OCP CHINA DAY | June 25th 2019 Beijing

Open Accelerator Infrastructure(OAI) Project

Whitney Zhao, Technical Lead, Facebook

Richard Ding, AI System Architect, Baidu

Song Kok Hang, Principal Engineer, Intel

Siamak Tavallaei, Principal Architect, Microsoft

Outline

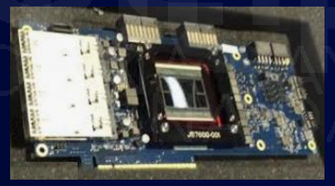
- **Motivation**
- **OAI(Open Accelerator Infrastructure)**
- **OAM(OCP Accelerator Module)**
- **UBB(Universal Baseboard)**
- **Examples**
- **Requesting Participation and Feedback**

OCP CHINA DAY June 25th 2019 Beijing

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)

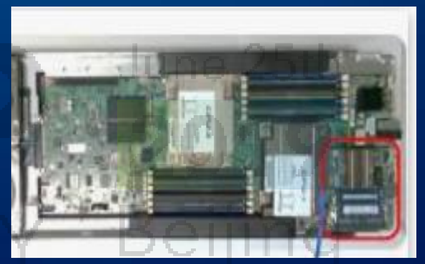
GPU | FPGA | ASIC | NPU | TPU | NNP | IPU | xPU…

Diverse Module and System Form Factors

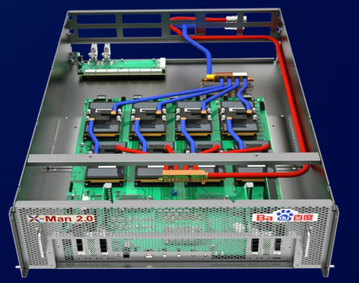
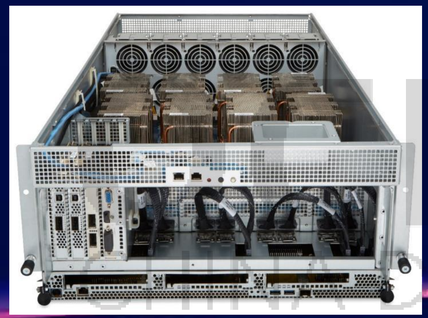








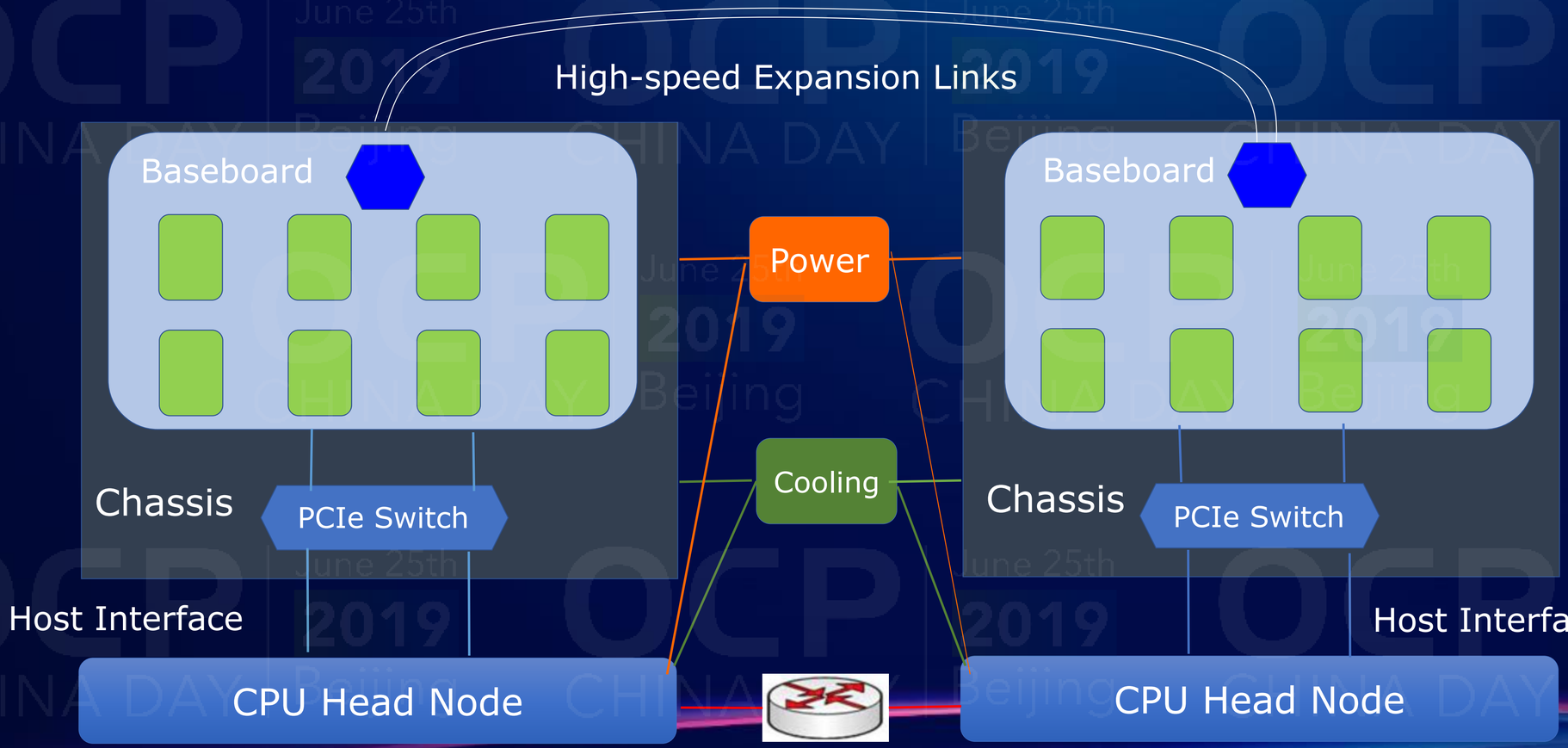



OCP CHINA DAY June 25th 2019 Beijing

Different Implementations

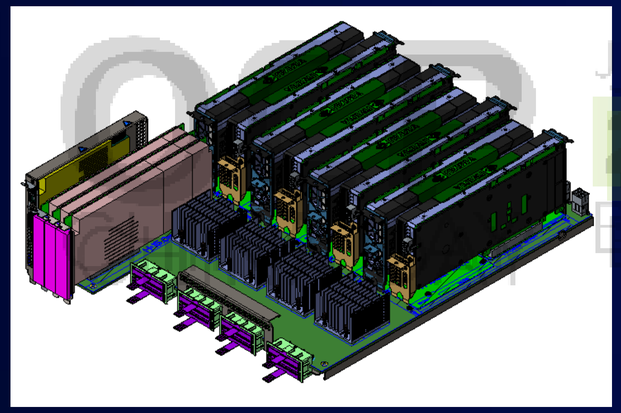
Targeting Similar Requirements!

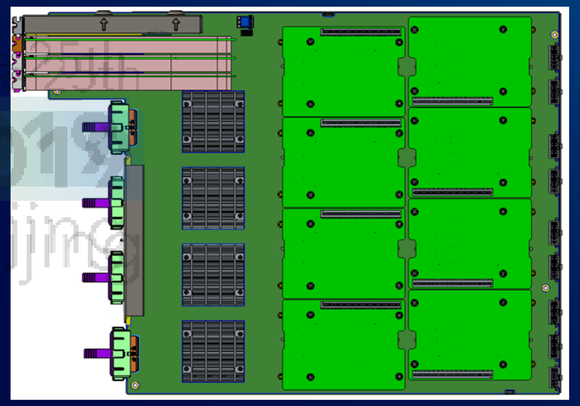
Logical Components for AI Hardware System

