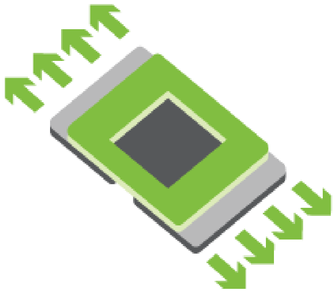

Open. Together.
OCP
SUMMIT

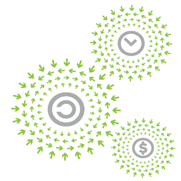
# OCP Accelerator Module (OAM)

## An Open Accelerator Infrastructure Project

HPC

Specifications





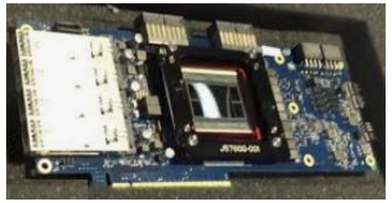
OPEN PLATINUM™

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

AI’s rapid evolution is producing an explosion of new types of hardware accelerators for Machine Learning (ML) and Deep Learning (DL)

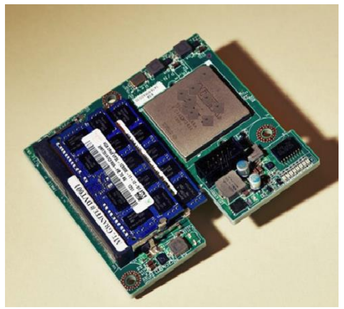
GPU | FPGA | ASIC | NPU | TPU | NNP | IPU | xPU…

# Varied Module and System Form Factors



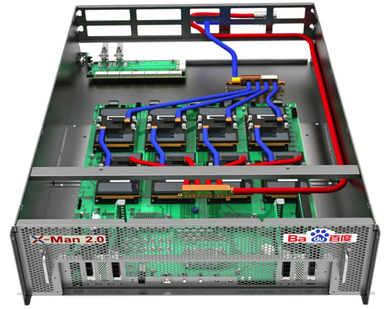
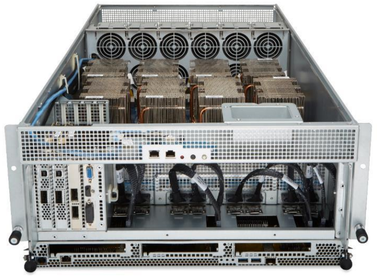
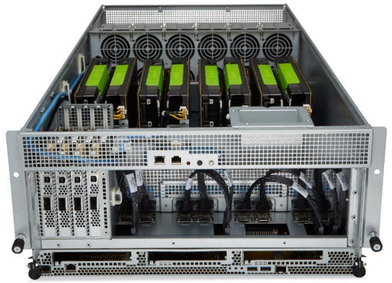











HPC

# Different Implementations

# Targeting Similar Requirements!

# Common Requirements

HPC

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

HPC

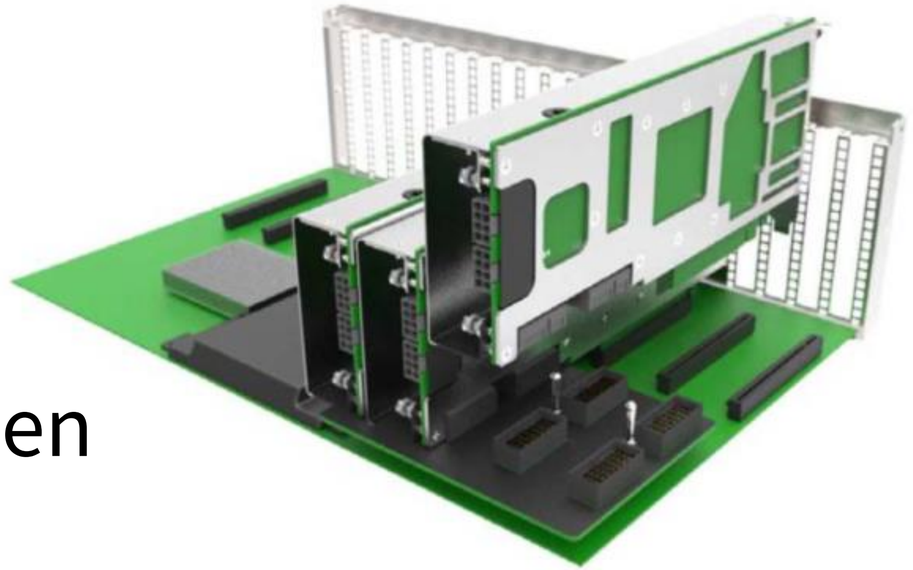
# PCIe CEM Form Factor

HPC

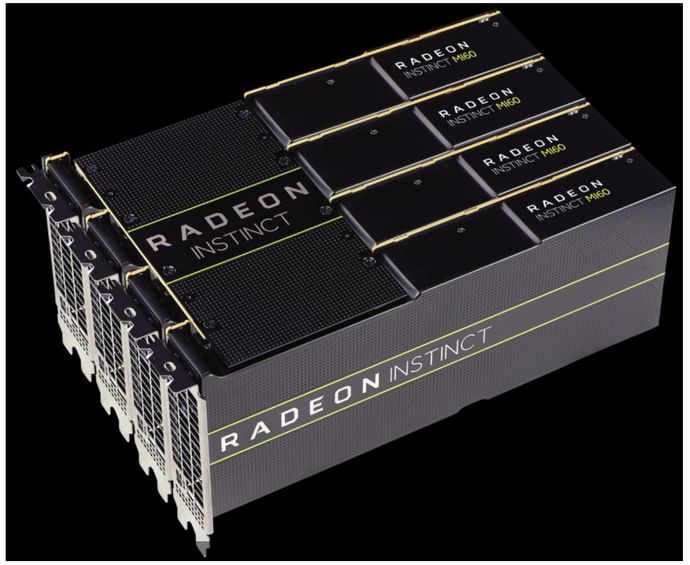
# PCIe CEM Form Factor

# is not it!

# Why not PCIe AIC?



- Multi Cards in system
- High Interconnect BW needed between Card to Card
- Too much signal loss from ASIC to HS Connectors in PCIe Form Factor
- Inter-Card Cabling is difficult and limited in supported topologies



HPC

# We need an

# Open Accelerator Infrastructure

# Our Proposal: Open Mezzanine Module

HPC

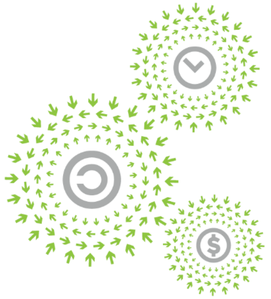
- High-density Connectors for input/output Links
- Low signal insertion loss → high-speed interconnect
- Enough space for Accelerators and associated local logic & power
- Flexible for heatsink design for air-cooled & liquid cooling
- Flexible inter-Module interconnect topologies

# Complementary Support

HPC

- **OAM** is an Open Accelerator Module supporting multiple suppliers
- A multi-OAM, Universal Baseboard (**UBB**) supporting 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**) supporting all Chassis, Trays, UBBs, and OAMs as well as the Hosting Head Node

