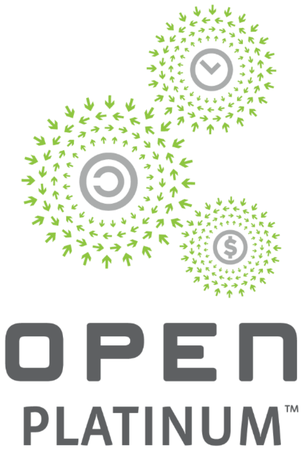

Open. Together.
OCP
SUMMIT

SERVER

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

Whitney Zhao, Hardware Engineer, Facebook
Siamak Tavallaei, Principal Architect, Microsoft
Richard Ding, AI System Architect, Baidu
Tiffany Jin, Mechanical Engineer, Facebook



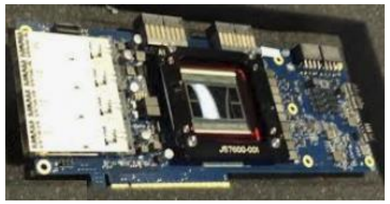
Outline

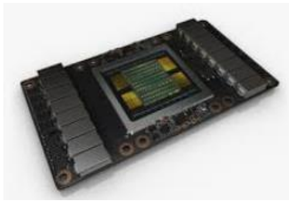
- Motivation
- Approach
- Examples
- Requesting Participation and Feedback

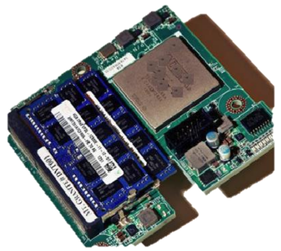
AI’s rapid evolution is producing an explosion of new types of hardware accelerators for Machine Learning (ML), Deep Learning (DL), and High-Performance Computing (HPC)

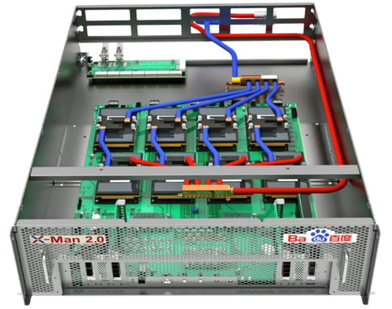
GPU | FPGA | ASIC | NPU | TPU | NNP | IPU | xPU…

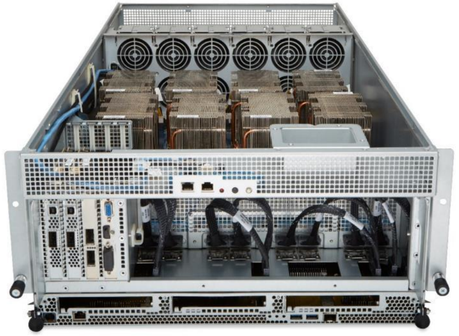
Diverse Module and System Form Factors

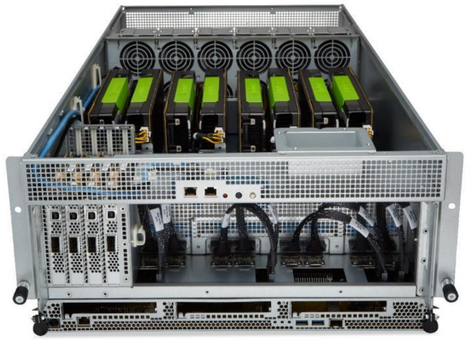








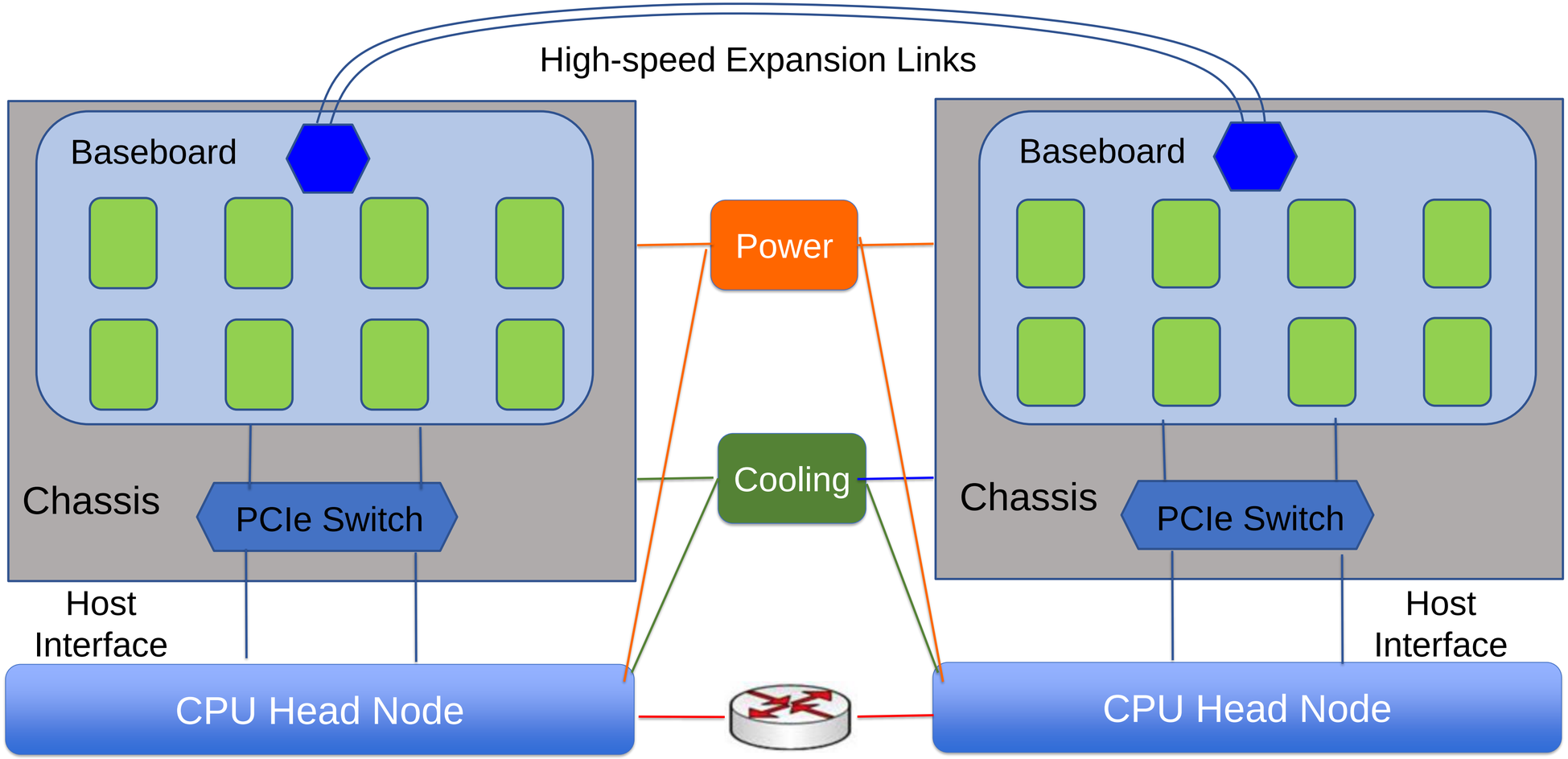




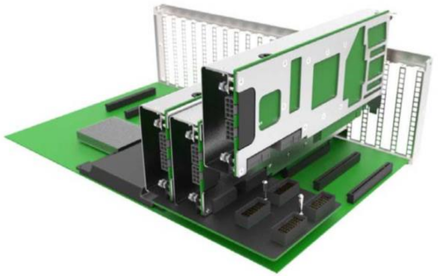
Different Implementations

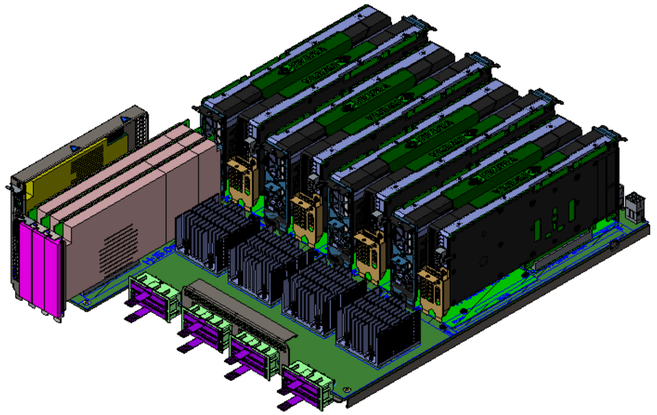
Targeting Similar Requirements!