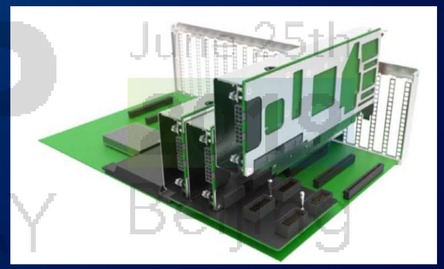


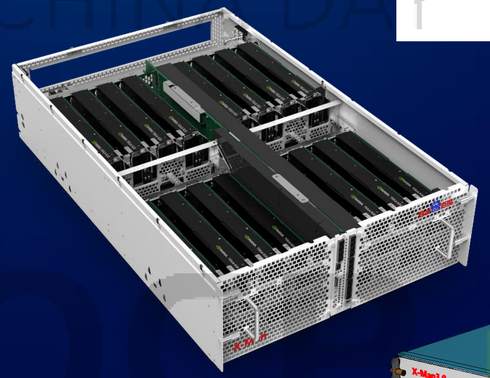
OCP CHINA DAY June 25th 2019 Beijing

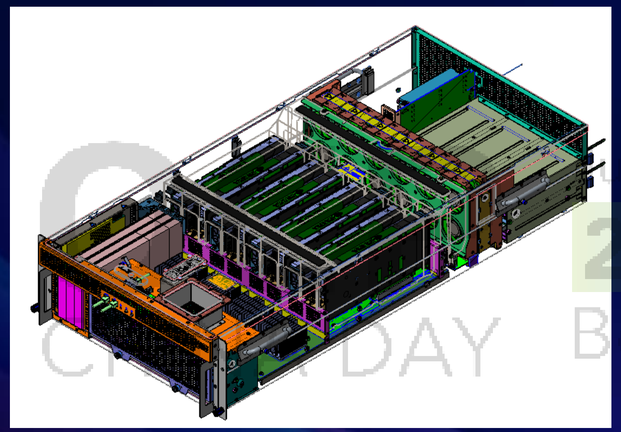
Accelerators in PCIe CEM Form Factor

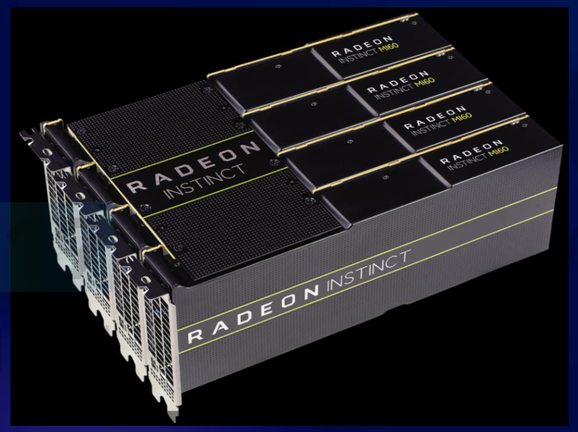






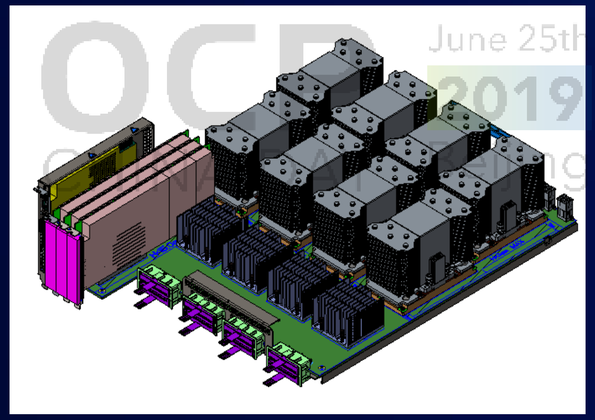


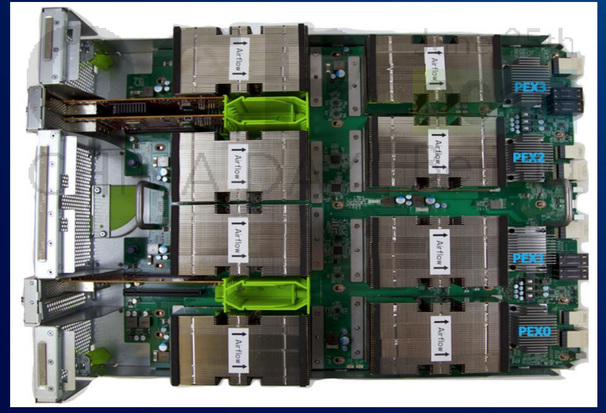


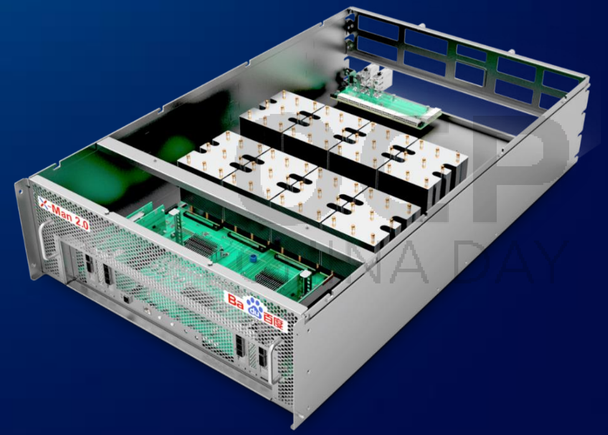


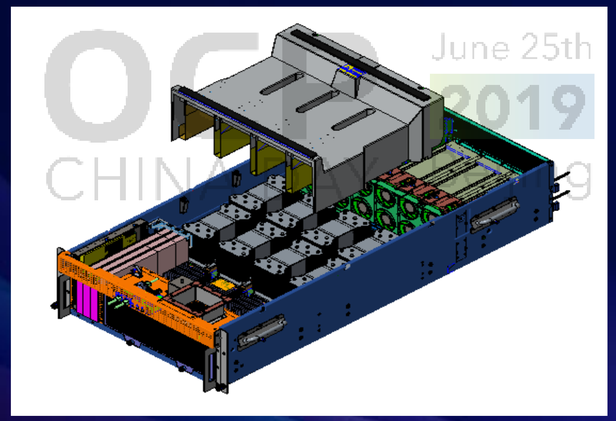
OCP CHINA DAY June 25th 2019 Beijing

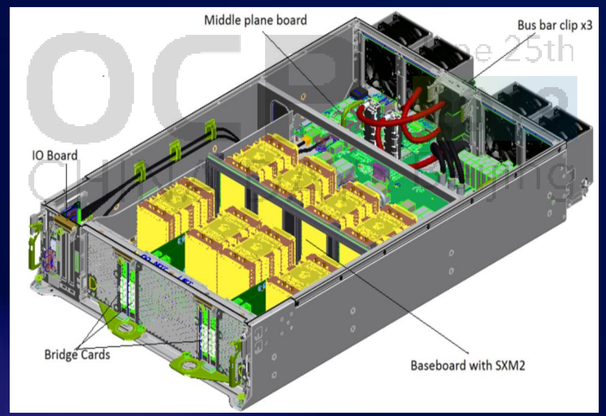
Accelerators in Mezzanine Form Factor on Baseboard