Specifications



OPEN PLATINUM™

HPC

# Different Neural Networks benefit from different Interconnect Topologies

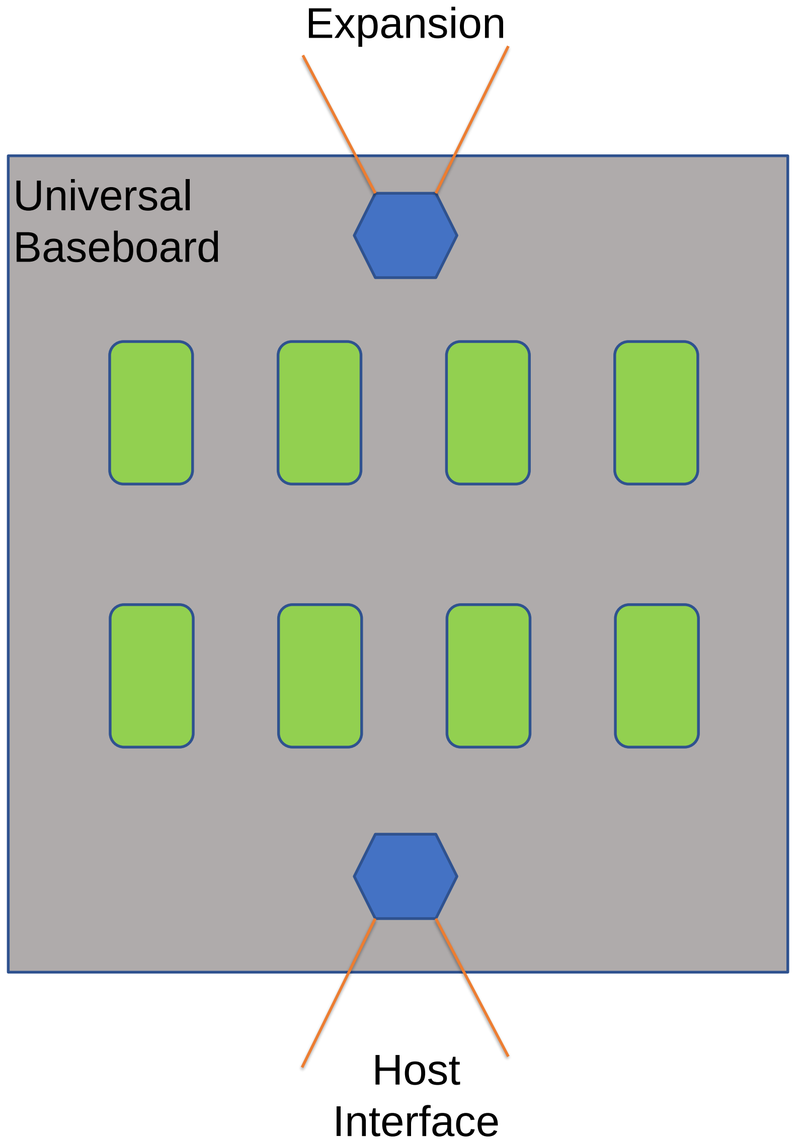
# Universal Baseboard (UBB)

Consider a Grid of Planar OAM sites

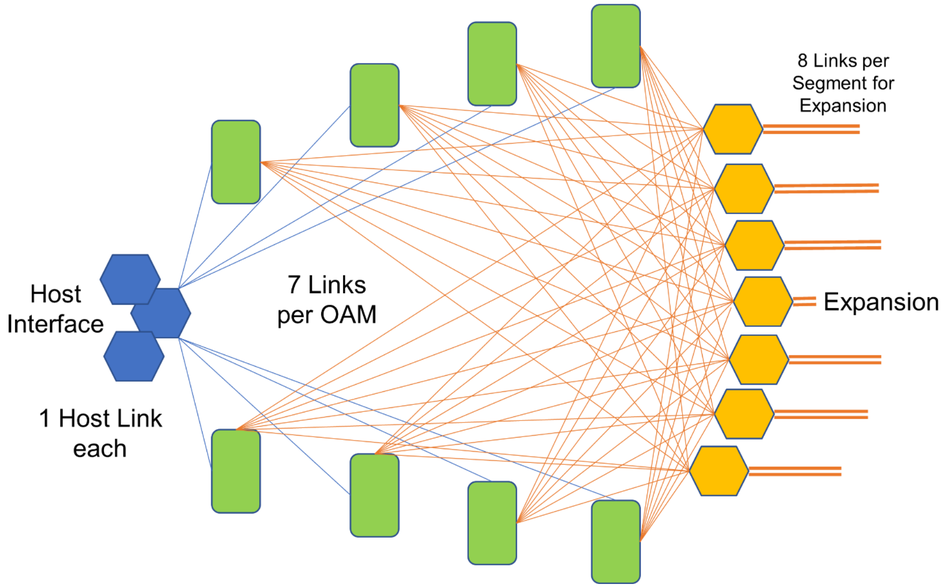
Standard Volumetric

Protocol Agnostic Interconnects

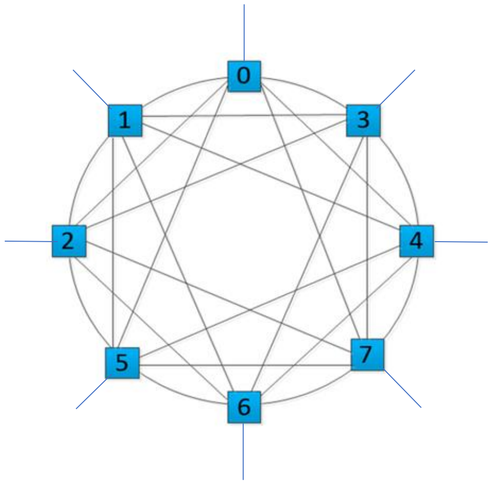
Wires are Wires!



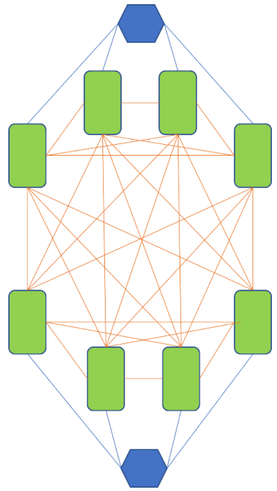
# With different interconnect topologies