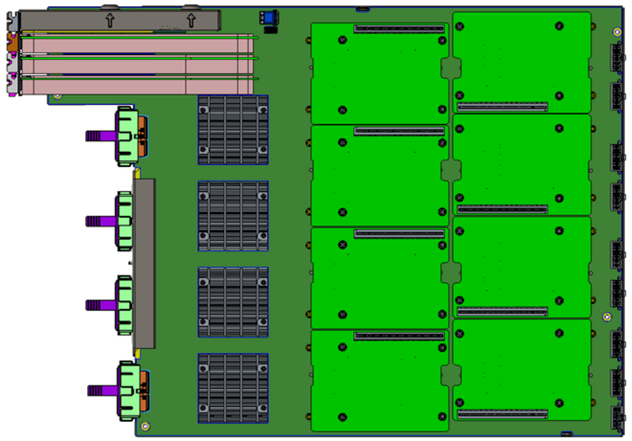
Logical Components for AI Hardware System

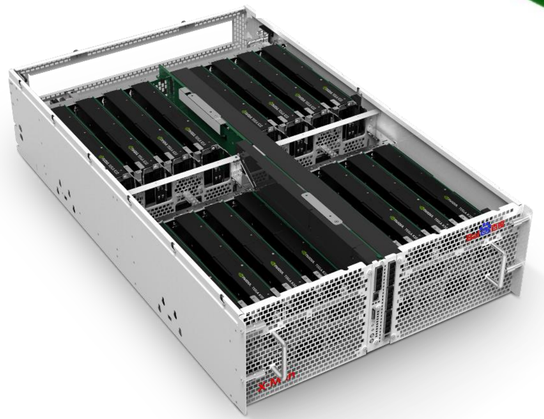


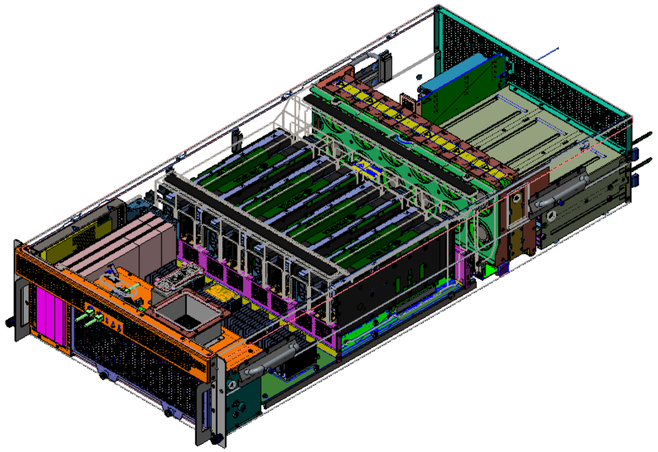
Accelerators in PCIe CEM Form Factor

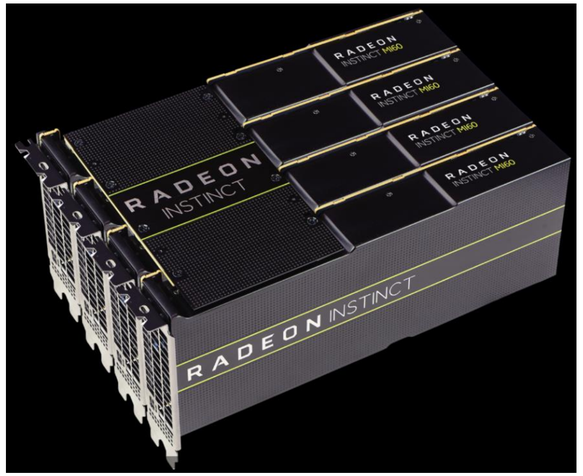


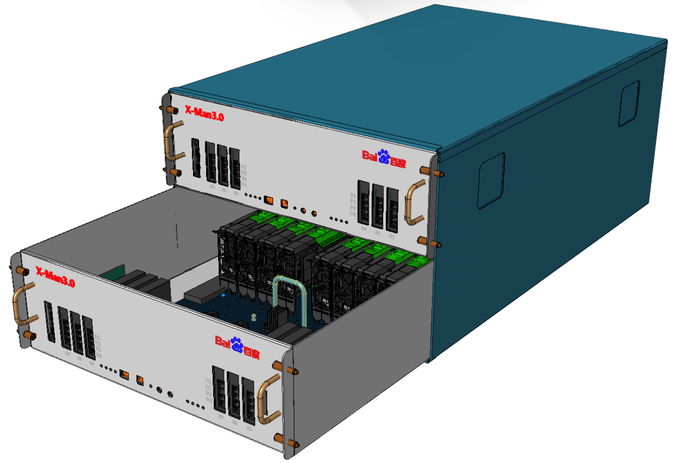




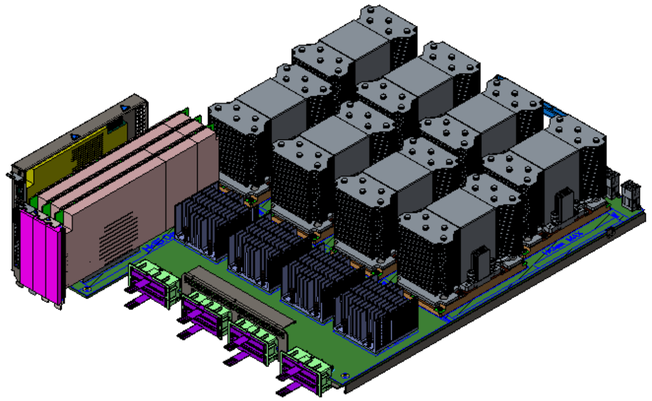


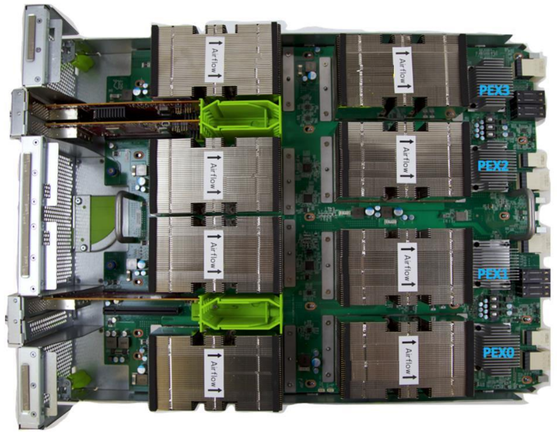


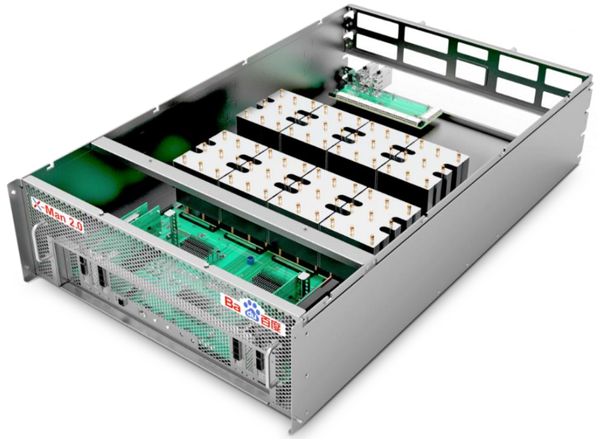


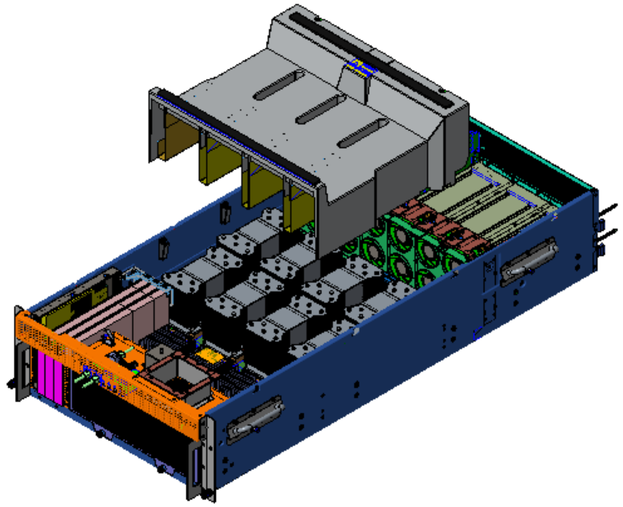


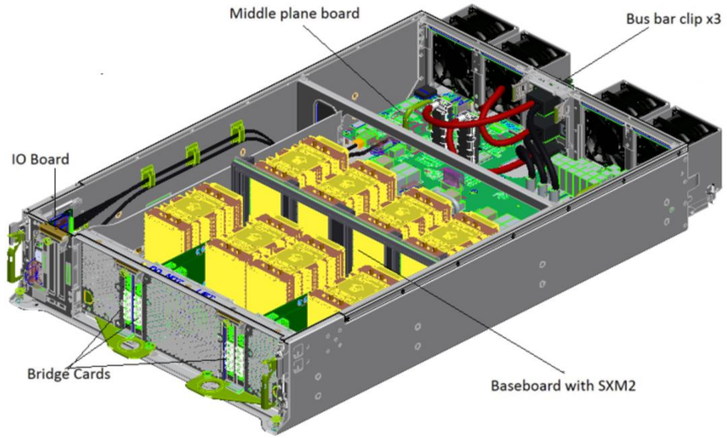
Accelerators in Mezzanine Form Factor on Baseboard