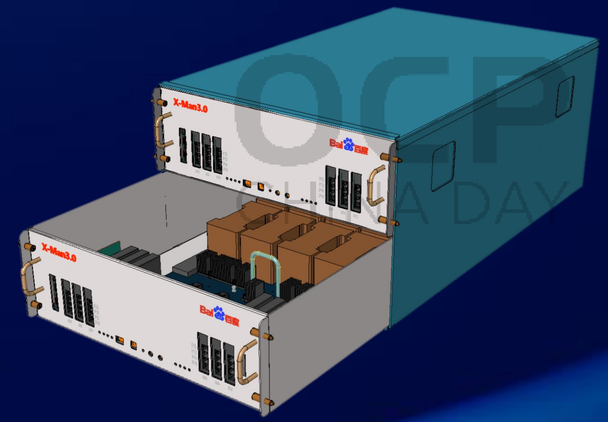




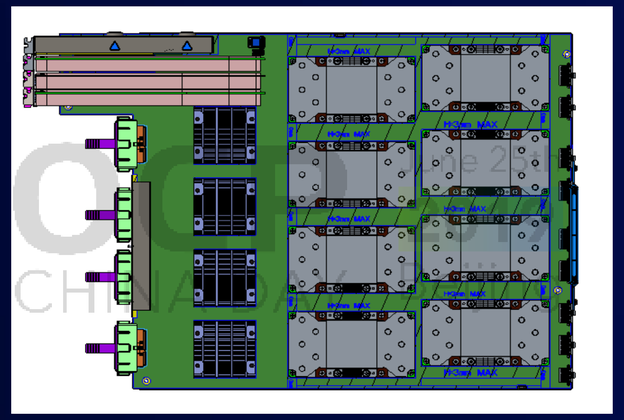


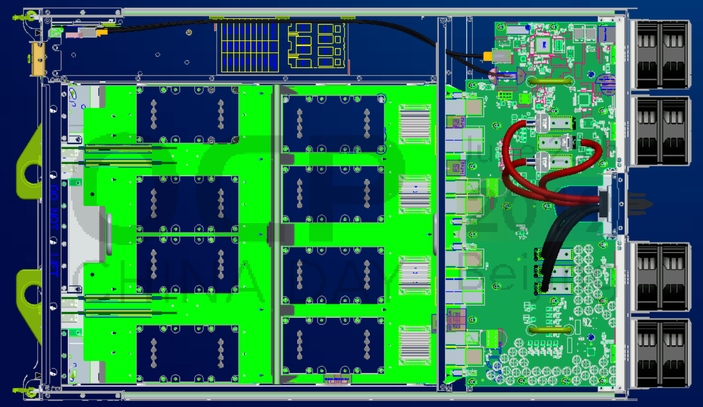


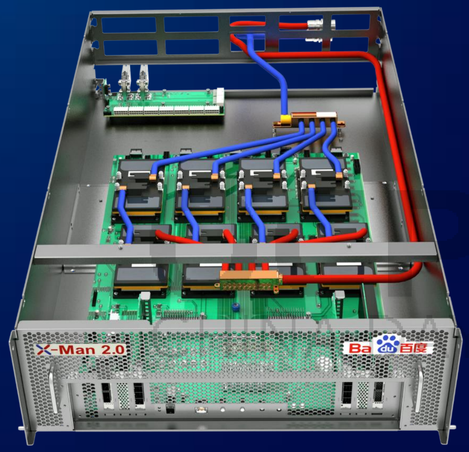


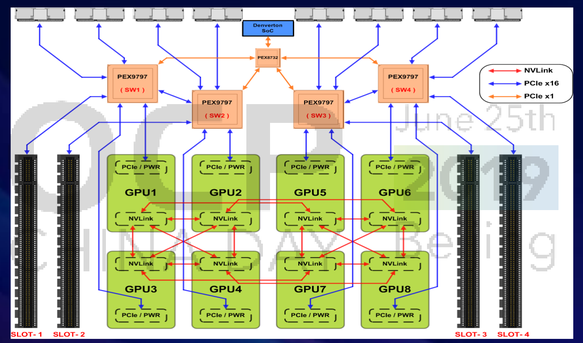


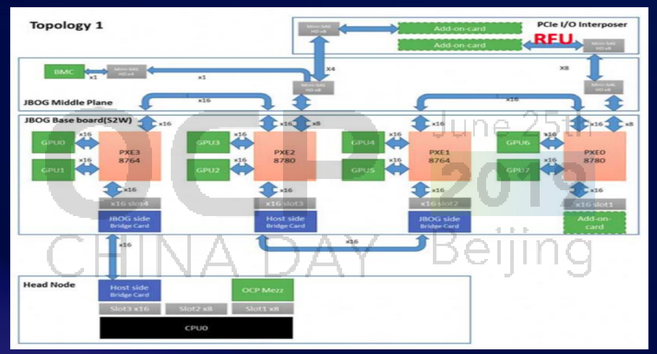
Various PCIe Switch Topologies

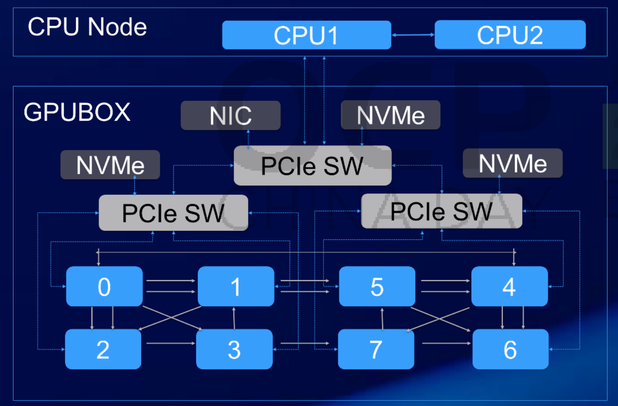




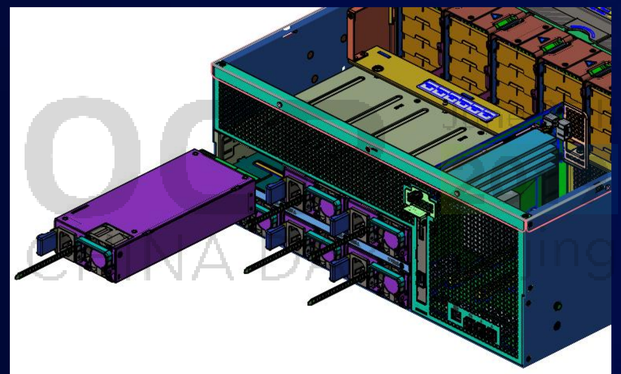




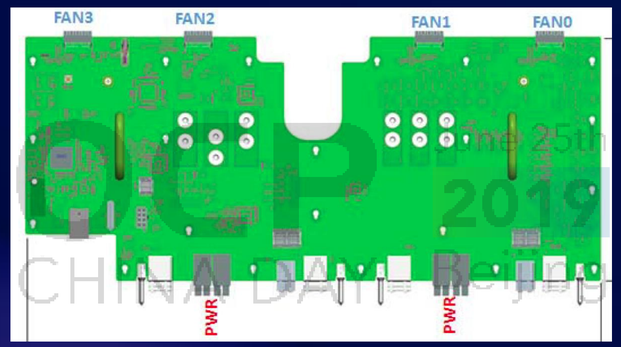




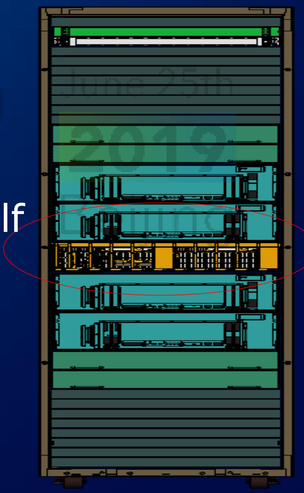
Power Delivery and Distribution



To Busbar

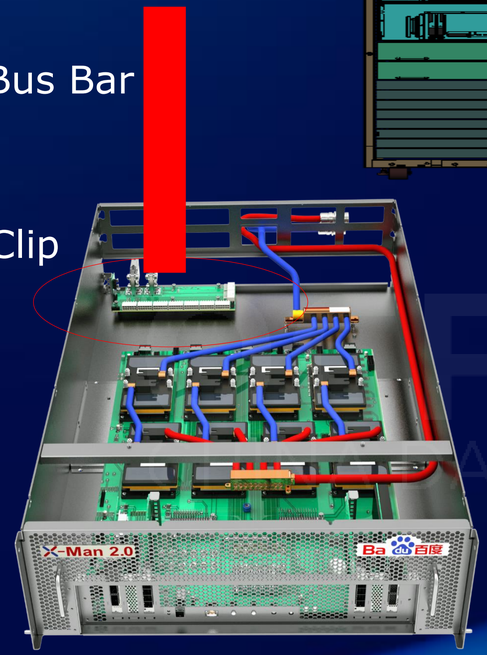


Shared Power Shelf



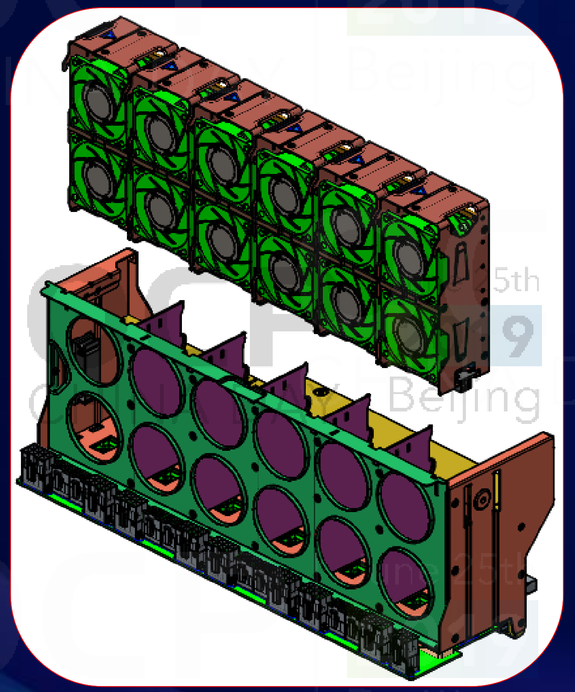
Bus Bar

Clip

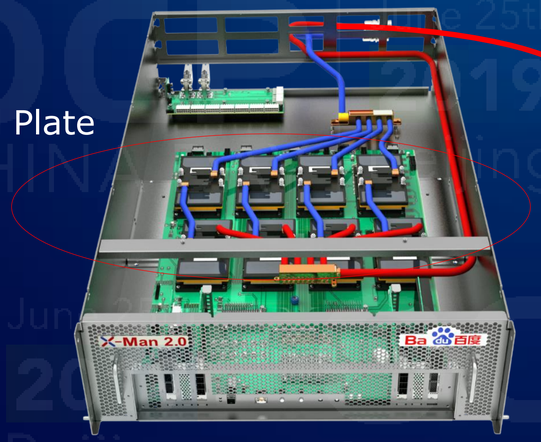


OCP CHINA DAY June 25th 2019 Beijing

Cooling Methods

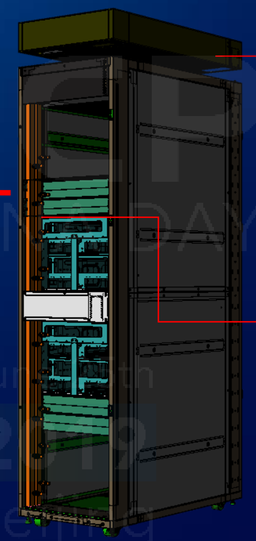


Cold Plate



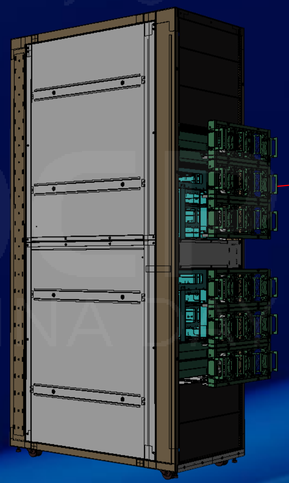
\+

Liquid Cooling

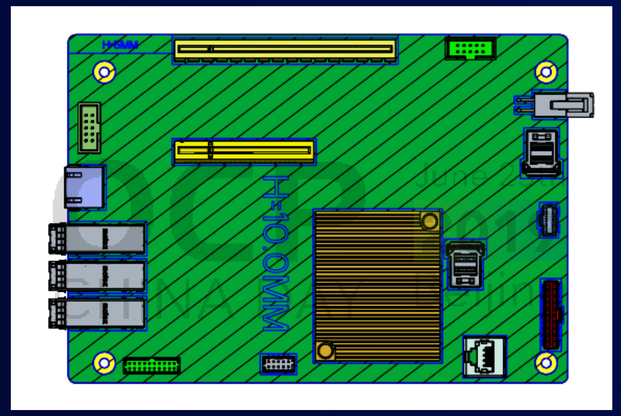


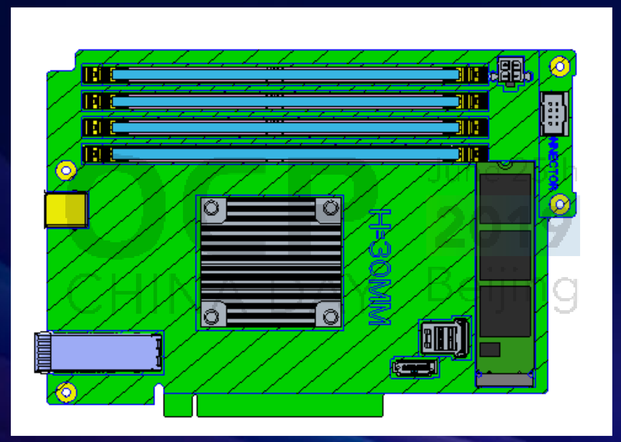
\+

Air Cooling

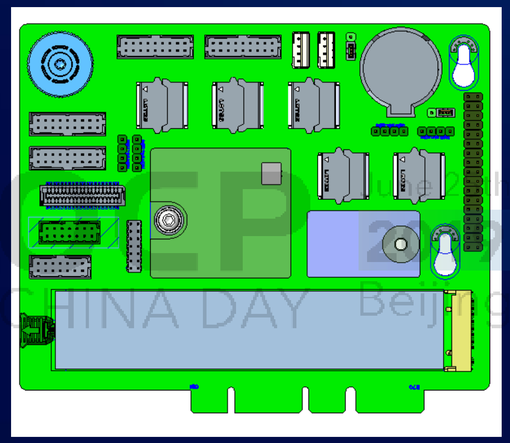


Management Module

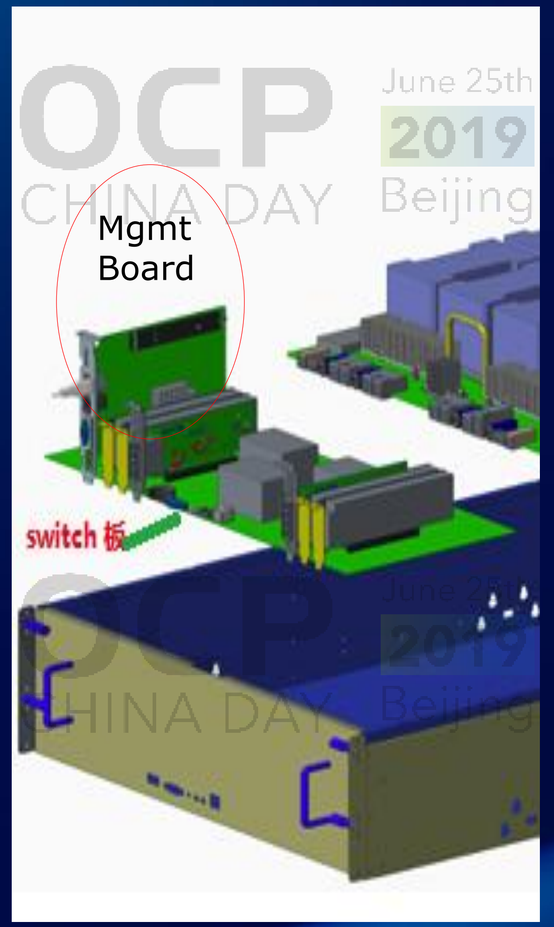




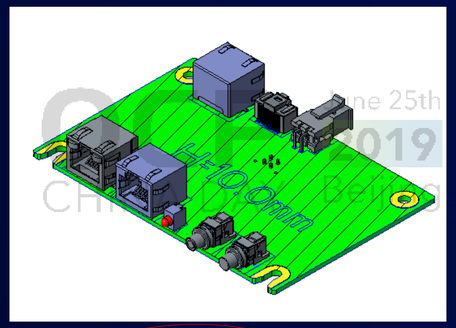
DC-SCM

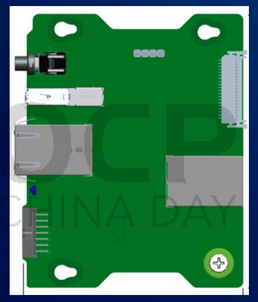


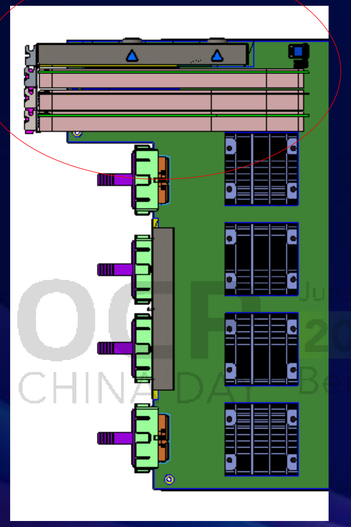
DC-SCI

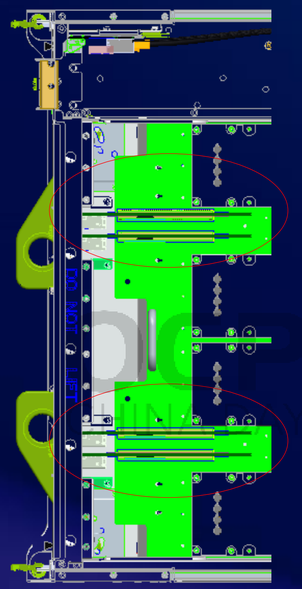


IO Board









Common Requirements

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

OCP CHINA DAY | June 25th 2019 Beijing

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

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

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

OCP CHINA DAY June 25th 2019 Beijing

We need an

Open

Accelerator Infrastructure

OCP CHINA DAY June 25th 2019 Beijing

Increase Interoperability

Accelerate Innovation

Via

Modular Building Block Architecture

Modular in everyway!

Open Accelerator Infrastructure(OAI) Project

Hierarchical Base Specification

Well-defined boundaries
Fostering Innovation

- OCP Accelerator Module (OAI-OAM)
- Universal Baseboard (OAI-UBB)
- Host Interface (OAI-HIB)
- Power Distribution (OAI-PDB)
- Expansion Beyond UBB (OAI-Expansion)
- Security, Control, and Management (OAI-SCM)
- Tray
- Chassis

Open Accelerator Infrastructure(OAI) Project

Hierarchical Base Specification

Well-defined boundaries
Fostering Innovation

- OCP Accelerator Module (OAI-OAM)
- Universal Baseboard (OAI-UBB)
- Host Interface (OAI-HIB)
- Power Distribution (OAI-PDB)
- Expansion Beyond UBB (OAI-Expansion)
- Security, Control, and Management (OAI-SCM)
- Tray
- Chassis

Designs and Products may be compliant to any or all specifications

Open Accelerator Infrastructure(OAI) Project

Hierarchical Base Specification

Well-defined boundaries
Fostering Innovation

- OCP Accelerator Module (OAI-OAM)
- Universal Baseboard (OAI-UBB)
- Host Interface (OAI-HIB)
- Power Distribution (OAI-PDB)
- Expansion Beyond UBB (OAI-Expansion)
- Security, Control, and Management (OAI-SCM)
- Tray
- Chassis

Designs and Products may be compliant to any or all specifications

OCP CHINA DAY June 25th 2019 Beijing

OCP Accelerator Module (OAM)

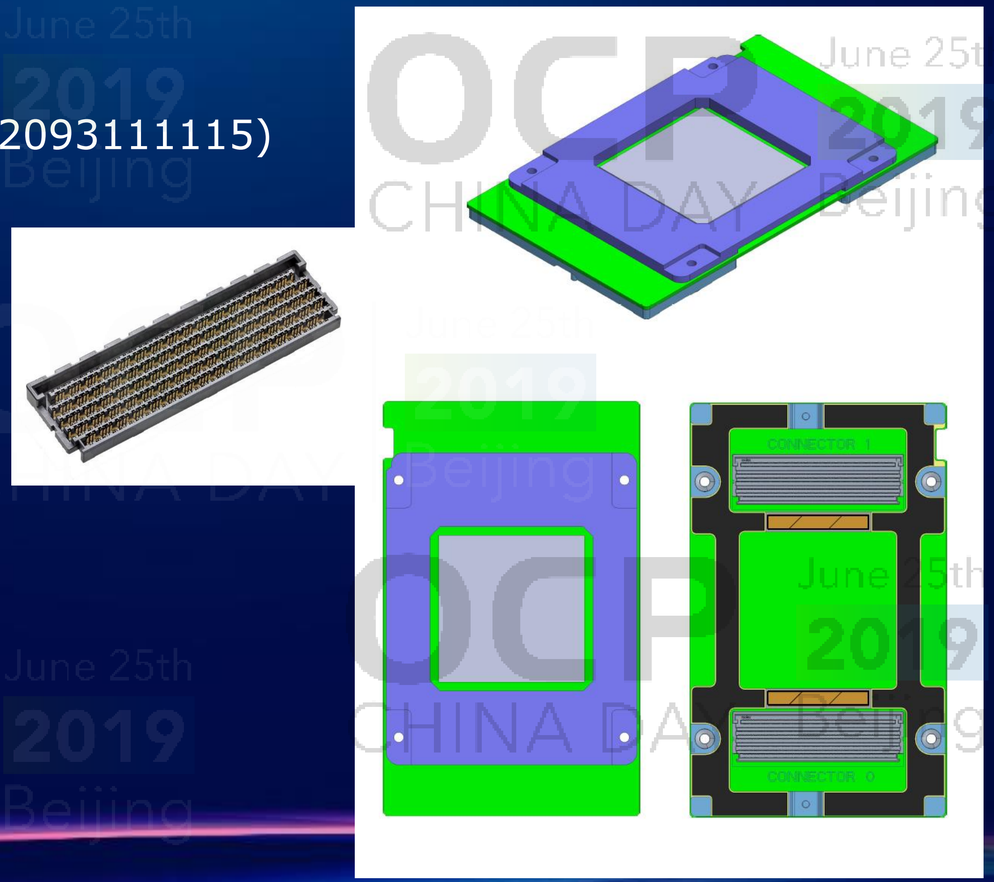
Why OAM

Go beyond what's possible with PCIe CEM form factor

- High-density connectors to increase # of input/output Links
- Low signal insertion loss → high-speed interconnect
- Enough space for Accelerators and local logic & power
- Flexible for heatsink design for air- and liquid-cooling
- Flexible inter-Module interconnect topologies
- Maximize WL performance -> support higher power >300W