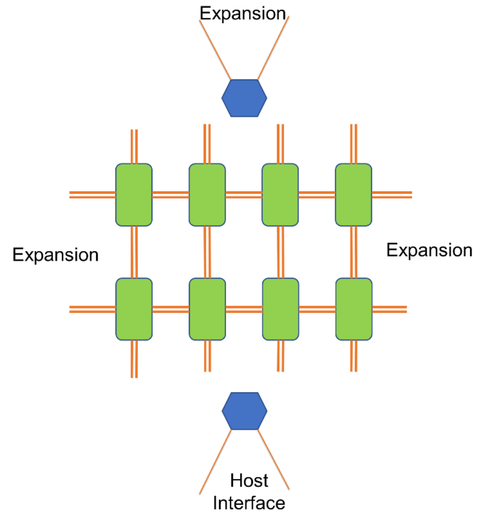

A Grid of interconnected OAMs,
Max Bisection BW
One Hop Away
Ready for Expansion




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

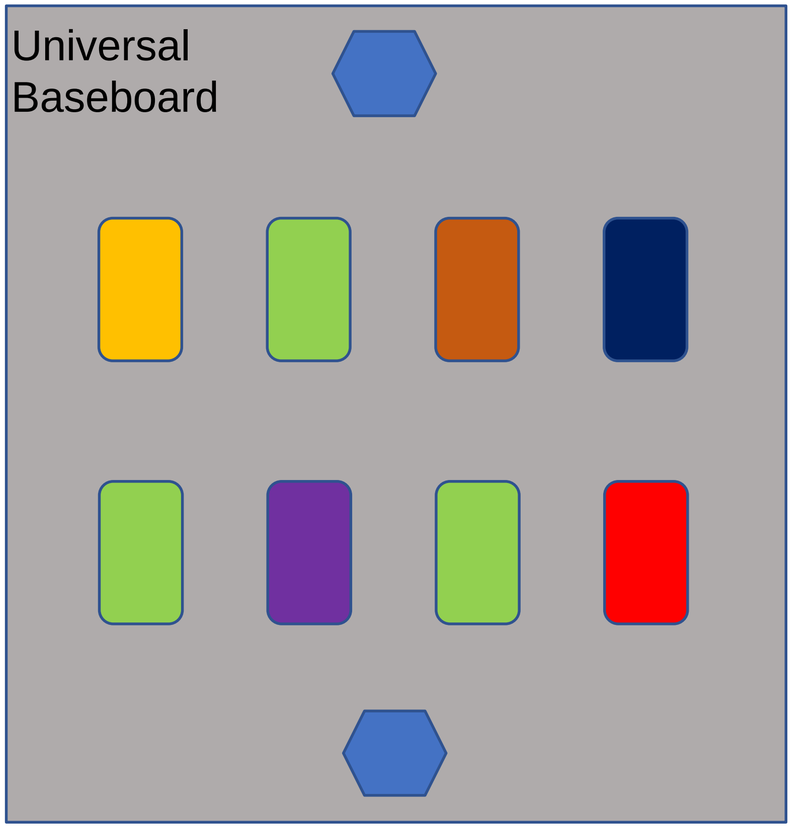
# Heterogenous OAMs

These Modules need not be of the same type

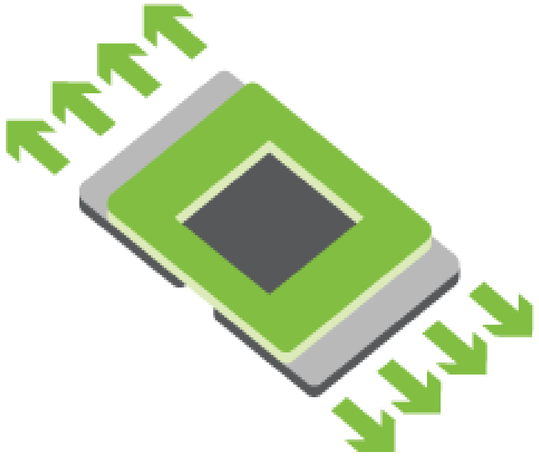
Each one may be suited for a specific application/task

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

Chained, pipelined processing stages



# Current Work:  OAM Spec

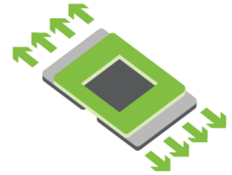


HPC

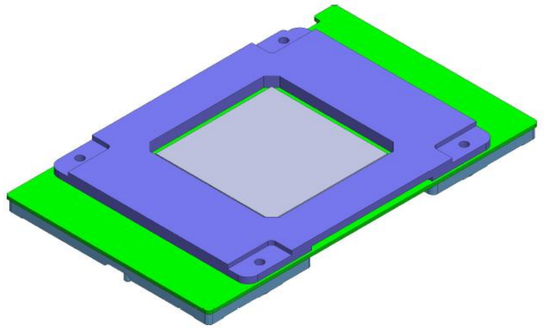
Specifications

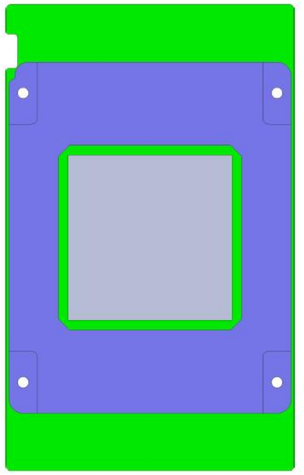
OPEN PLATINUM™

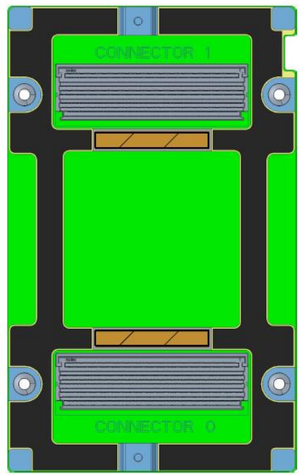
# OCP Accelerator Module Spec



- Support both 12V and 48V as input
- Up to 350w(12V) and up to 700w(48V) TDP
- 102mm x 165mm
- 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
- Expect to support up to 450W (air-cooled) and 700W (liquid-cooled)
- Up to 8* Modules per system
- System management and debug interfaces