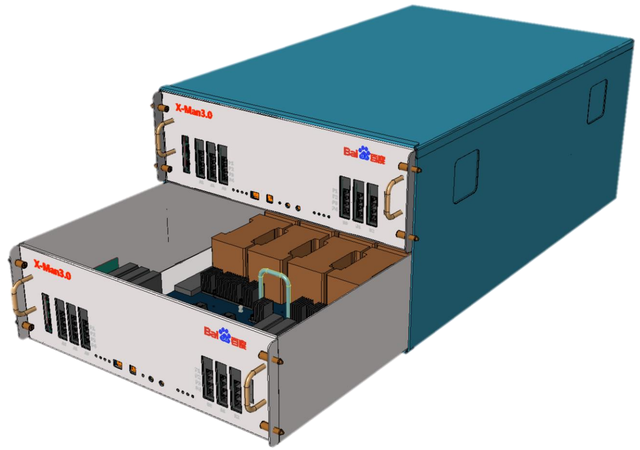




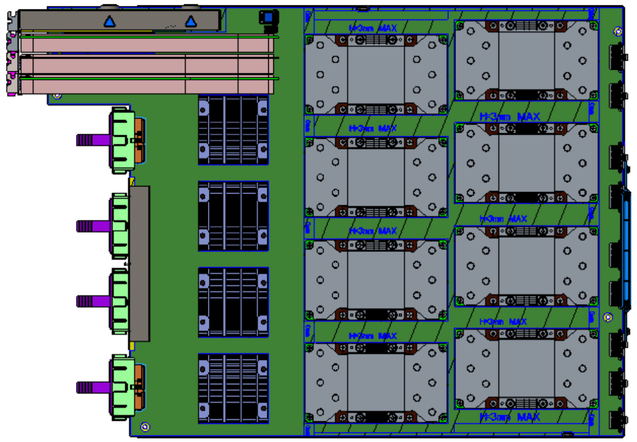


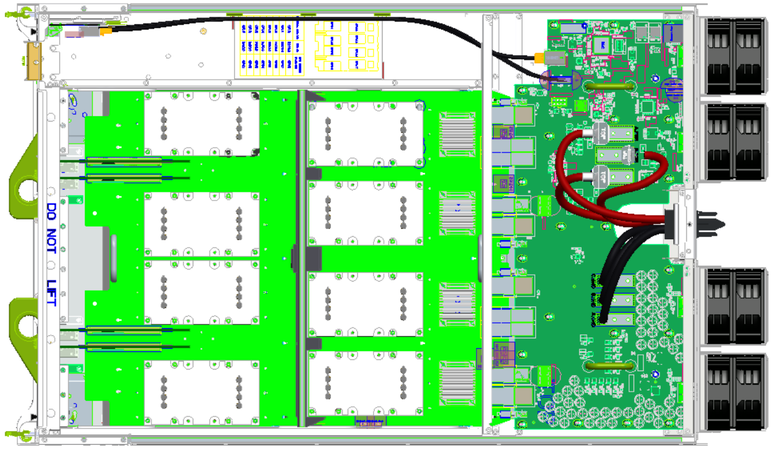


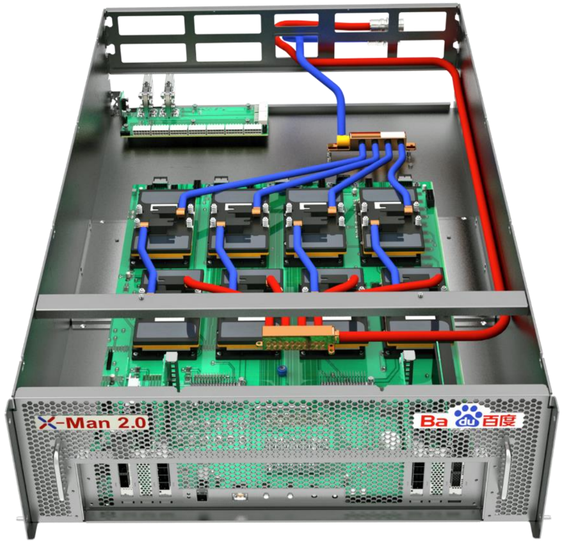


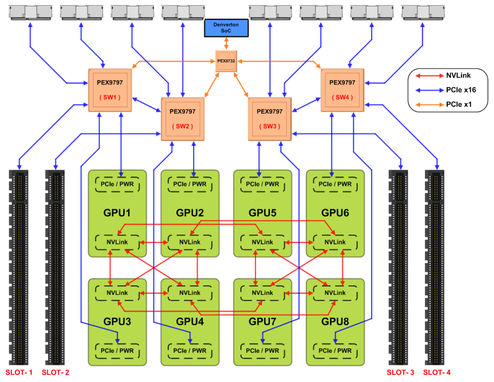


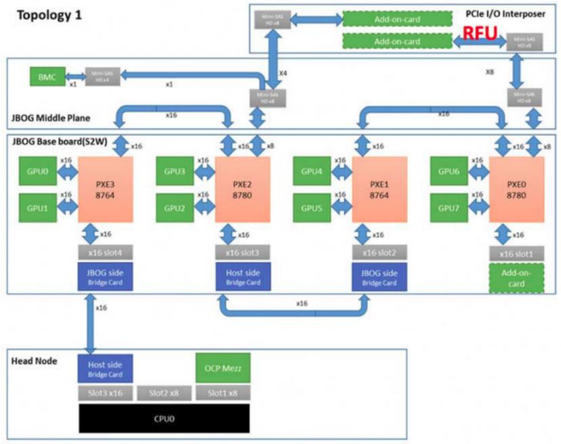
Various Interconnect Topologies

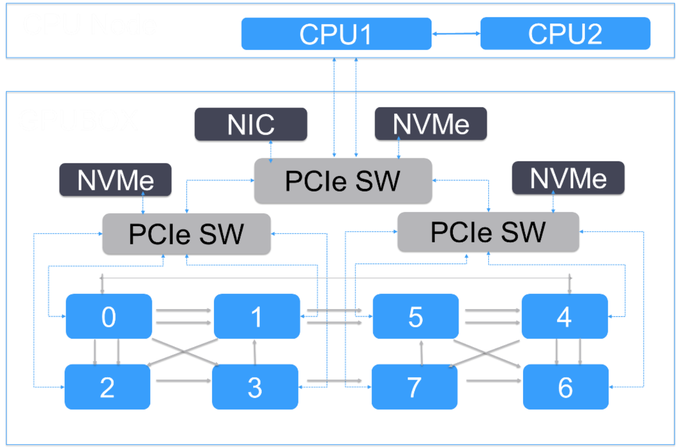




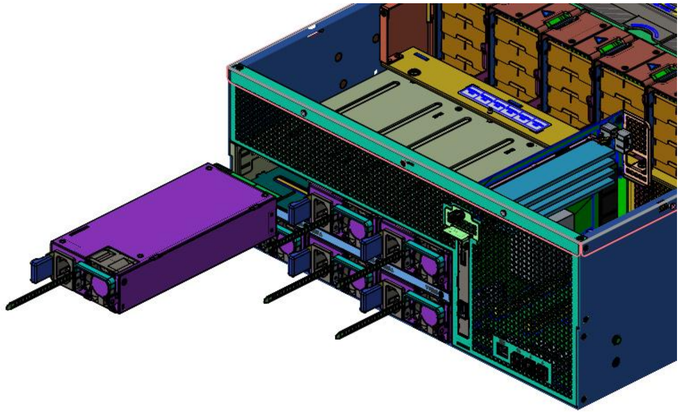


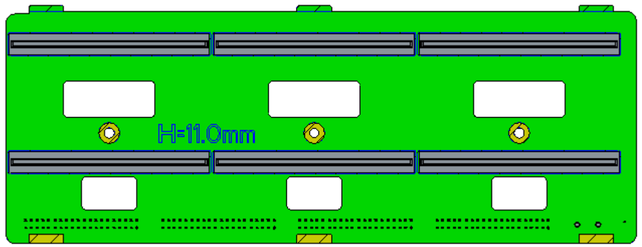


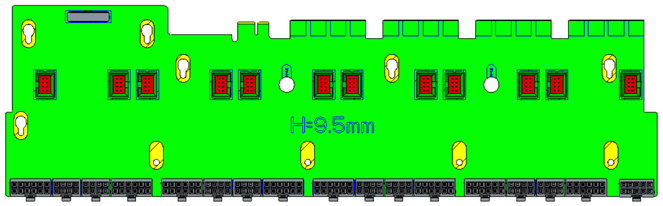




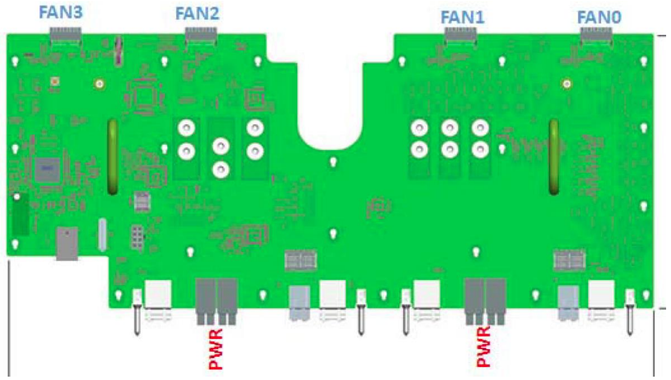
Power Delivery and Distribution



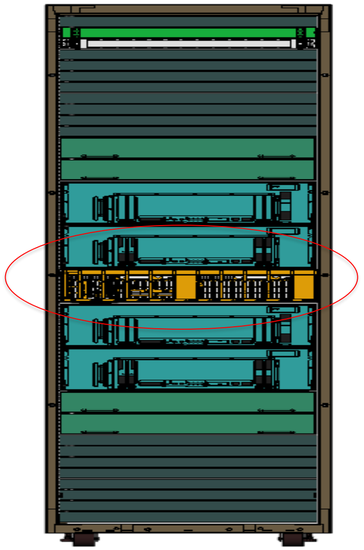




To Busbar

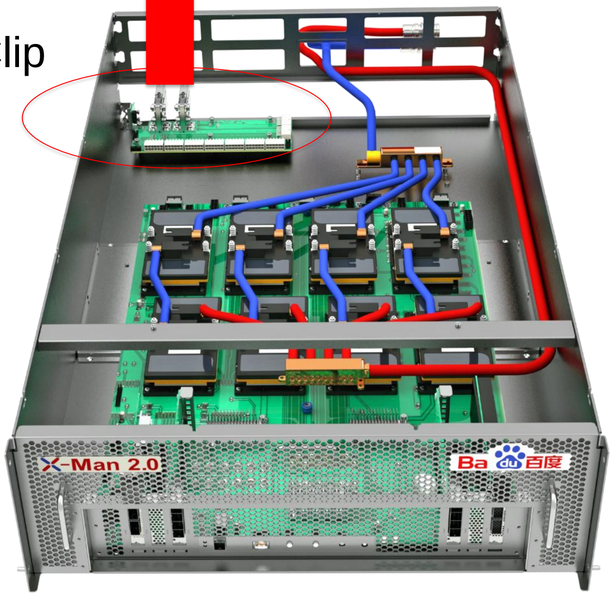


Shared Power Shelf

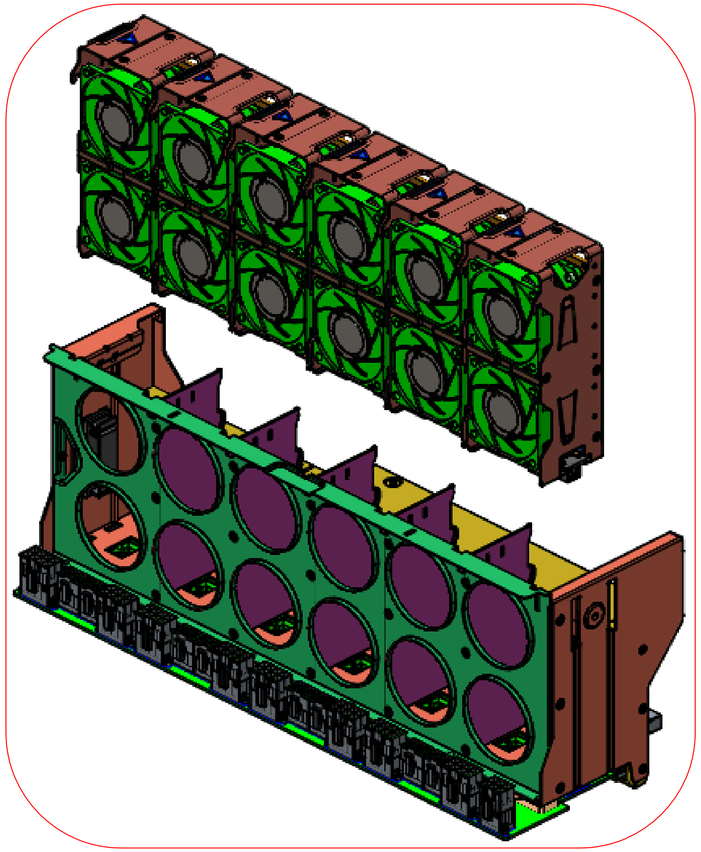


Bus Bar

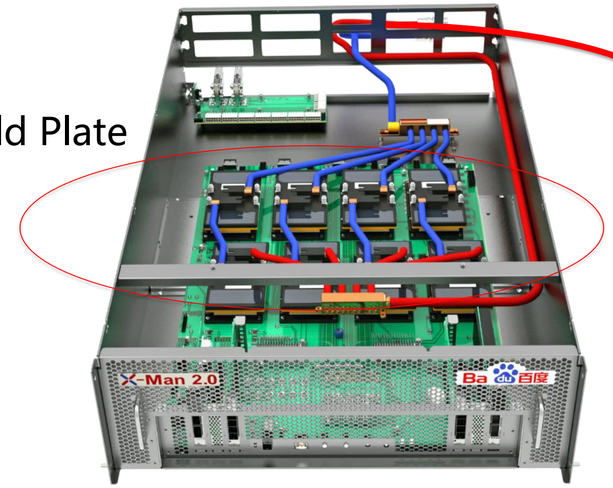
Clip



Cooling Methods

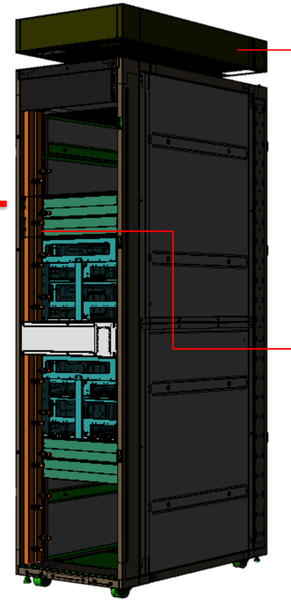


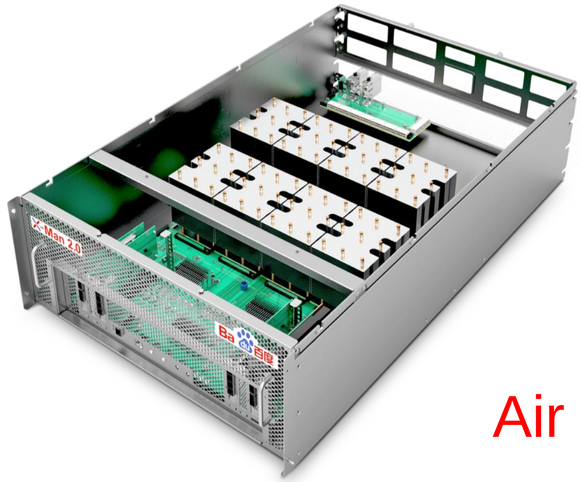
Cold Plate



+

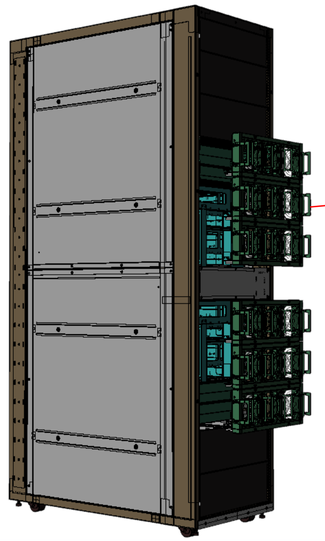
Liquid Cooling