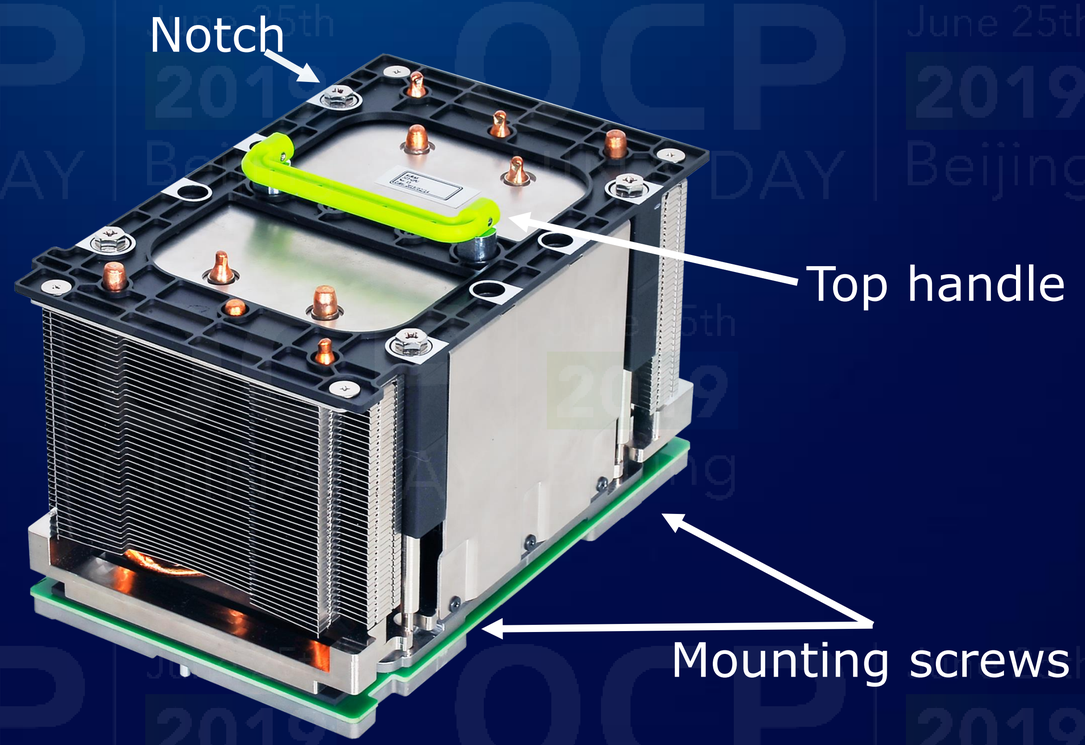
Heterogenous OAMs

- 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
- SerDes speed up to 56G PAM4
- 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



Facebook, Baidu and MSFT Contributed OAM Spec at OCP 2019

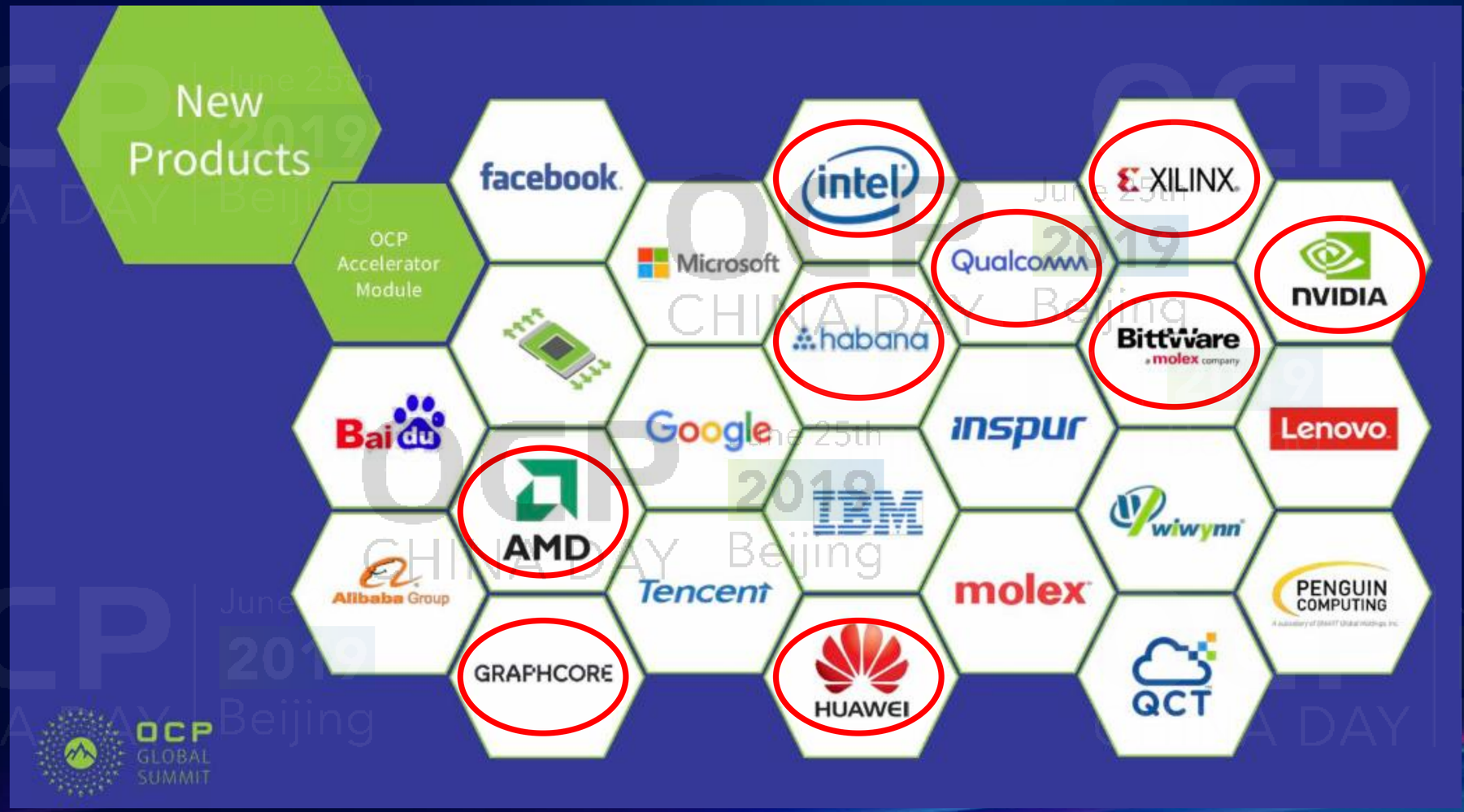
OCP CHINA DAY June 25th 2019 Beijing

OAM Supporter List

OCP GLOBAL SUMMIT

OCP CHINA DAY June 25th 2019 Beijing

OAM Supporter List



OCP GLOBAL SUMMIT

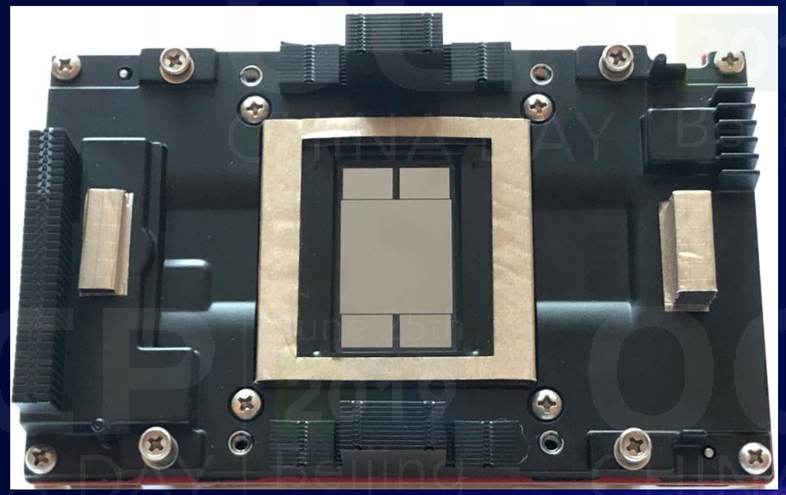
OCP CHINA DAY June 25th 2019 Beijing

OAM Supporter List

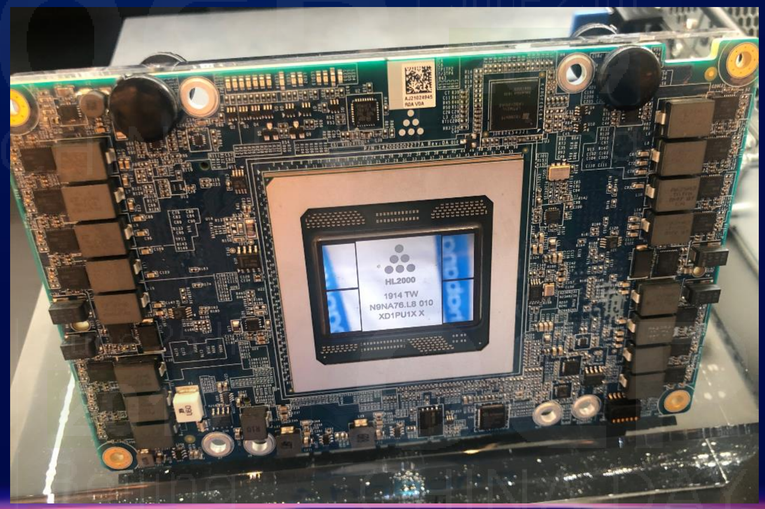
OCP GLOBAL SUMMIT

OAM Status

- Spec v0.9 already released
- Spec v1.0 target to be released by end of June.
- We're working with multiple OAM suppliers to enable OAM based accelerators, GPUs etc.



Intel® Nervana™ NNP-T OAM



Habana Gaudi OAM

OCP CHINA DAY June 25th 2019 Beijing

The Universal Baseboard (UBB)

OCP CHINA DAY June 25th 2019 Beijing

Different Neural Networks and Frameworks for

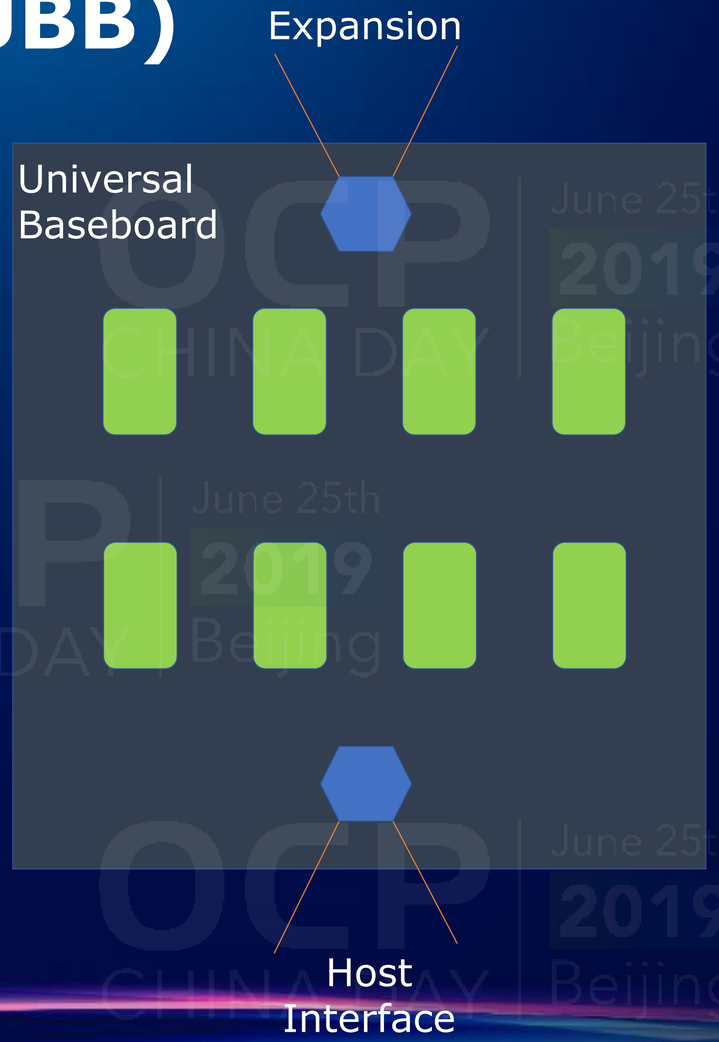
Model or Data Parallelism

Benefit from different

Interconnect Topologies

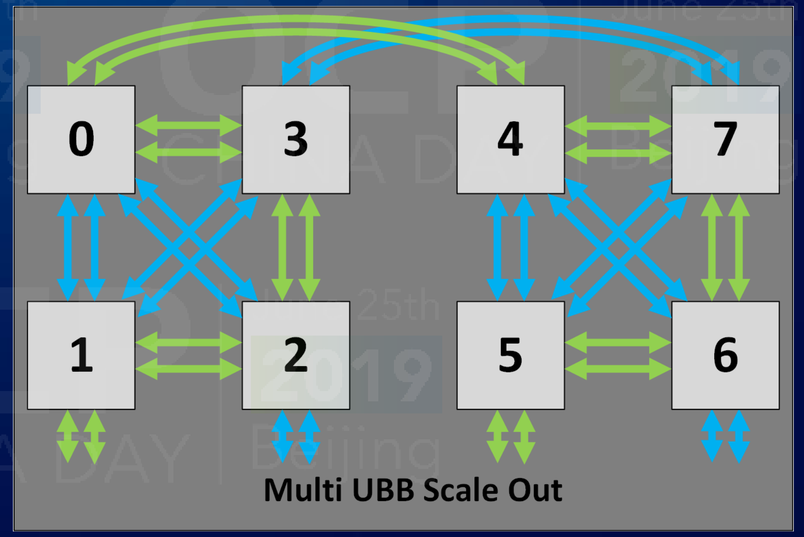
Universal Baseboard (UBB)

- Consider a Grid of Planar OAM sites
- Standard Volumetric
- Protocol Agnostic Interconnects
- Common Components and Connectors
- Wires are Wires!