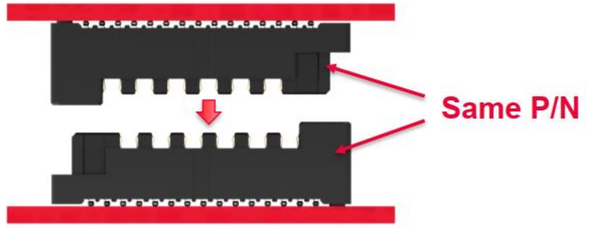


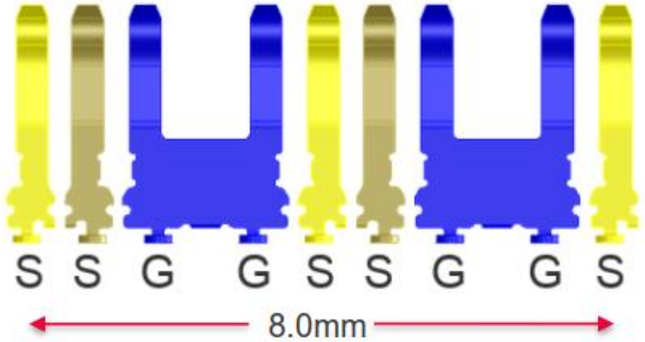




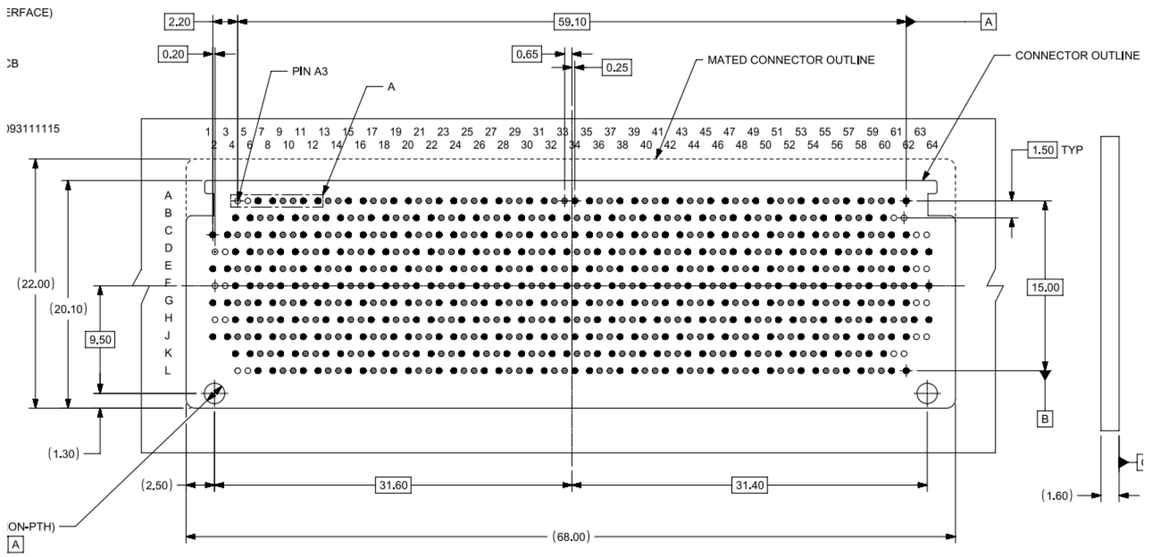
# Molex Mirror Mezz Connector

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



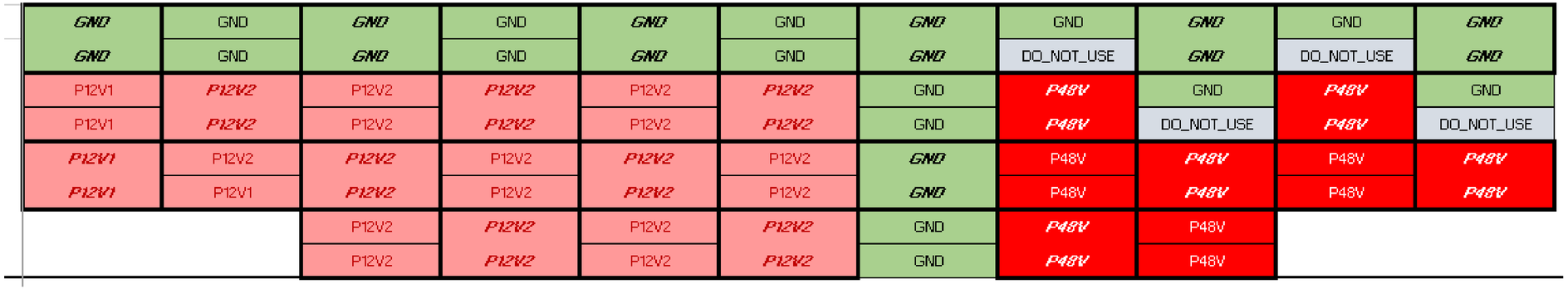

Images courtesy of Molex



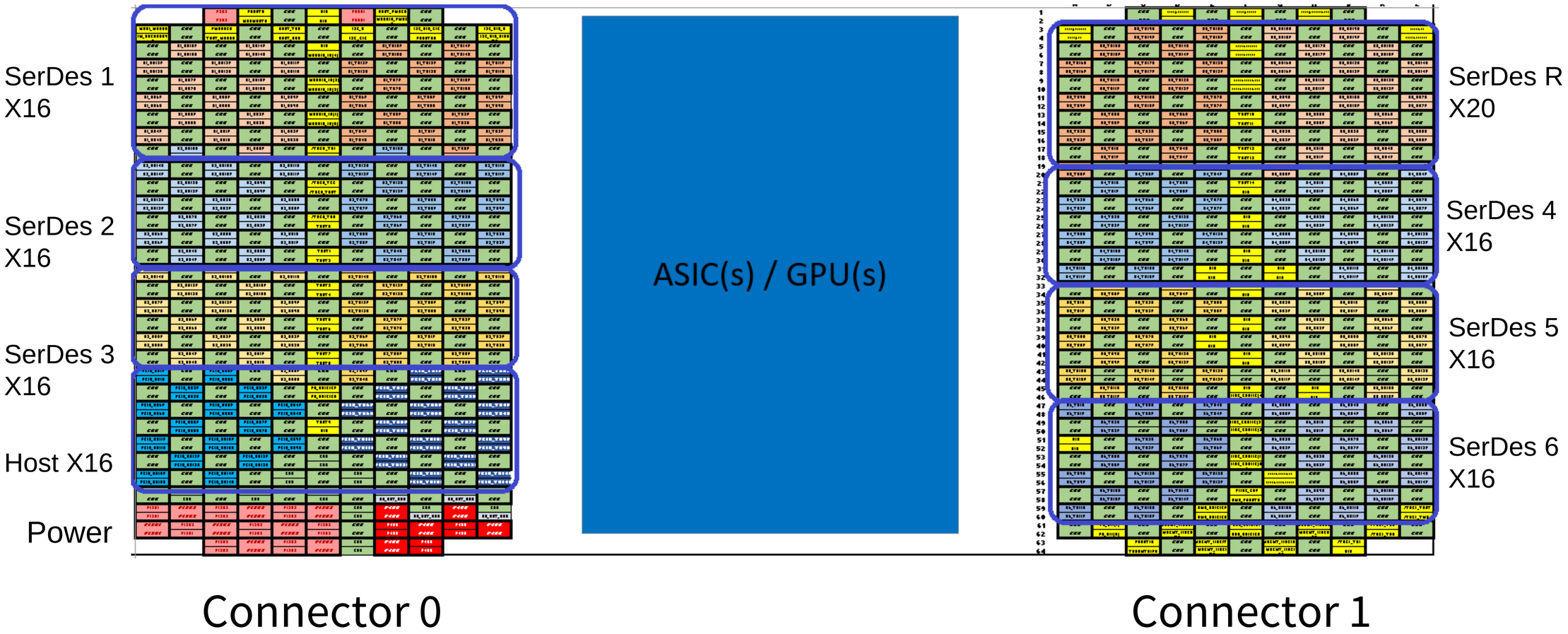
# Module Power

- Support both 12V and 48V as input
  - 12V to support up to 350w TDP
  - 48V to support up to 700w TDP

| Power Rail | Voltage Tolerance | # of pins | Current Capability | Status |
|---|---|---|---|---|
| P12V | 11V min to 13.2V max | 27 | 27A (when at 11V) | Normal Power |
| P12V Mandatory | 11V min to 13.2V max | 5 | 5A (when at 11V) | Normal Power |
| P48V | 44V min to 60V max | 16 | 16A (when at 44V) | Normal Power |
| P3.3V | 3.3V±10% (max) | 2 | 2A | Normal Power |