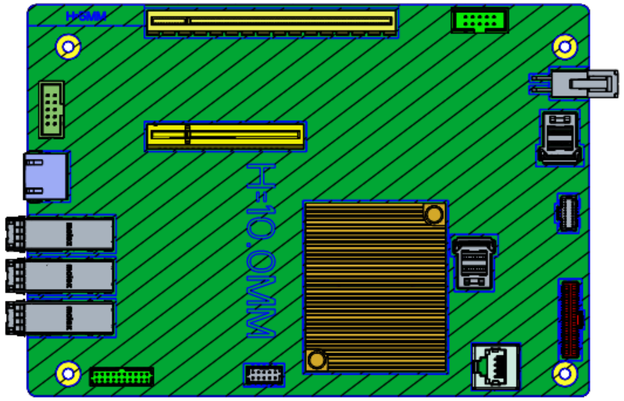


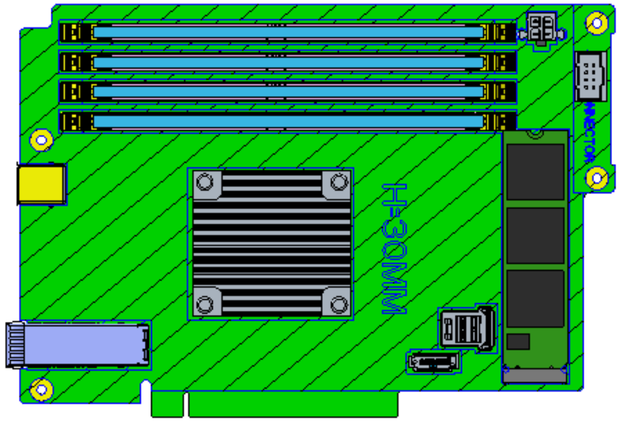
+

Air Cooling

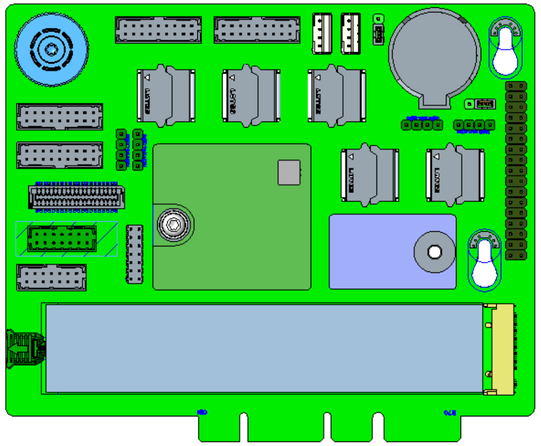


Management Module

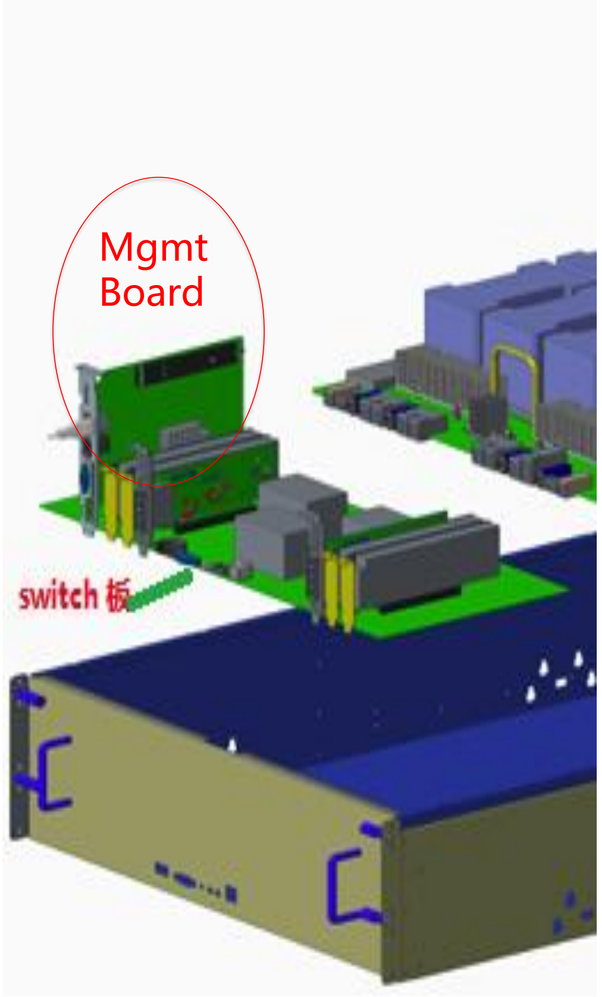




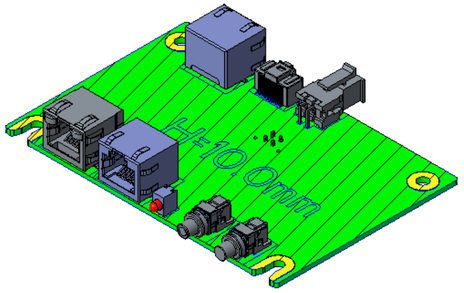
DC-SCM

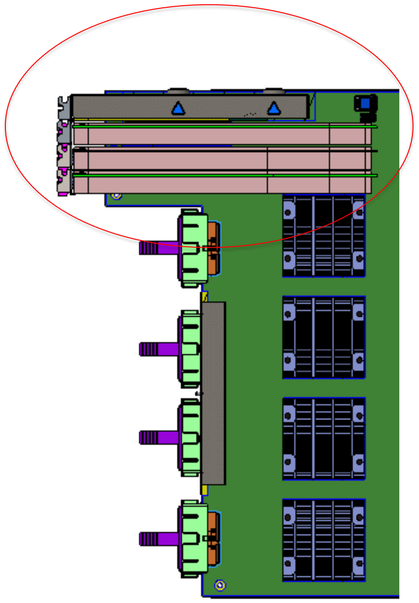


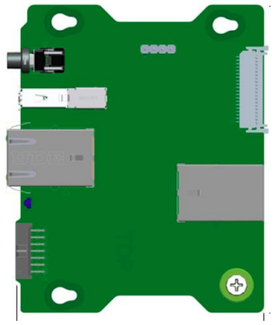
DC-SCI

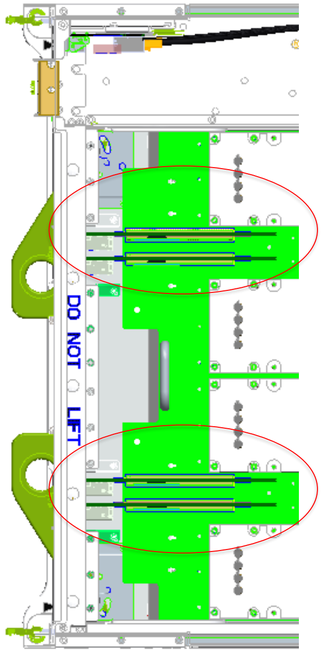


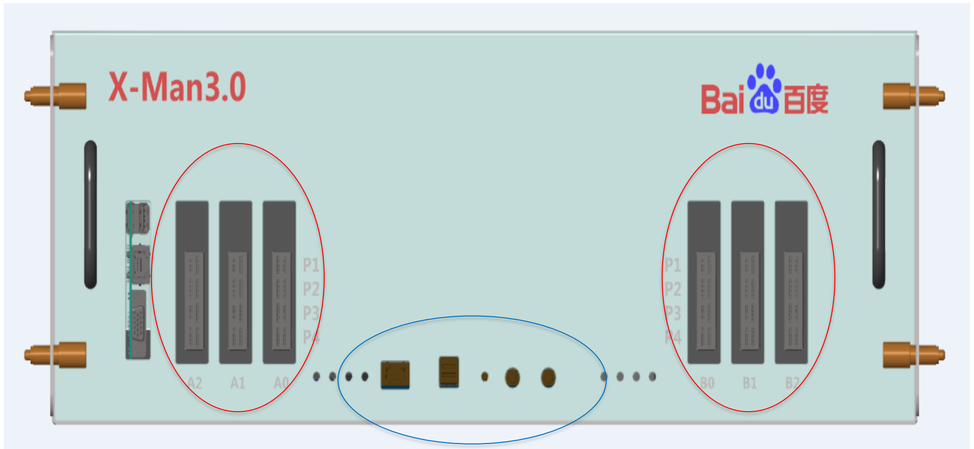
IO Board











Common Requirements

- Power & Cooling
- Flexibility
- Reliability & Serviceability
- Configuration, Programming, and Management
- Inter-module Communication to Scale Up
- Input / Output Bandwidth to Scale Out

“If you want to go *Fast*, go *Alone*;
If you want to go *Far*, go *Together*”

We have done *Fast for Short-term result*;

It is time to go *Far* at OCP for
Long-term gain!

We need an

Open

Accelerator Infrastructure

Increase Interoperability

Accelerate Innovation

Via
Modular Building Block Architecture