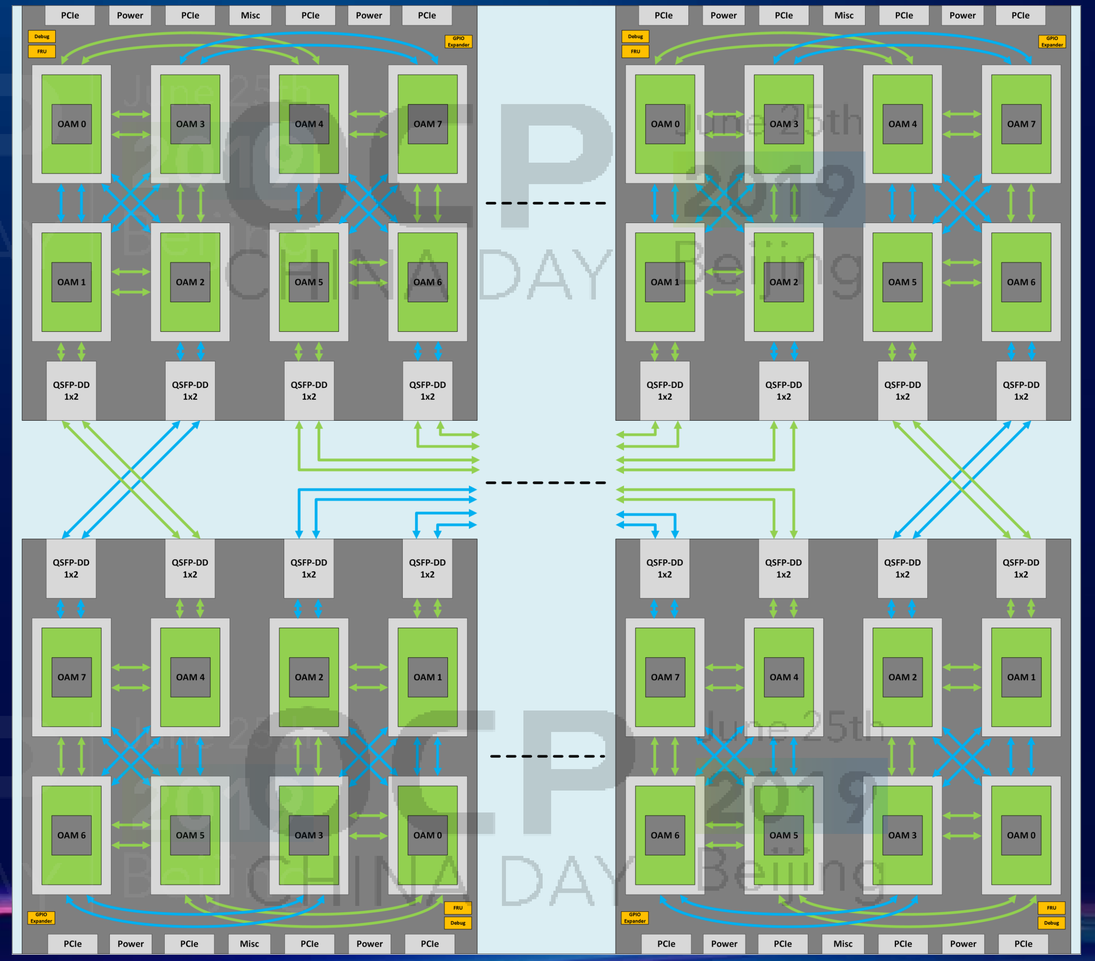
Hybrid Cube Mesh (HCM)

- Eight (x8) Inter-OAM Links
- One Host Link
- Multi UBB Scale Out
- Optimized for Bandwidth
- Optimized for Large Training Clusters



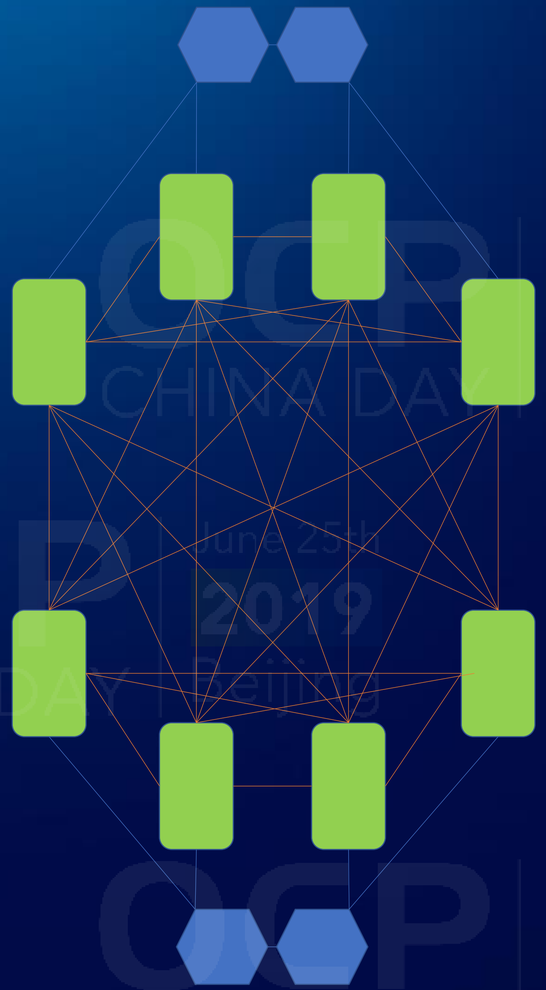
Multi HCM UBB Scale Out

- 16/32/64/128...OAMs
- 2/4/8/16...UBBs



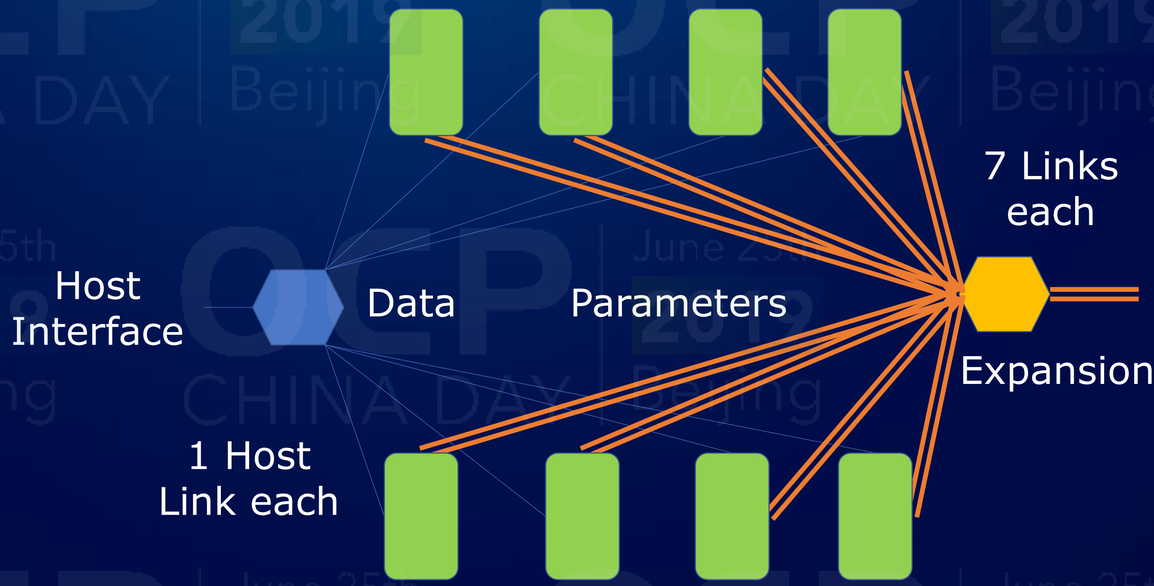
Fully-connected OAMs

With ***seven*** inter-OAM Links and one Host Link



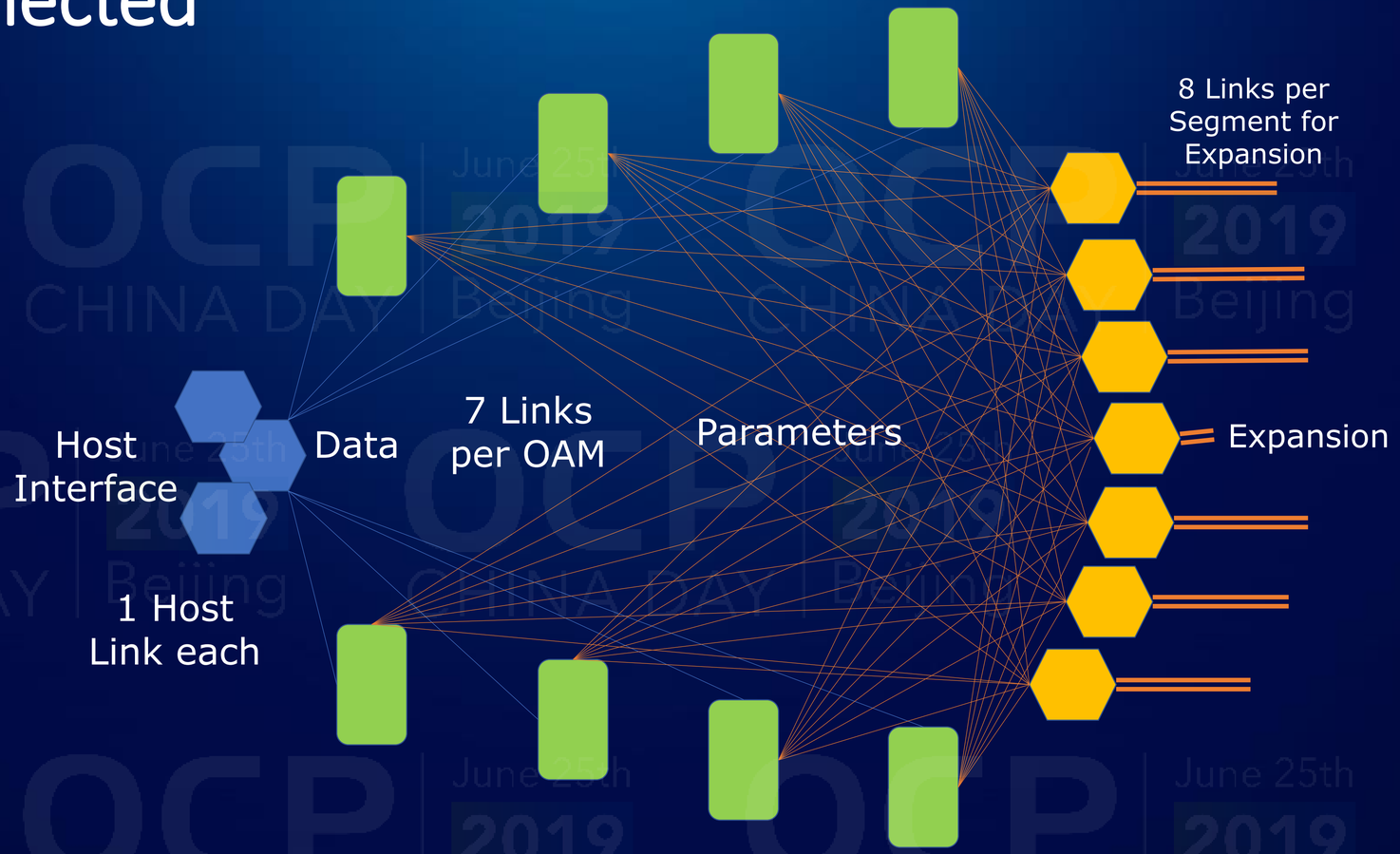
Fully-connected

- A Grid of interconnected OAMs
- Max Bisection BW
- One Hop Away
- Concurrent, Non-blocking
- Ready for Expansion