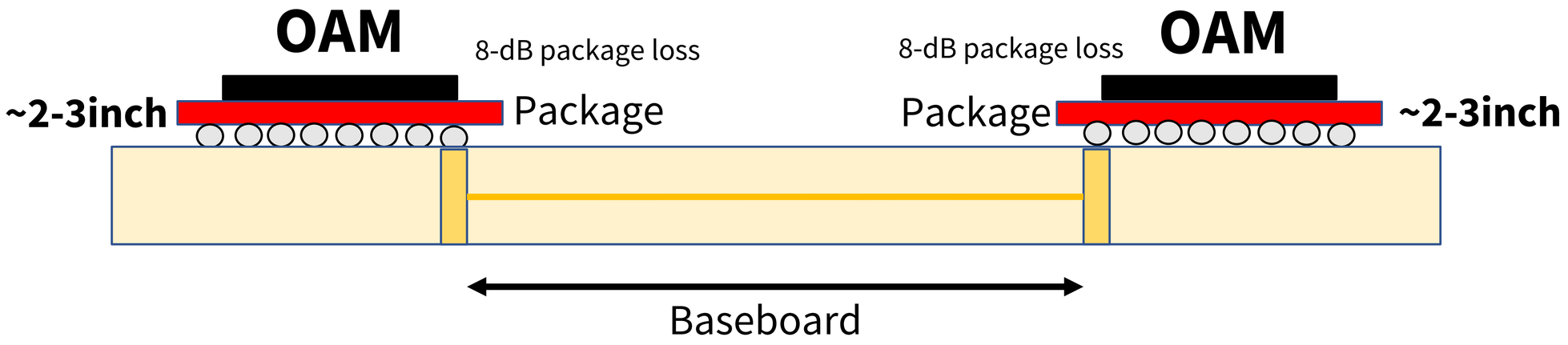


# Module Pin Map



# Interconnect end-to-end Channel Loss

- The module interconnection channel total insertion loss @28Gbps should not be over -8dB
- System baseboard IL budget = Die to Die IL from each OAM supplier – 16dB

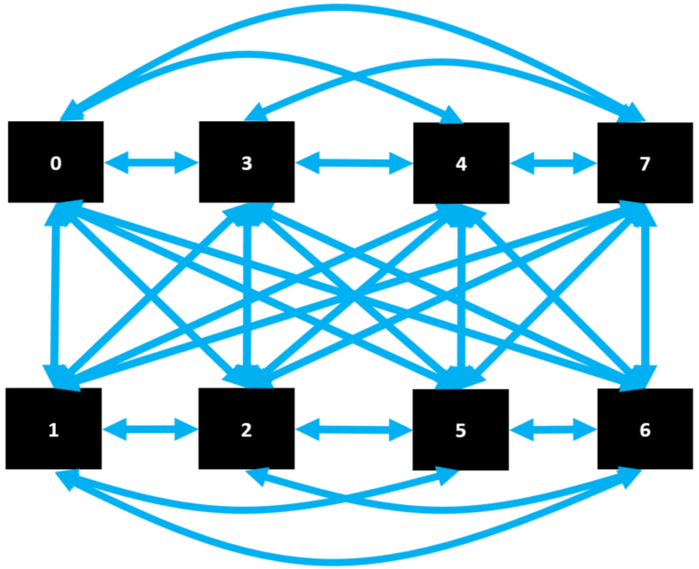


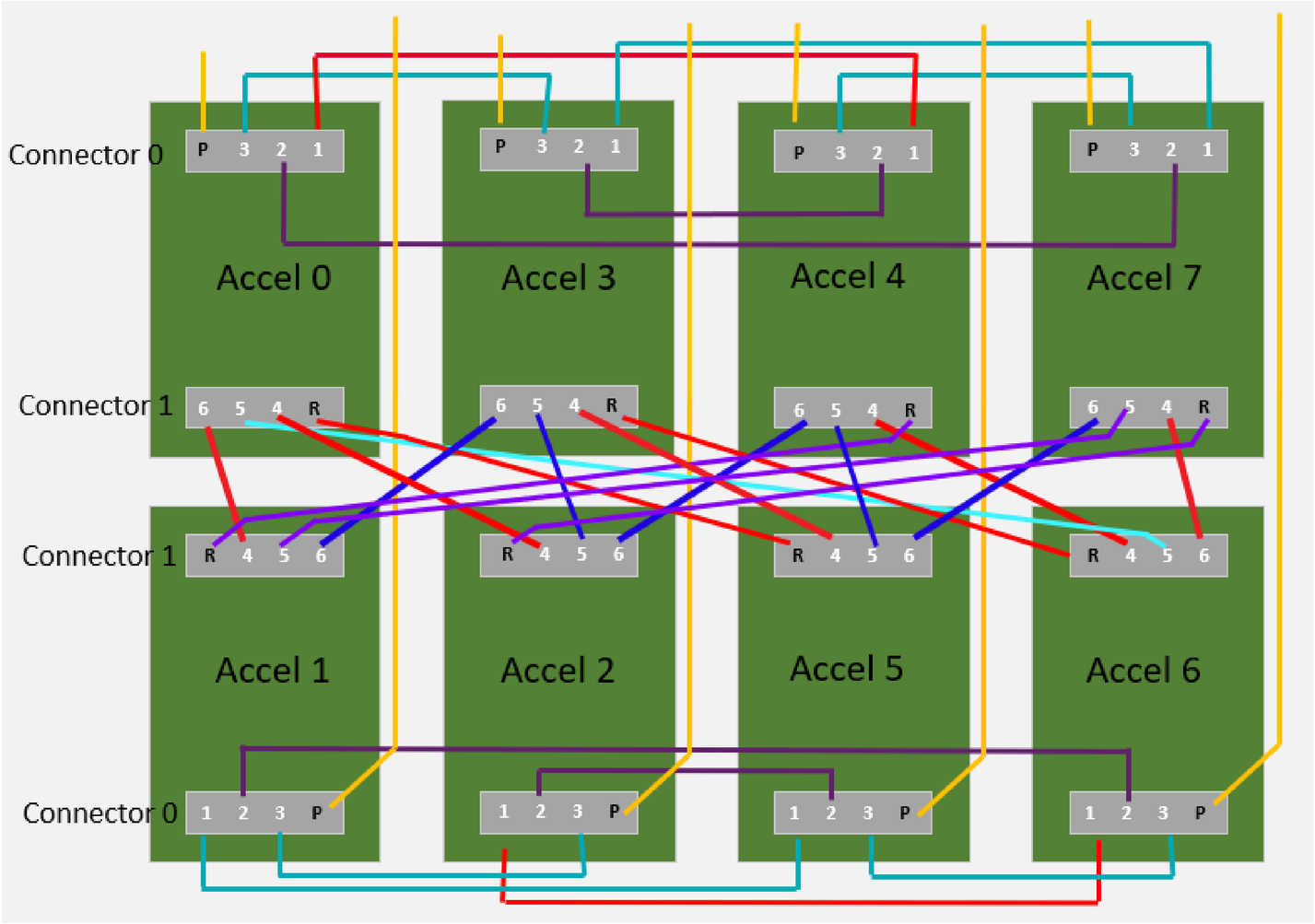
# System Management/Debugging

- Sensor reporting
- Error monitoring/Reporting
- Firmware Update
- Power Capping
- FRU Information
- IO Calibration
- JTAG/I2C/UART interfaces for debugging