We started with an *OCP Accelerator Module*

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

Then, we will add *Infrastructure Support*

SERVER

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

Modular in everyway!

Hierarchical Base Specification

Well-defined boundaries

Fostering Innovation

SERVER

- Power and Cooling
- Mechanical
- Electrical
- Security & Management
- OAM
- UBB (Interconnect Topology)
- Tray
- DC-SCM

Hierarchical Base Specification

Well-defined boundaries

Fostering Innovation

SERVER

- Power and Cooling
- Mechanical
- Electrical
- Security & Management
- OAM
- UBB (Interconnect Topology)
- Tray
- DC-SCM

Designs and Products may be compliant to any or all specifications

The Universal Baseboard (UBB)

Different Neural Networks and Frameworks for

Model or Data Parallelism

Benefit from different

Interconnect Topologies

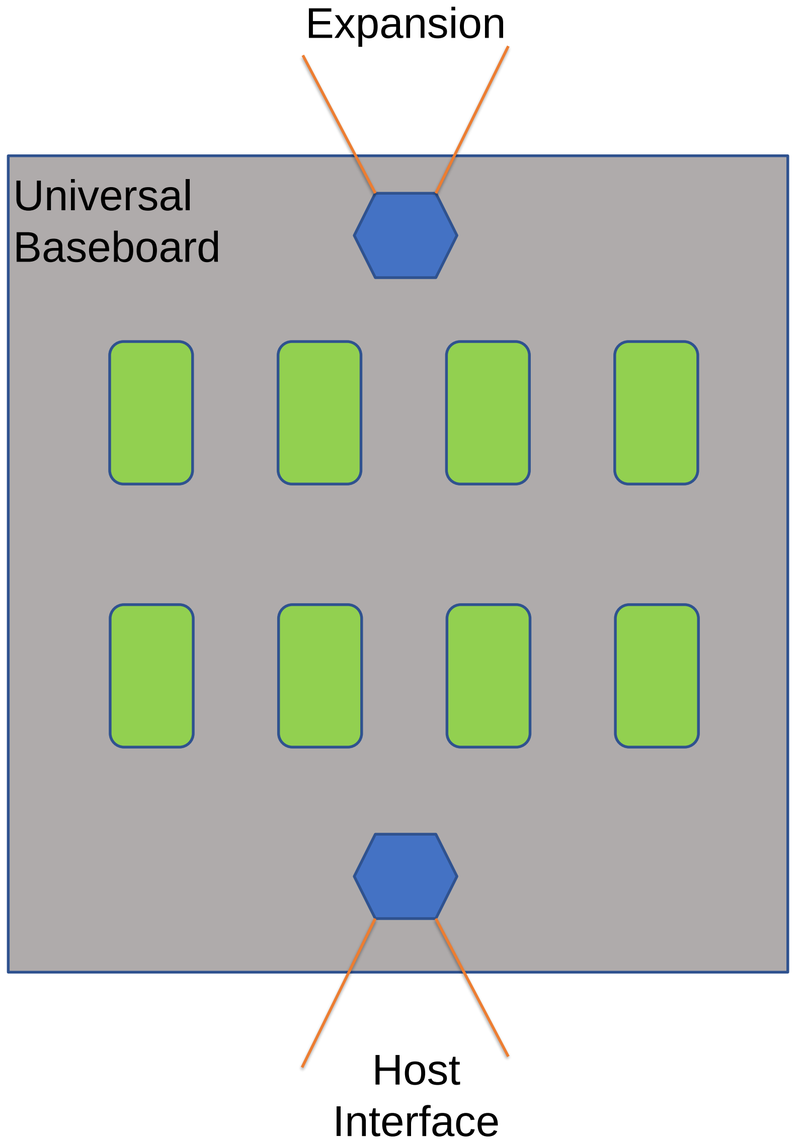
Universal Baseboard (UBB)