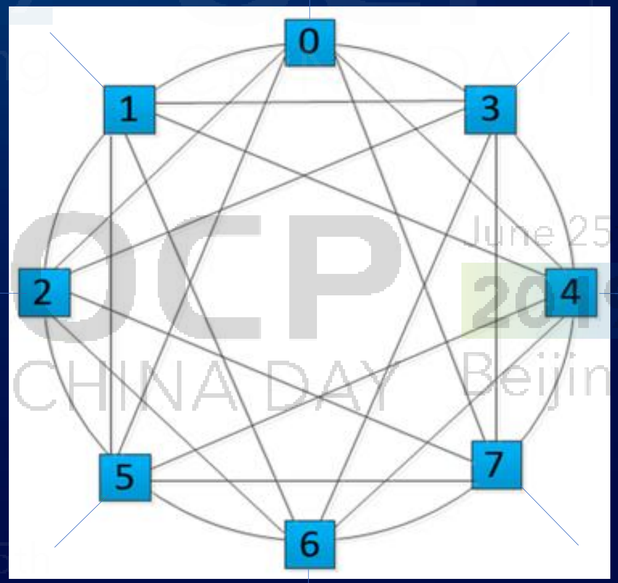
Fully-connected

- A Grid of interconnected OAMs
- Max Bisection BW
- One Hop Away
- Concurrent, Non-blocking
- Ready for Expansion



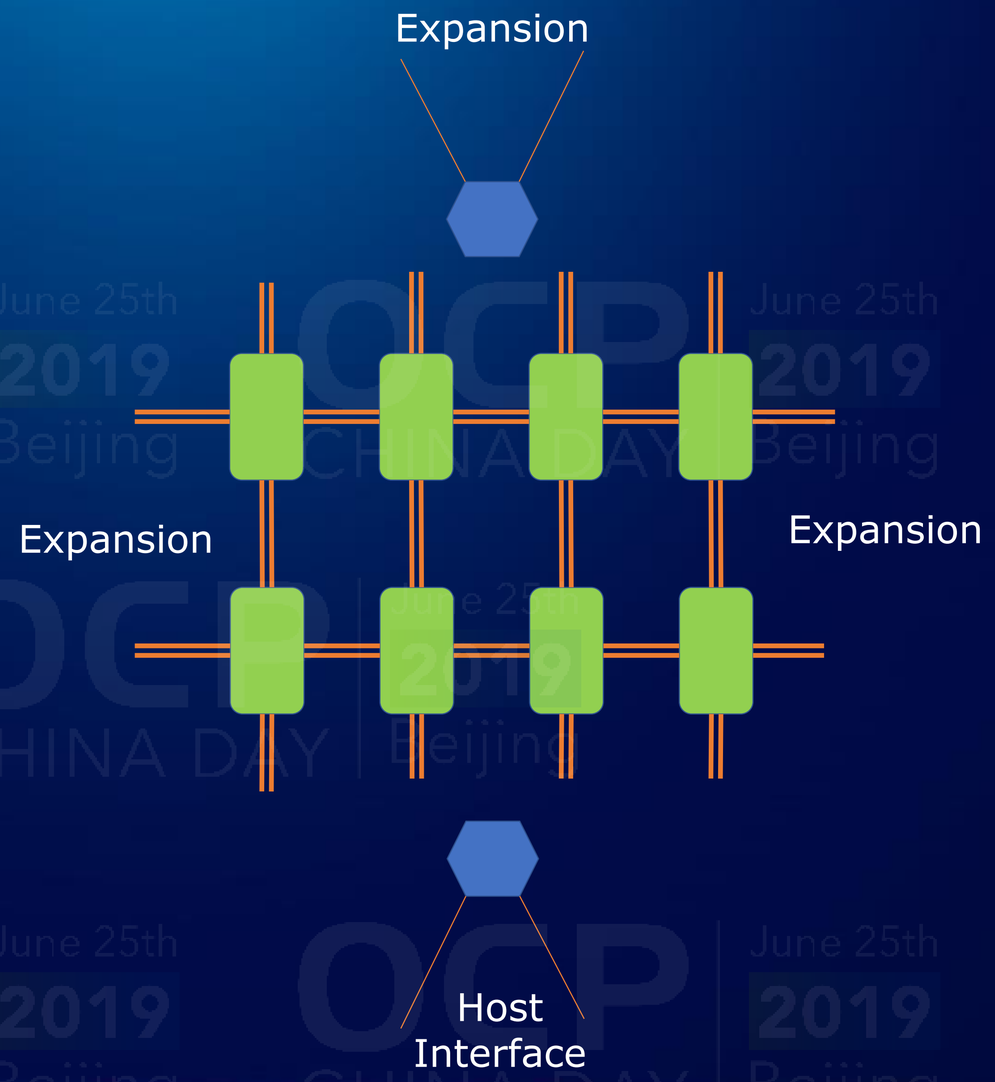
Topologies with 6 Links per OAM

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



2D Mesh-connected

- Consider expansion beyond one UBB



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, …

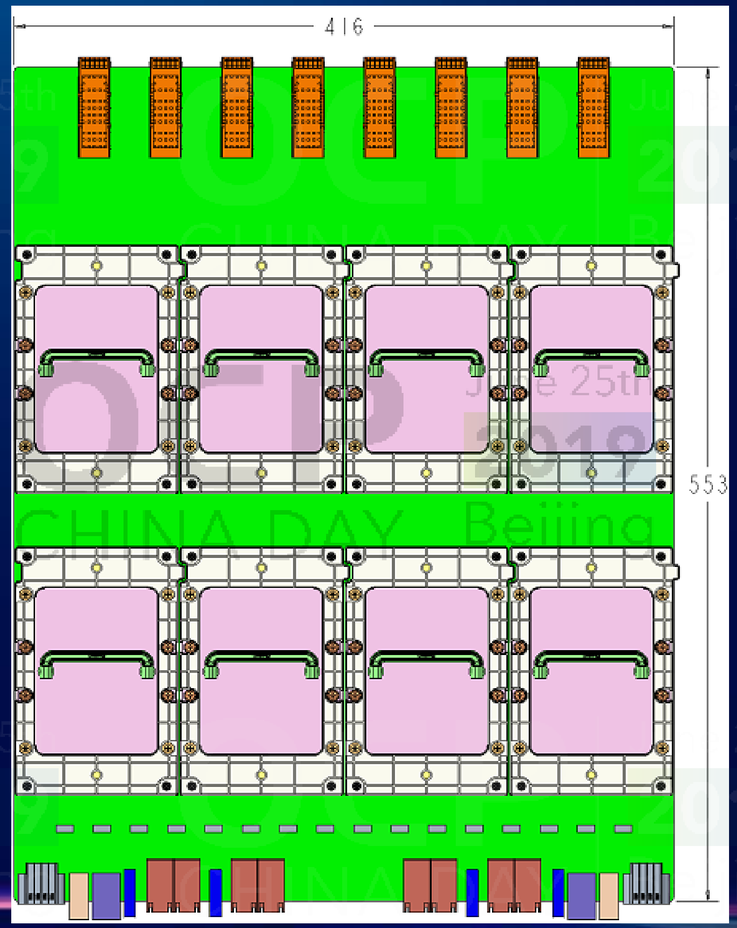
Chained, pipelined processing stages

UBB Status

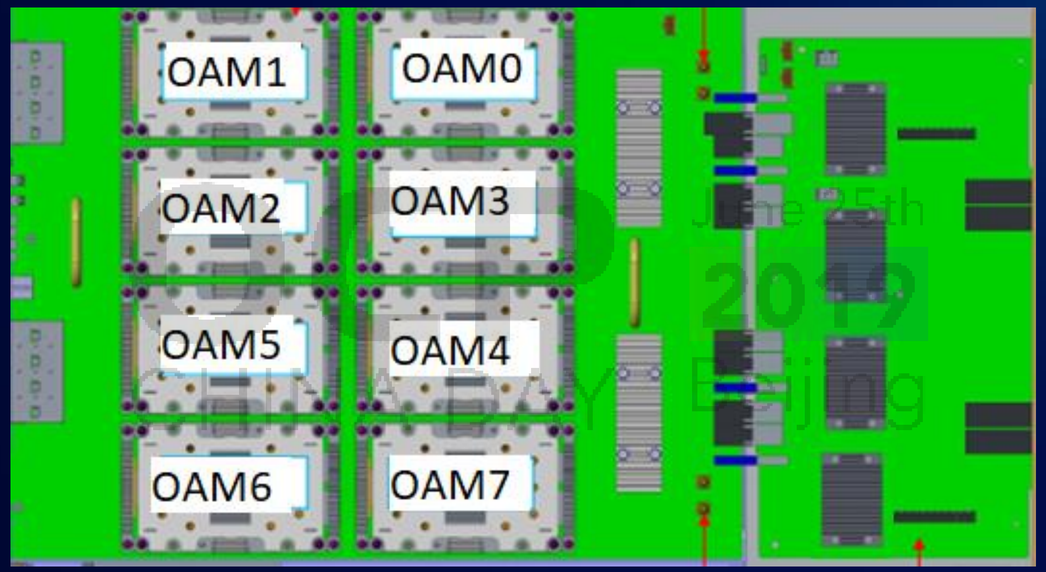
- UBB will be the major focus for current stage
- OAI-JDA subgroup is actively working on the spec
- Inspur, HyveDesignSolutions and ZT Systems are the 3 volunteer system providers to build UBB and OAM reference systems.
- UBB spec lock down meeting during 6/26-6/27.
- Target to release the UBB spec in Sept

UBB Status Update

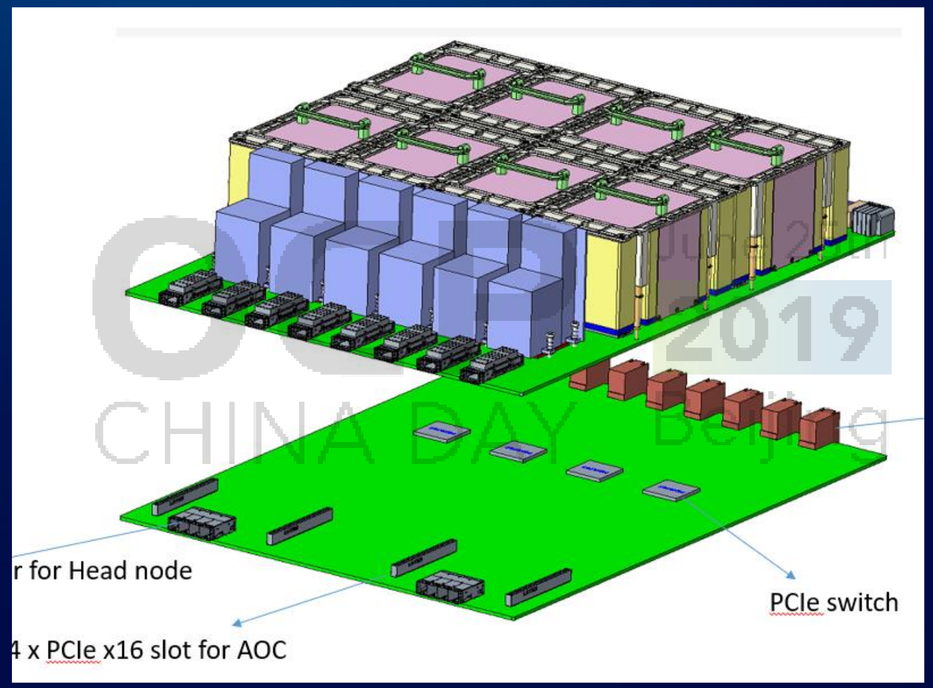
- UBB Dimension ~16.7’’ x21’’
- UBB Interconnect link: X8
- PCB material: ultra low loss
- 18L-22L
- 12V support up to 300w
- 54V support up to 450-500w



