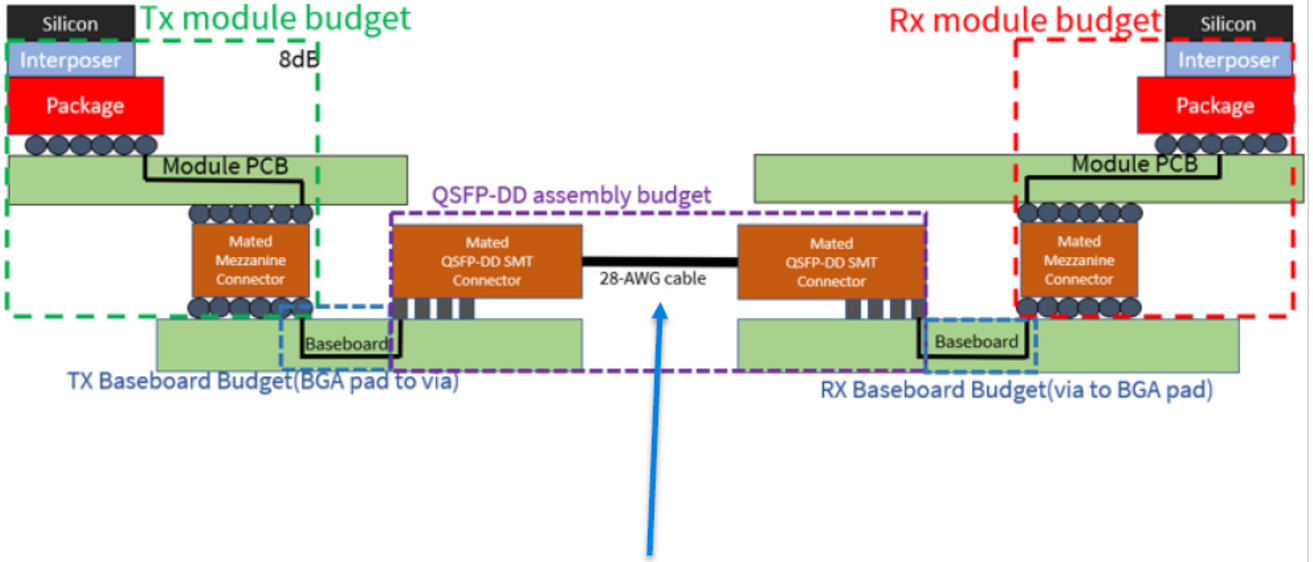


Passive cable support 675mm (4.84dB)

QSFP-DD with Flyover and cutouts PCB

SIDE VIEW





Total loss: OAM+2” trace+cable
(8+2.6+2.98)x2 = 27.16
Cable budget: 30 – 27.16 = 2.5 dB

Passive cable support 190mm (2.51dB)

Well-defined boundaries (OAI-Expansion)

- OAI-UBB provides network-based expansion via NICs in PCIe Slots of the OAI-HIB

- UBB provides native expansion via eight x8 connectors (Eight QSFP-DD connectors as the primary target)
- OAI-Tray provides mechanical space for sixteen QSFP (or QSFP-DD) connectors in two rows
- OAI-UBB provides enough space for re-timers close to the QSFP-DD connectors
- OAI-UBB provides enough space for cabled connection from OAMs to QSFP-DD connectors
  - Option 1: cabled from OAM to QSFP-DD without UBB-PCB traces
  - Option 2: cabled from near OAM Mezz connector to QSFP-DD (bottom side of UBB)
  - Option 3: cabled directly from OAM to QSFP-DD, QSFP, or OSFP