Consider a Grid of Planar OAM sites

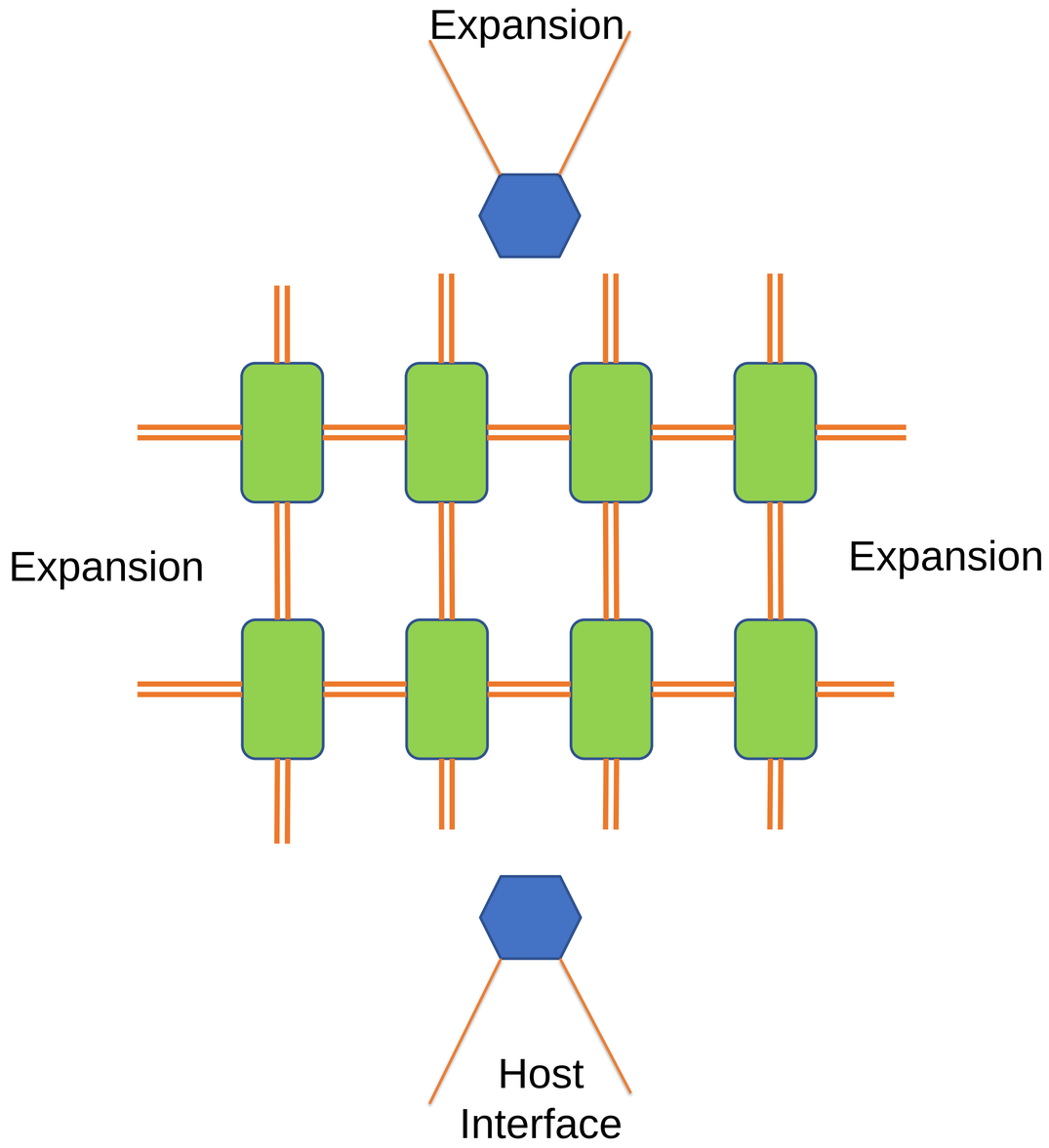
Standard Volumetric

Protocol Agnostic Interconnects

Wires are Wires!



2D Mesh-connected

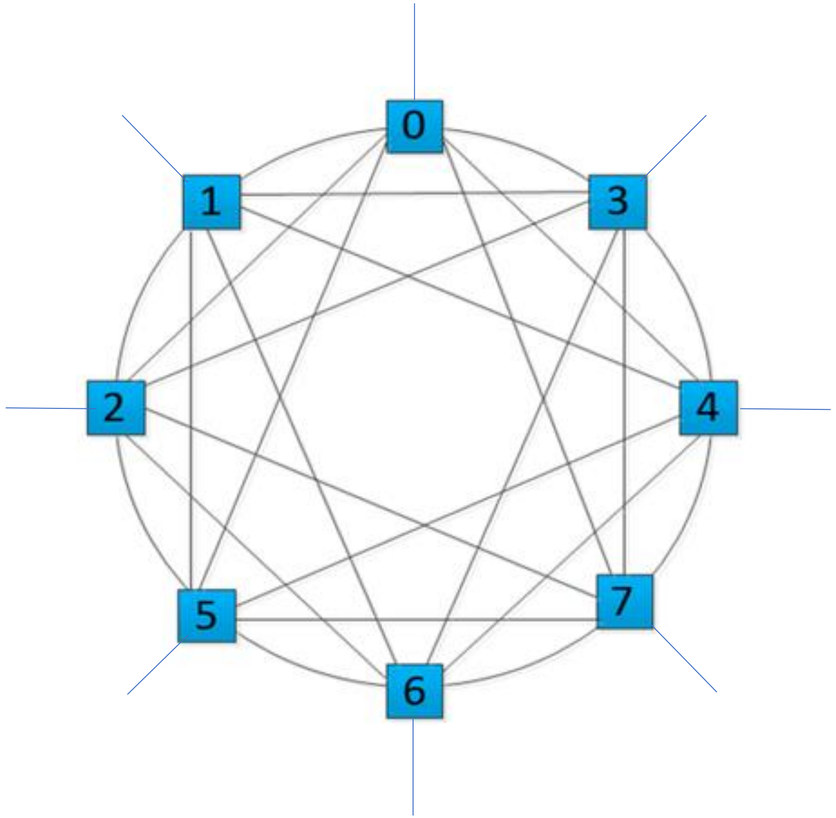


Consider expansion beyond one UBB

Topologies with 6 Links per OAM

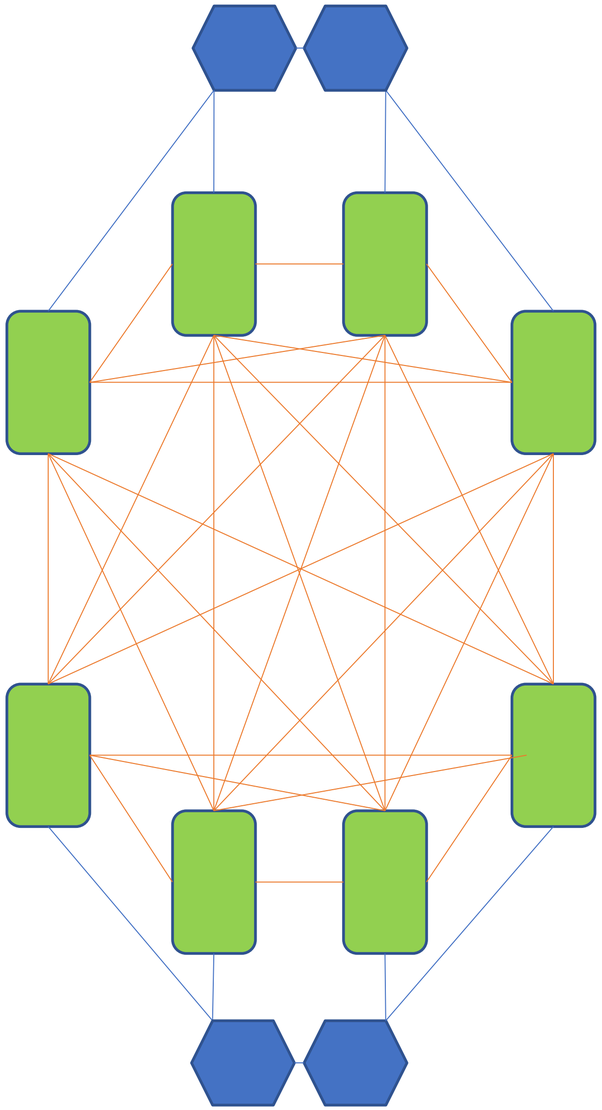
Six inter-module Links may create a 3D Mesh or Torus