# OAM Topology Examples

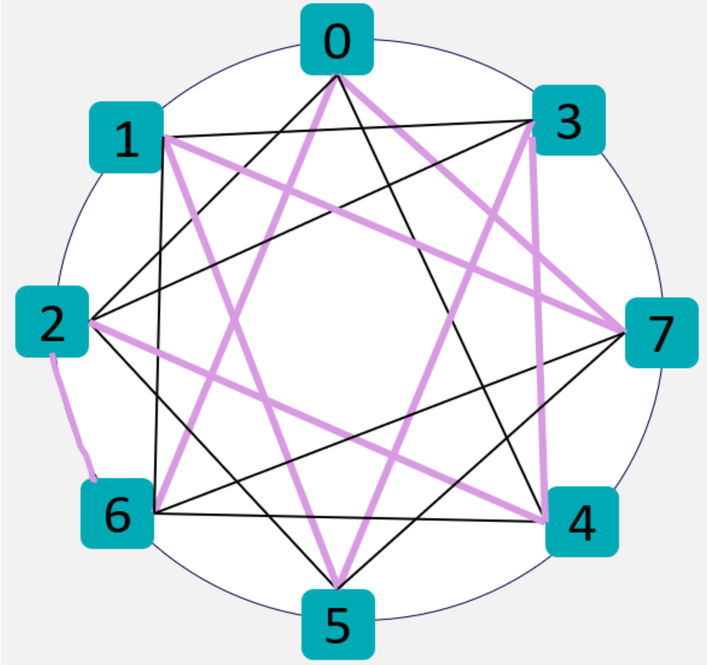
## Fully Connected w/ 7 links

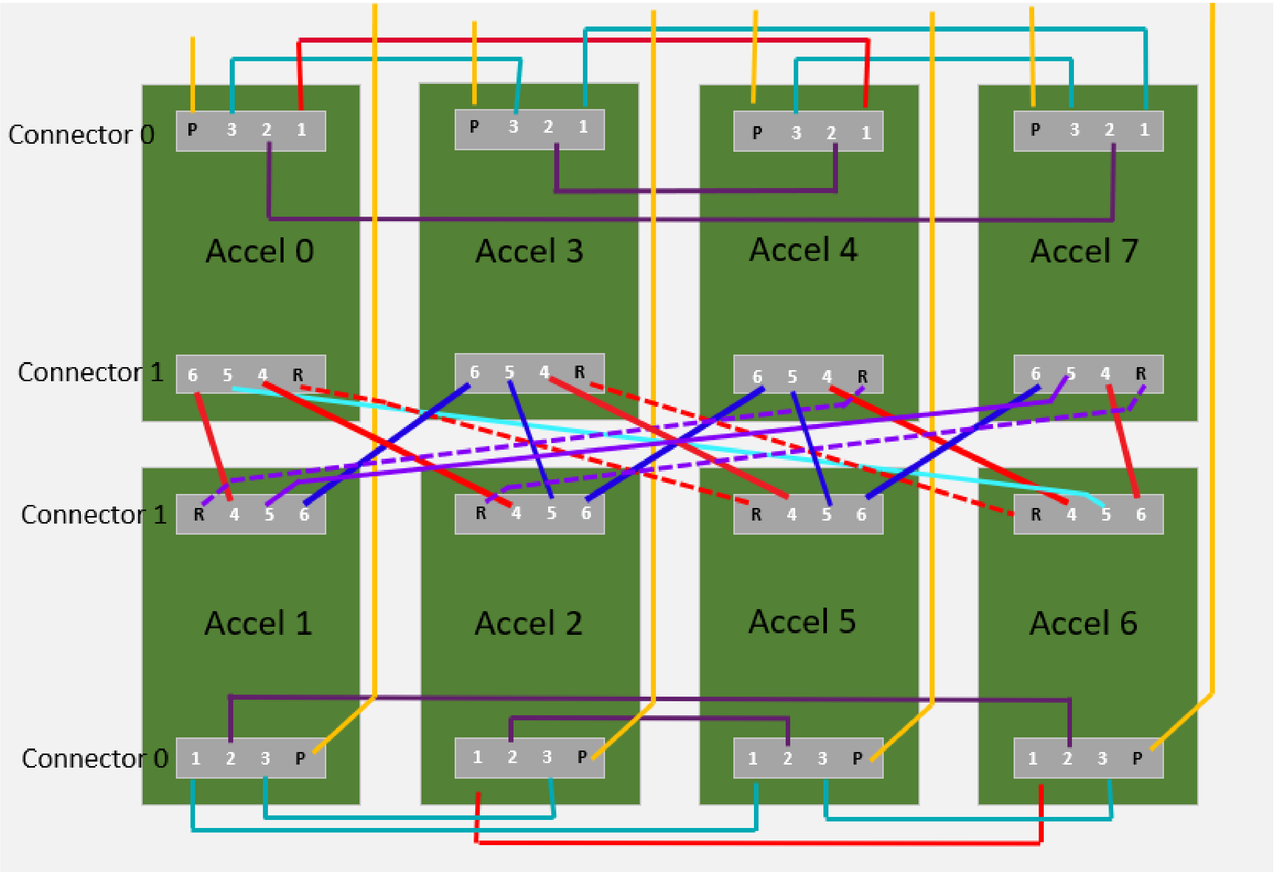


# OAM Topology Examples

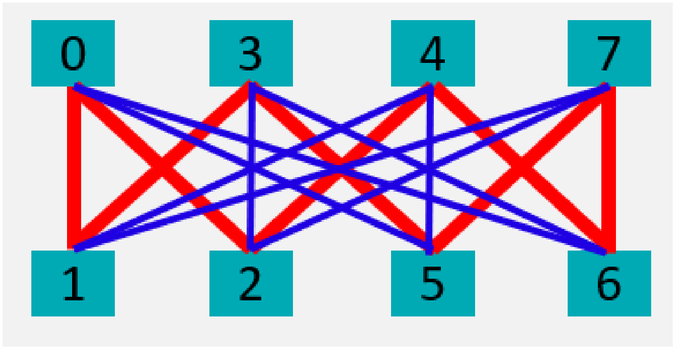
## Almost Fully Connected w/ 6 links

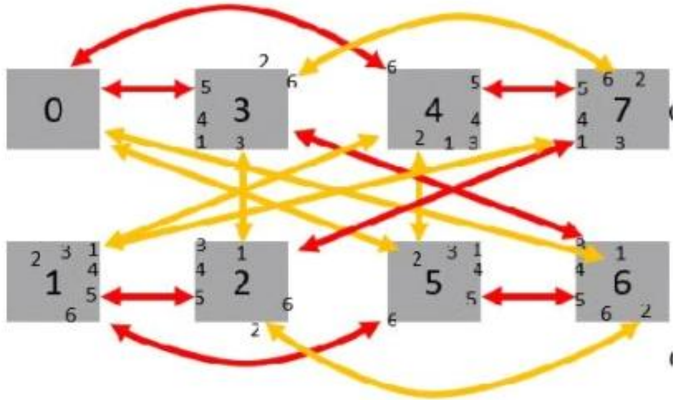