OAM Reference Systems Consideration



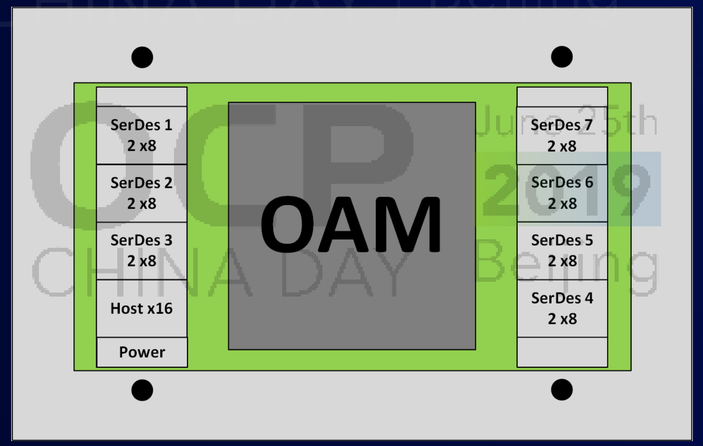
Coplanar like System



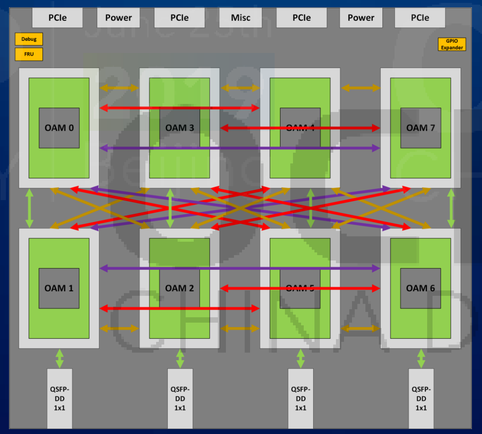
Stack like System

OAM and UBB Examples

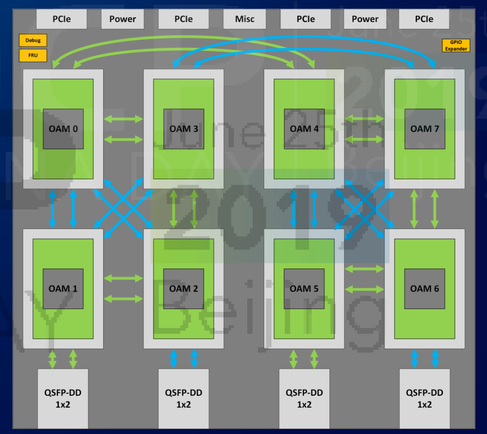
Single OAM Design



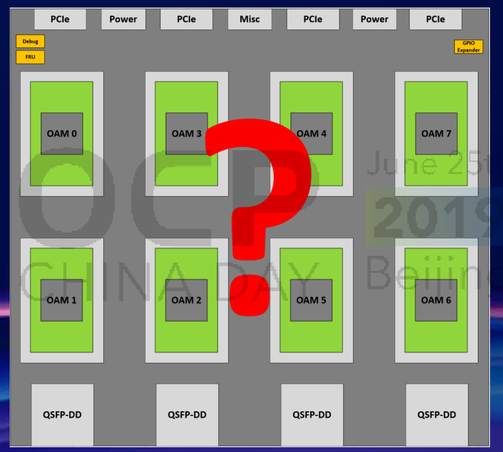
FC



HCM



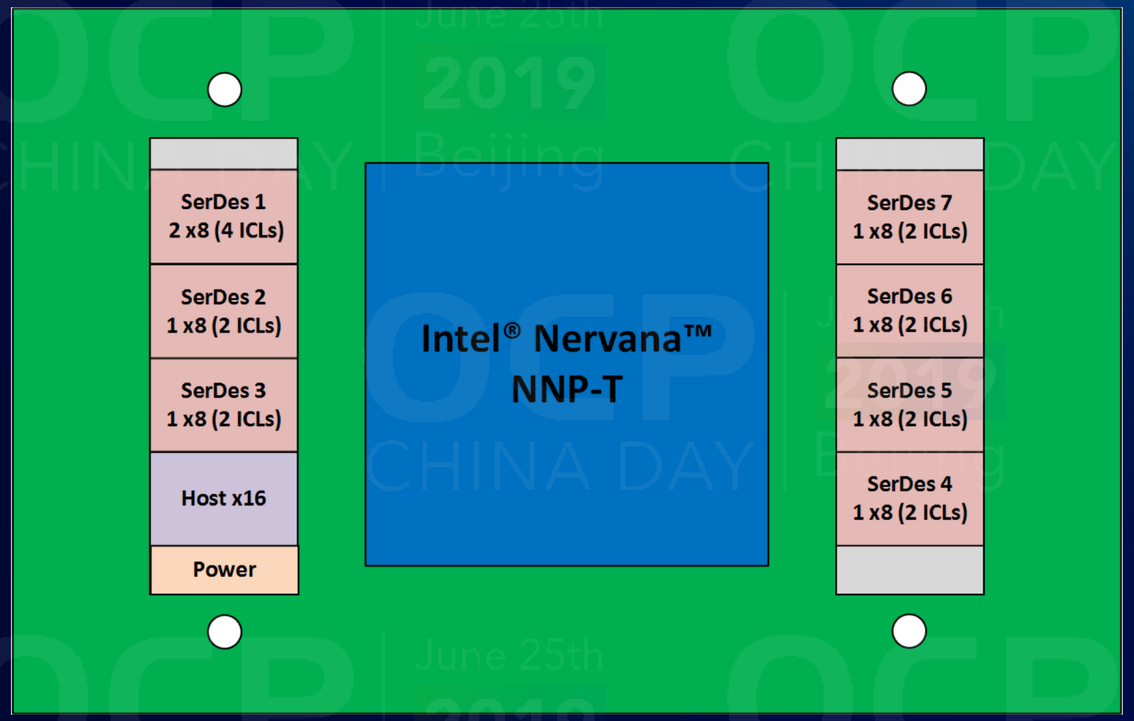
Future

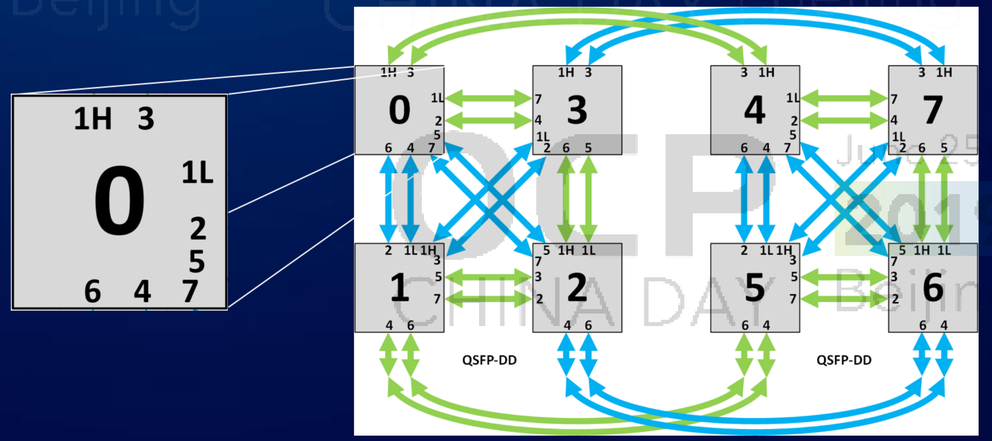


Intel® Nervana™ NNP-T OAM Specification

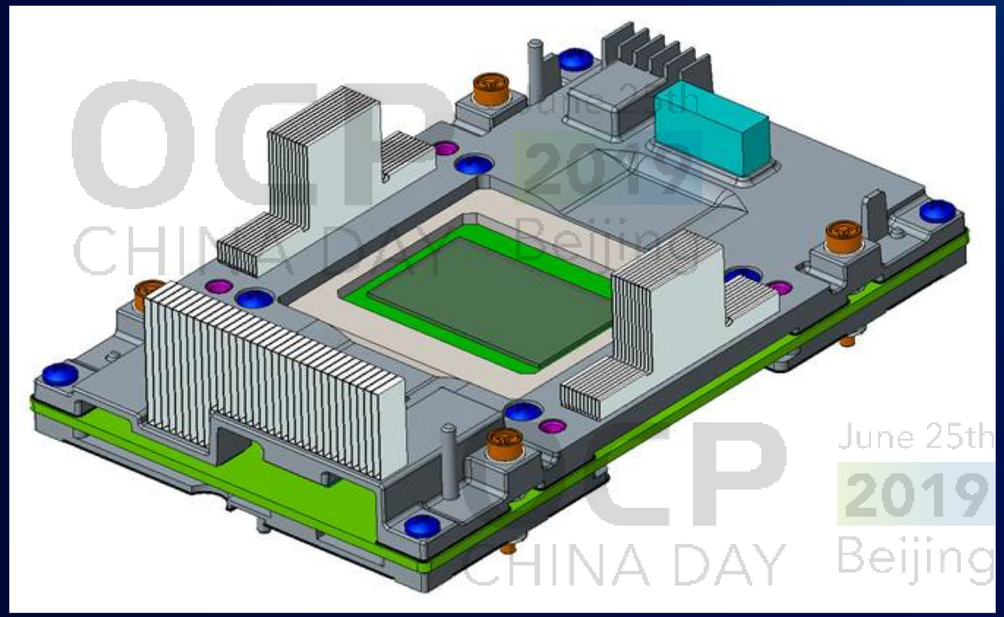
- OAM Complaint
- Power Consumption - up to 425W
- High Speed Inter-Chip Link (ICL) SerDes
- 16 ICL SerDes ports
- Each ICL SerDes port is x4 lanes

Intel® Nervana™ NNP-T OAM Pinmap

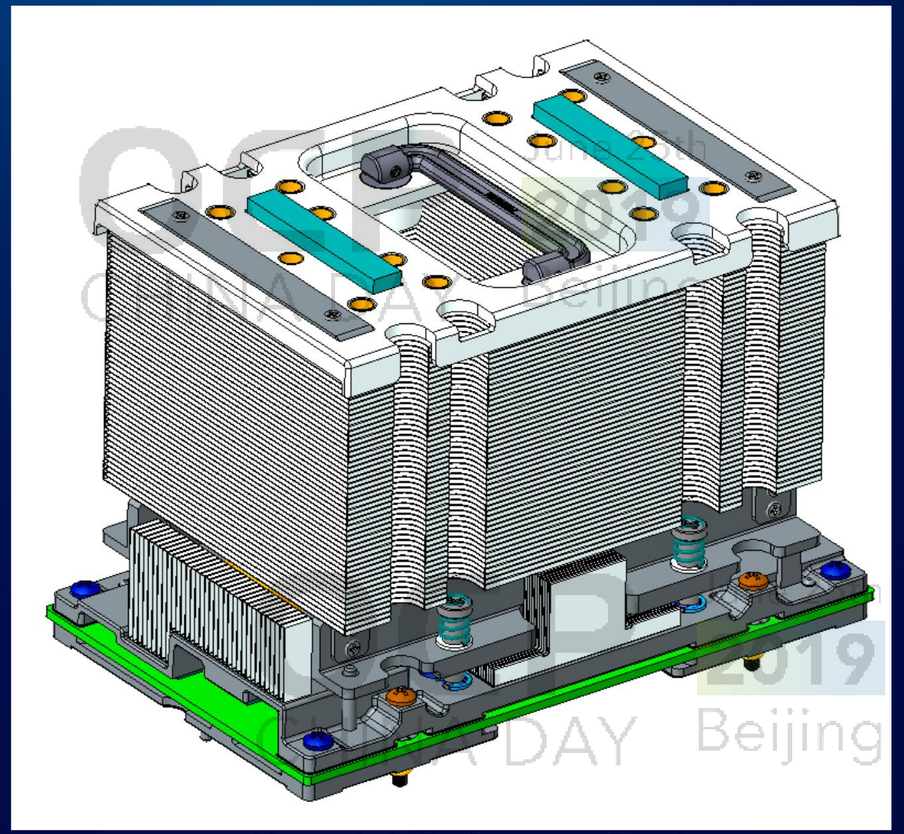




Intel® Nervana™ NNP-T OAM



Intel® Nervana™ NNP-T OAM



Heatsink Reference Design

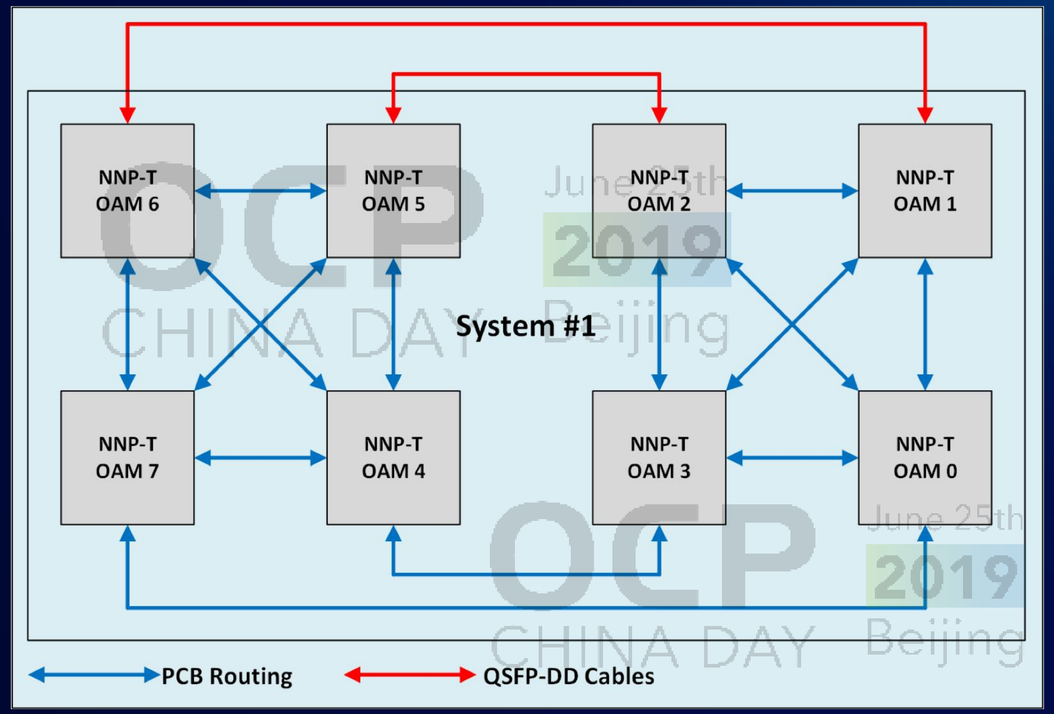
8x Intel® Nervana™ NNP-T OAM System Implementation