One Host Link



Fully-connected OAMs

With seven inter-OAM Links and one Host Link



Fully-connected

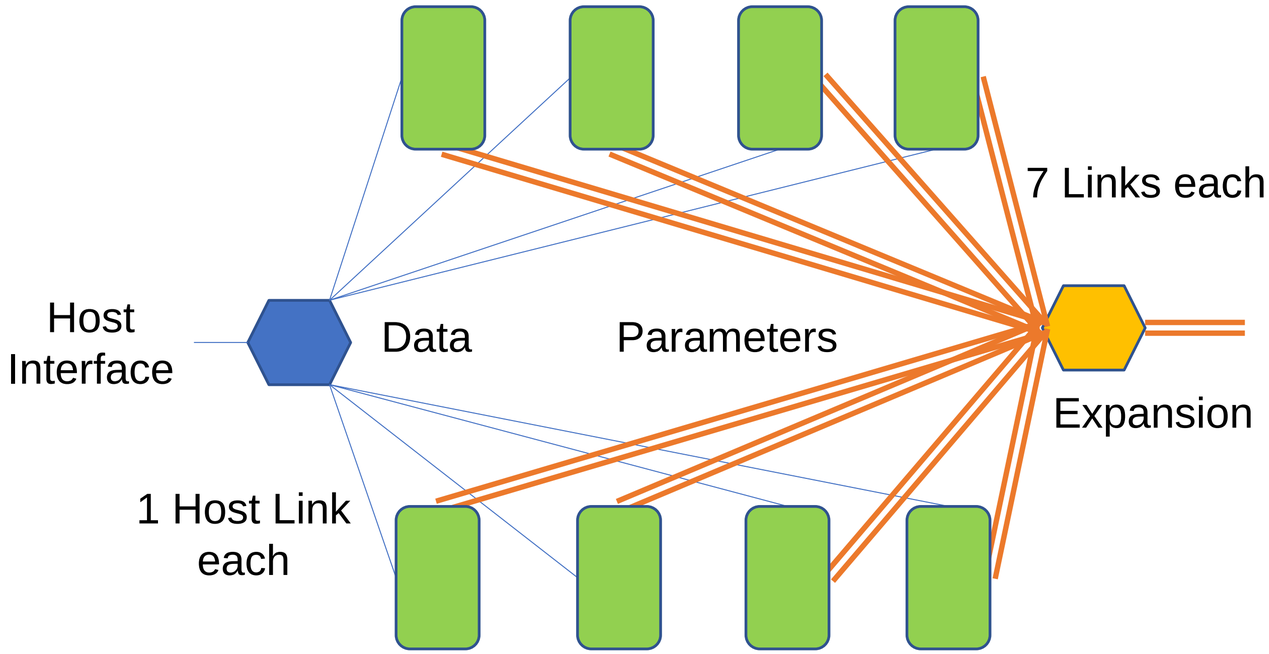
A Grid of interconnected OAMs

Max Bisection BW

One Hop Away

Concurrent, Non-blocking

Ready for Expansion



Fully-connected

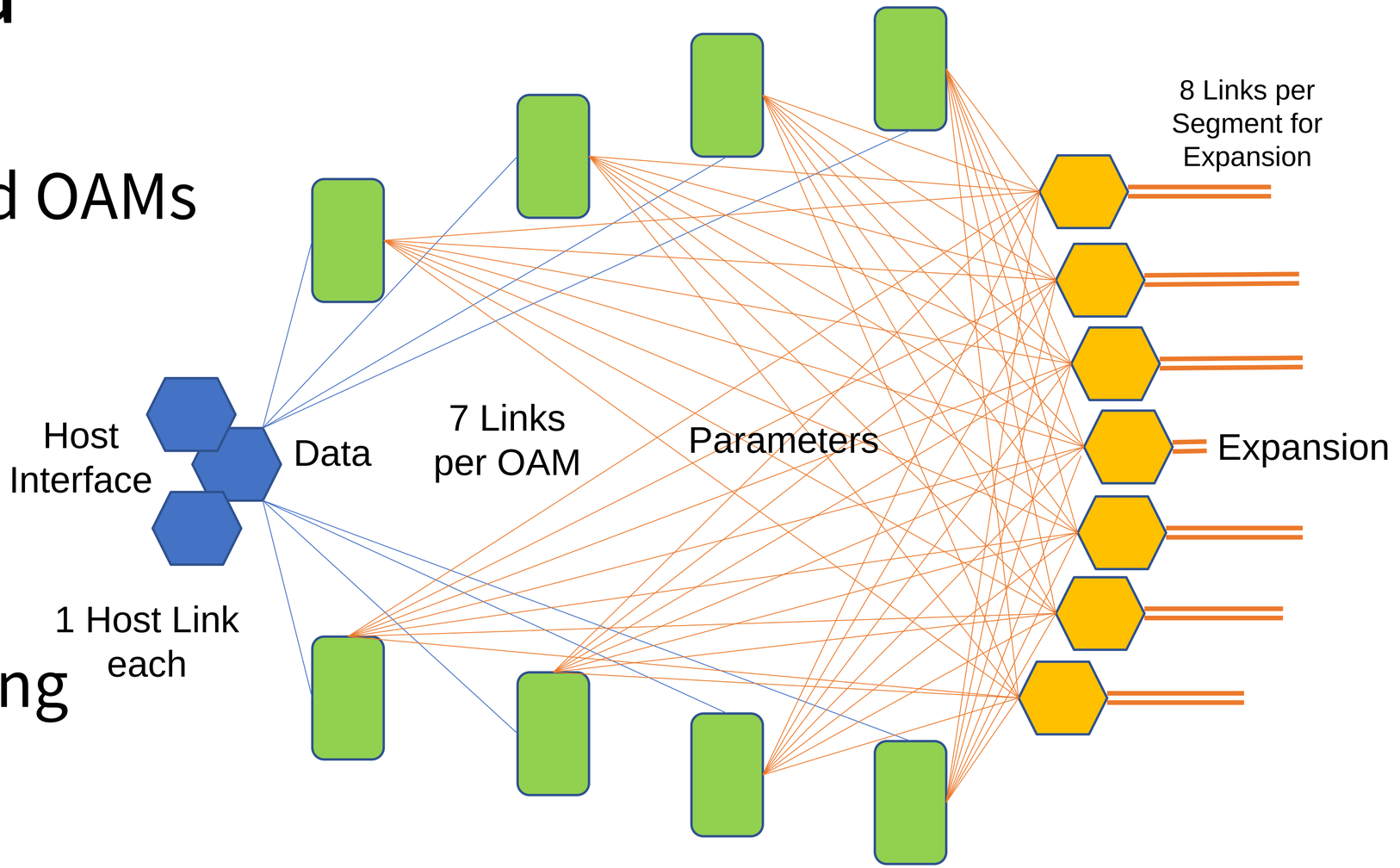
A Grid of interconnected OAMs

Max Bisection BW

One Hop Away

Concurrent, Non-blocking

Ready for Expansion



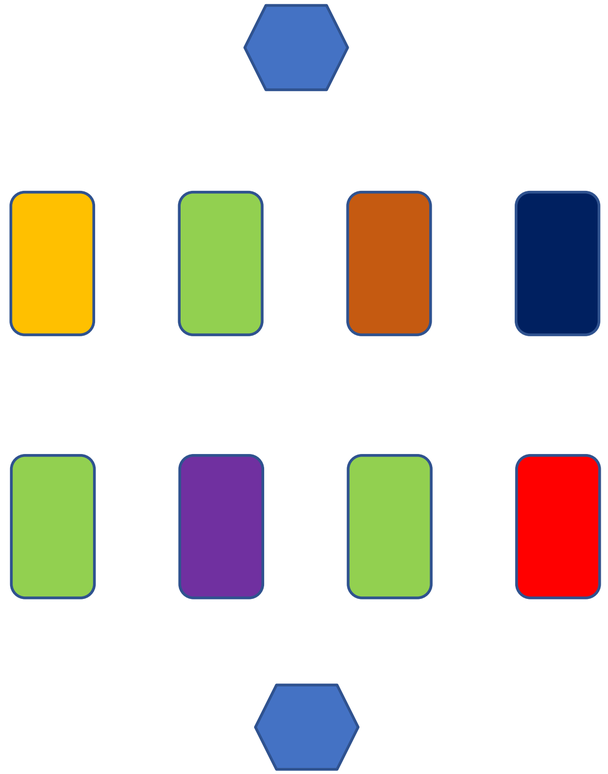
Heterogenous OAMs

These Modules need not be of the same type

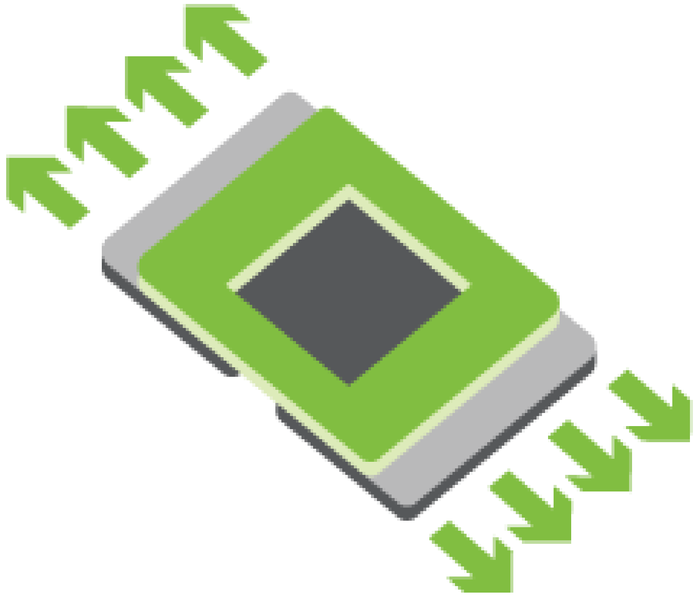
Each one may be suited for a specific application/task