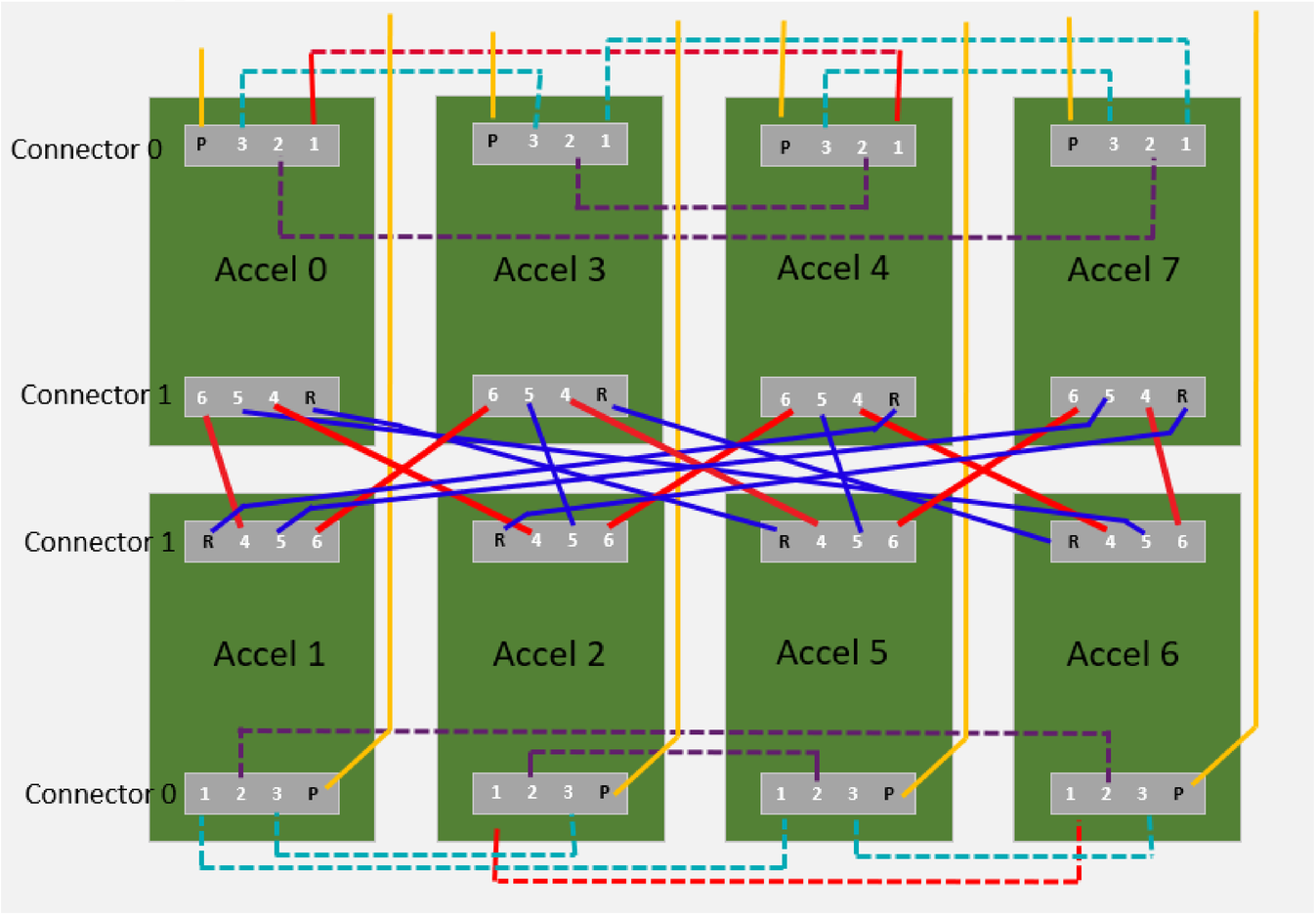



# OAM Topology Examples

## Rings w/ 4 links



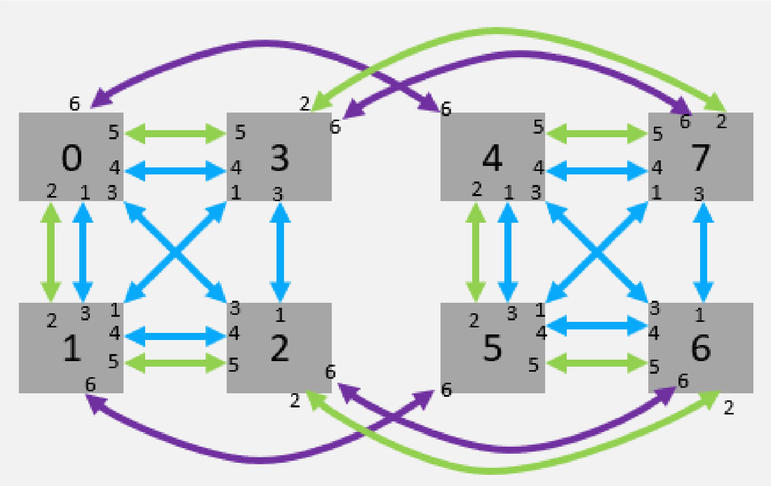



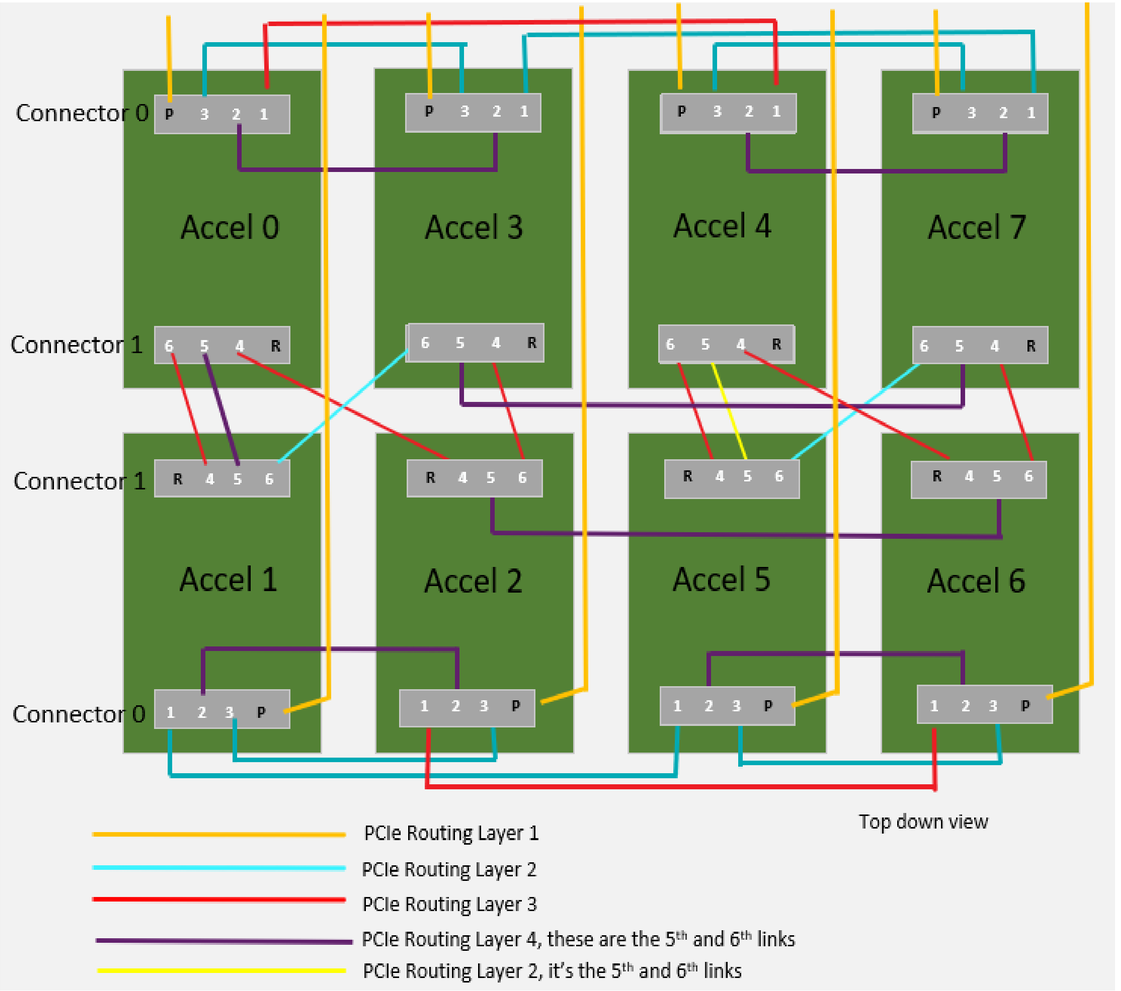

**Port 1/3/5/R for AISC which has 4 links on both Conns**

**Port 4/5/6/R for AISC which has 4 links on Conn1 Only**