- Voltage and Power Requirement
  40-60V and 3.3V Voltage Input
  Total of up to 3400W with 8x 425W Intel® Nervana™ NNP-T OAM

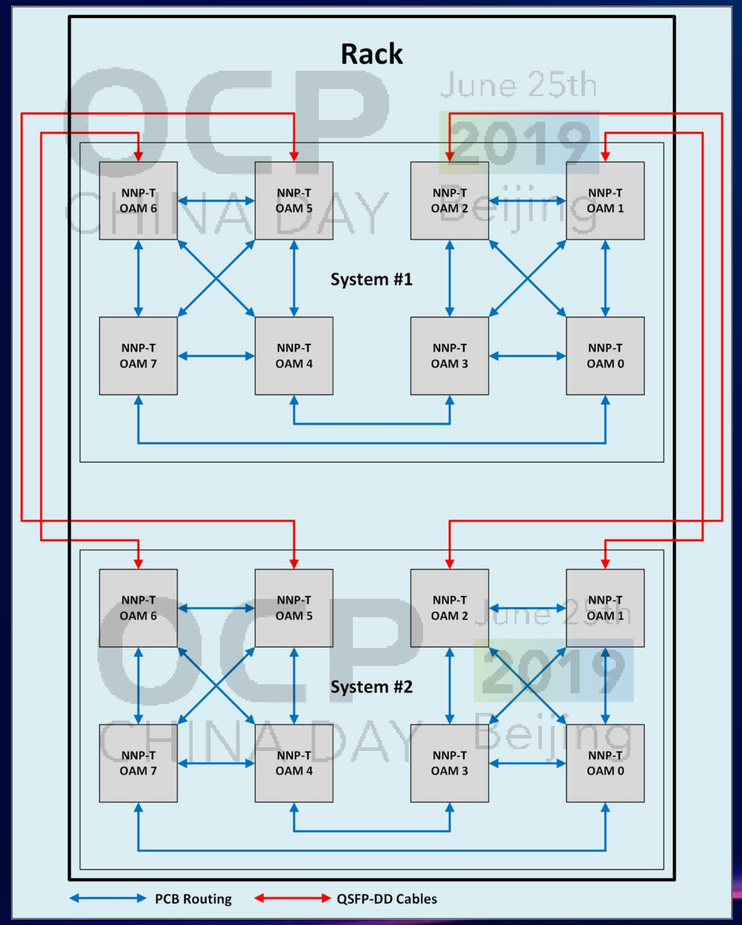
- Thermal Solution
  Support 3U/3OU Passive Air Cooled up to 35C ambient temperature

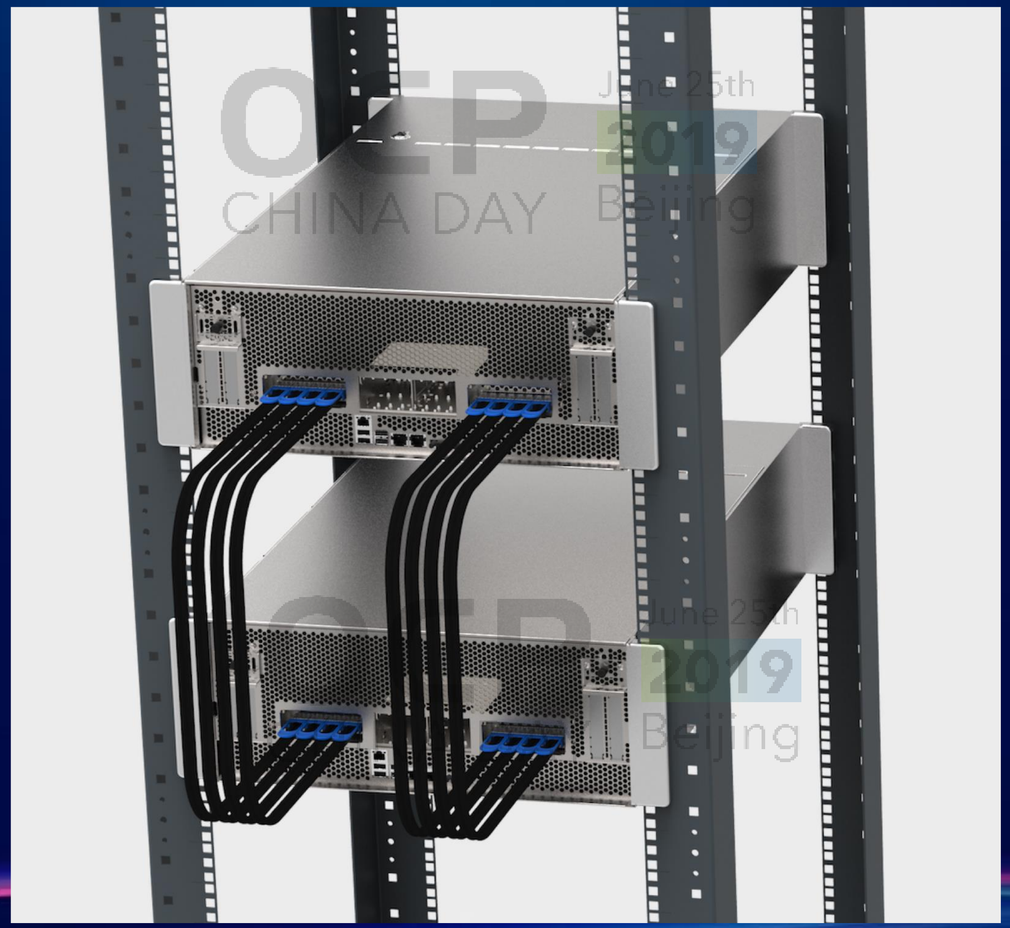
- Multi-Module Deep Learning Topology and Connectivity

Single Chassis HCM Topology

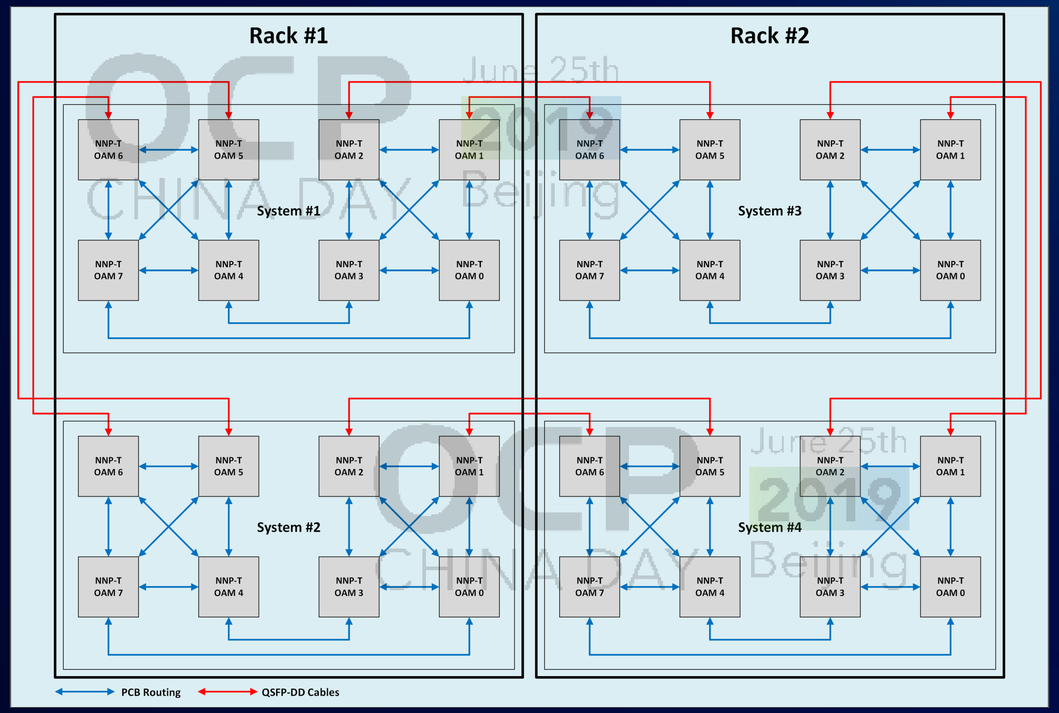


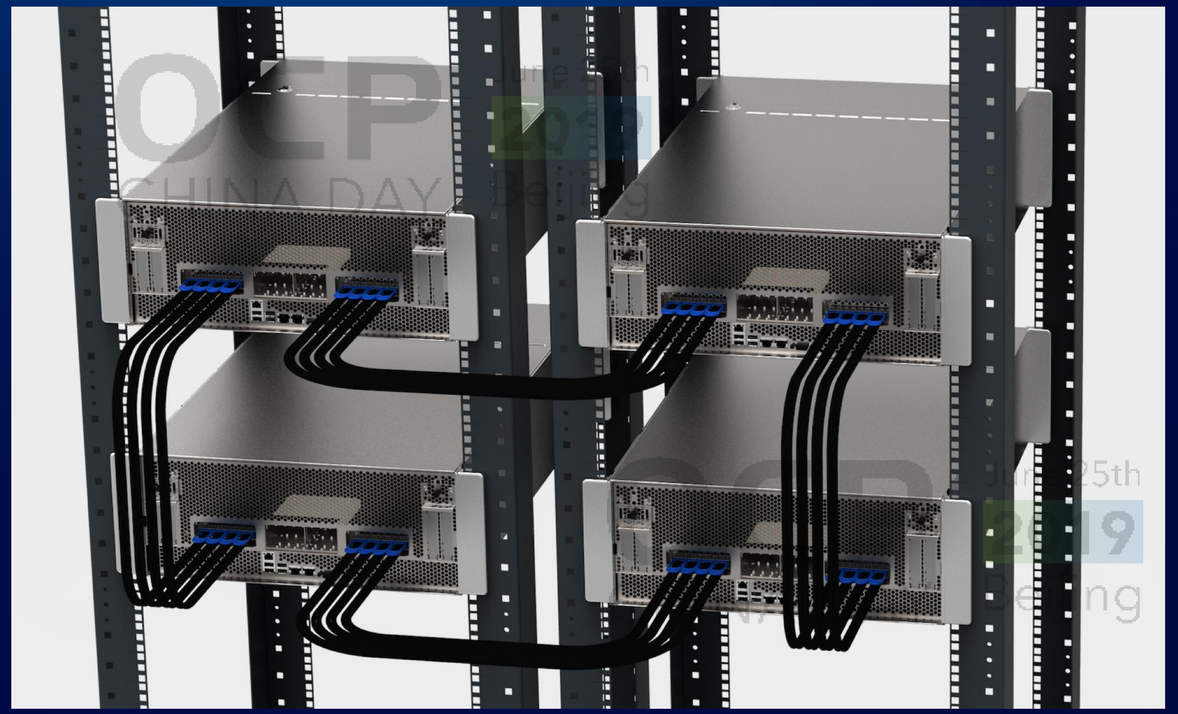
Multi-Chassis Scale Out HCM Topology





Multi-Rack Multi-Chassis Scale Out HCM Topology





Summary

- OAI group is to define the open accelerator infrastructure to accelerate adoption of new AI accelerators.
- OAM is a common form factor for AI accelerators.
- UBB is the universal baseboard for different OAMs.
- HIB and other system level module spec under evaluation
- Target to release UBB spec and show case OAI reference systems @

OCP Regional Summit
Amsterdam, Netherlands
September 26–27, 2019

Call to Action

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

Subscribe for the Mailing List:

https://ocp-all.groups.io/g/OCP-OAI

Wiki under OCP Server Project:

https://www.opencompute.org/wiki/Server/OAI

Presenters

- Whitney Zhao is a tech lead in infrastructure hardware group in Facebook. Part of her main focus is architecting HW system design for Facebook's main AI workloads. She has been driving multiple hardware-software co-design initiatives across both Facebook's training/inference areas and infrastructure challenges. 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. She has been leading the effort to form the industry standard common form factor for upcoming accelerators. She is currently co-leading the OCP OAI group and driving the open accelerator infrastructure effort, to build the ecosystem for the common form factor accelerator modules.
- Richard Ding is AI System Architect for heterogeneous computing in Baidu Technical Group. 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.
- Song Kok Hang is Principal Engineer at Intel. He was part of the original Nervana team. His main role is System Architect for high speed interconnect and high power Deep Learning platform and system architecture.
- 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.

OCP CHINA DAY June 25th 2019 Beijing

Thank you