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

Chained, pipelined processing stages



Current Work: OAM Spec

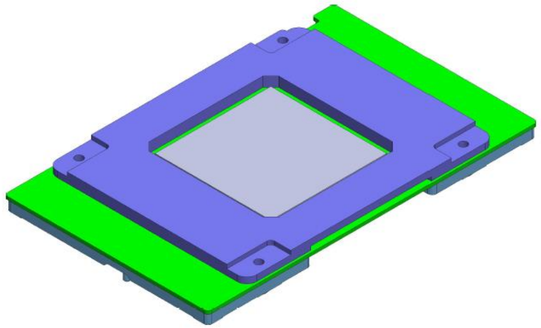


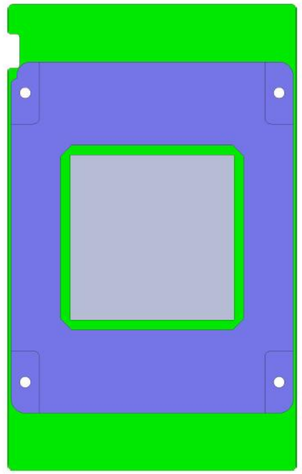
SERVER

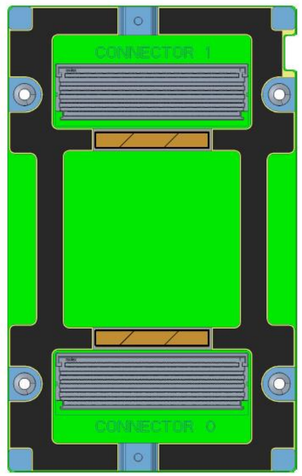
Specifications

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
- Up to 8* Modules per Baseboard
- System management and debug interfaces

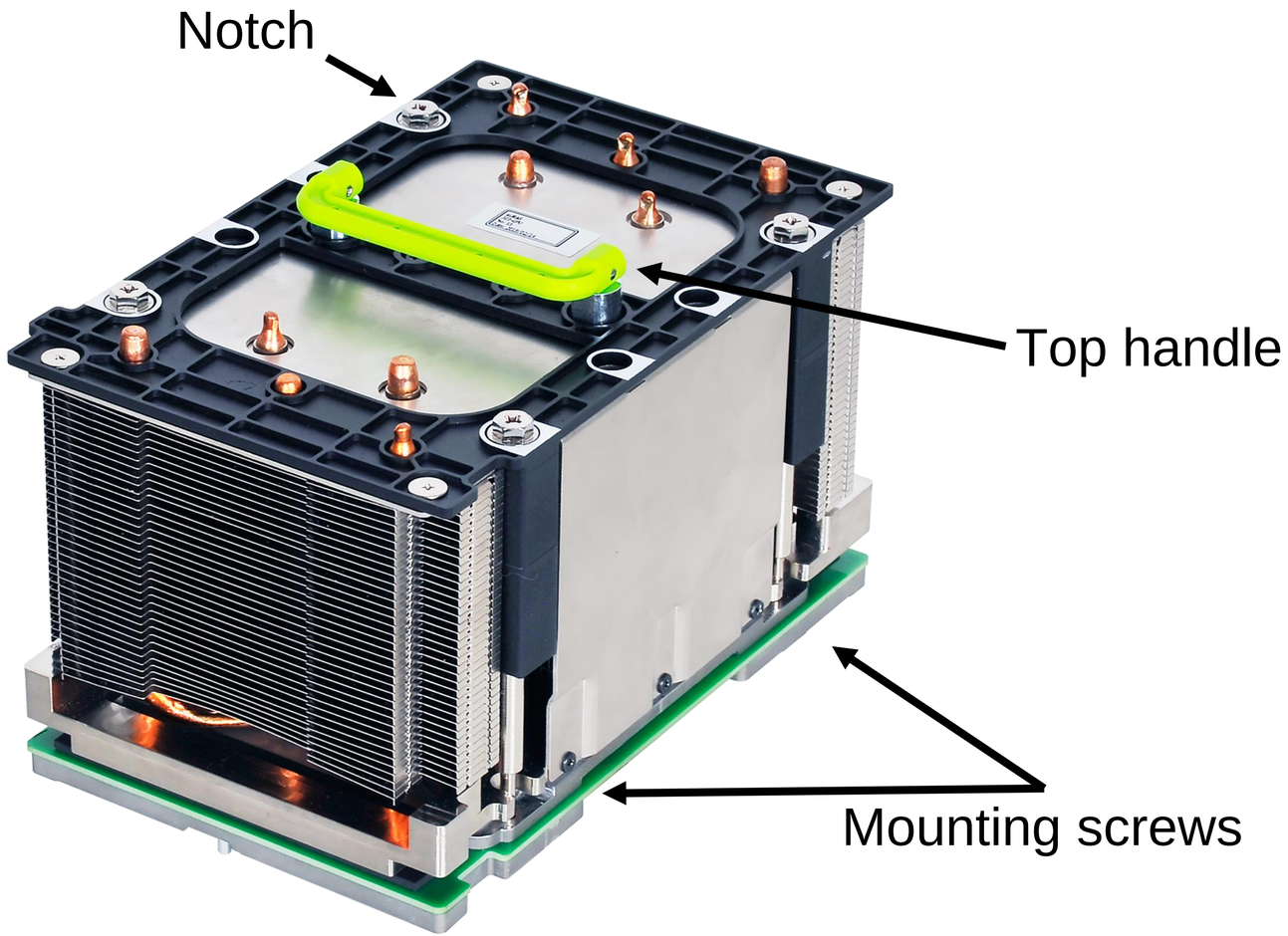






ME Recommendations – HS Reference Design

- 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



OAM Talk at HPC Track

Attend our OAM Talk at the HPC Track for
Mechanical, Electrical, and Thermal details on OAM

https://2019ocpglobalsummit.sched.com/event/M85K/ocp-open-accelerator-module-oam

Summary

- Rev 0.85 of the OAM spec is available
- We are forming a sub-group within Server Project to receive feedback and contributions
- Contributors will sign a License and Legal Agreement

Join the Project and further develop interoperable Modules for an ***Open Accelerator Infrastructure (OAI)***:

- **OAM** as an open accelerator module supporting multiple suppliers
- Universal Baseboard (**UBB**) supporting different interconnect topologies
- **Tray** supporting different UBBs
- System Chassis, Power, and Cooling supporting different Trays
- System- and Rack-level Management (**DC-SCM**) supporting all Trays, UBBs, and OAMs as well as the Hosting Head Nodes

Call to Action

We invite you to join the OAM subgroup for further collaboration:

Register for the Mailing List:
https://ocp-all.groups.io/g/OCP-OAI

Wiki under OCP Server Project:
https://www.opencompute.org/wiki/Server/OAI

Q&A

Presenters

- Siamak Tavallaei is a Principal Architect at Microsoft Azure and co-chair of OCP Server Project. 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 hardware, software, and management.
- Whitney Zhao is a seasoned hardware engineer leading AI/ML system design in Facebook. Whitney has led multiple hardware generations ranging from generous purpose 2S system such as Tioga Pass to ML JBOG Big Basin systems, all of which have been contributed to OCP. She has been driving multiple hardware-software co-design initiatives across both training and inference areas, She is leading the hardware system design for Facebook's main AI workloads. She is also instrumental in bringing industry partners together to solve common infrastructure problem of bringing efficient @scale AI/ML solution for everyone to benefit from.
- Richard Ding is AI System Architect for heterogeneous computing in Technical Group of Baidu. He leads architecture design of Baidu's AI computing platform X-MAN, the high-performance parallel file system FAST-F, and the large-scale training cluster KongMing. His research focuses on large-scale and distributed training system design and optimization, high-performance storage, and high-speed interconnect technologies, as well as hardware-software co-optimization for AI chips.
- Tiffany Jin is a mechanical engineer for datacenter hardware design at Facebook. She leads the mechanical design of multiple programs across hardware infrastructure, mainly compute platforms including AI/ML and 2S systems such as Tioga Pass. Tiffany holds a BS and MS in Mechanical Engineering from MIT and Stanford, respectively.


Open. Together.
OCP Global Summit | March 14–15, 2019
OCP
SUMMIT