# OAM Topology Examples

## Hybrid Cube Mesh

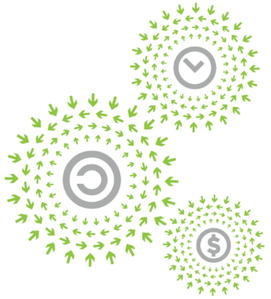


HPC

# OAM Mechanical/Thermal Features

Specifications



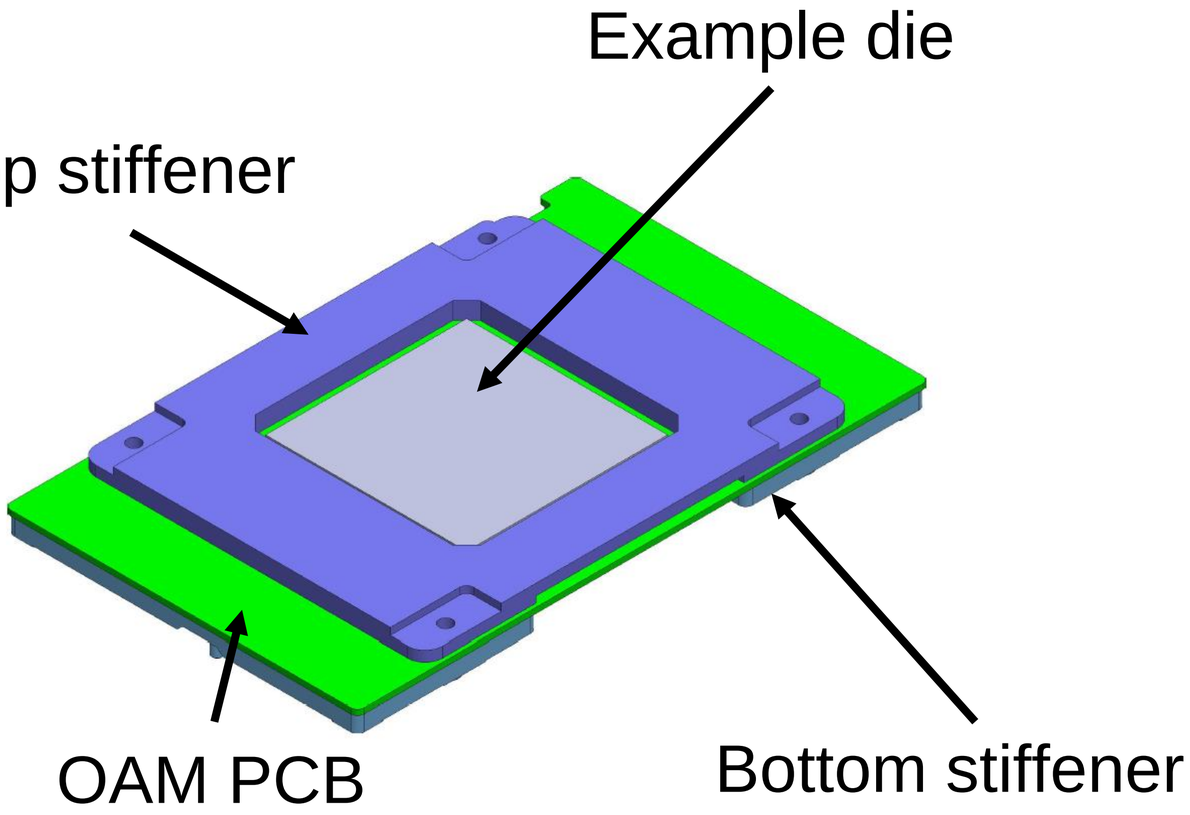
OPEN PLATINUM™

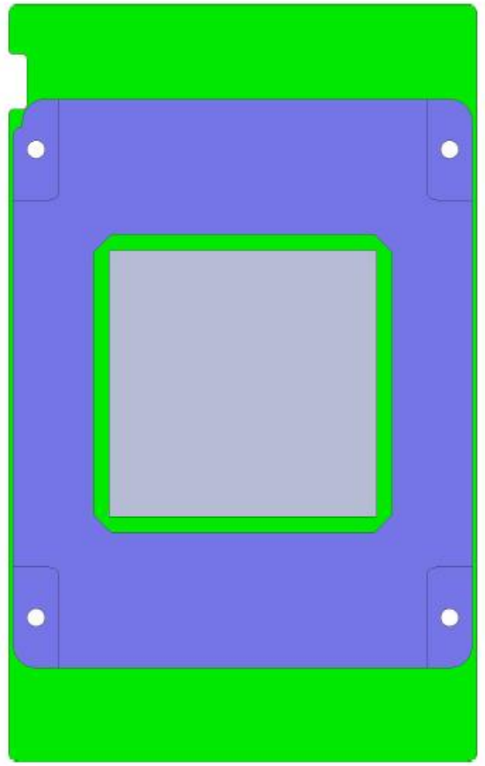
# Overview OAM: Mechanical/Thermal

Goals:

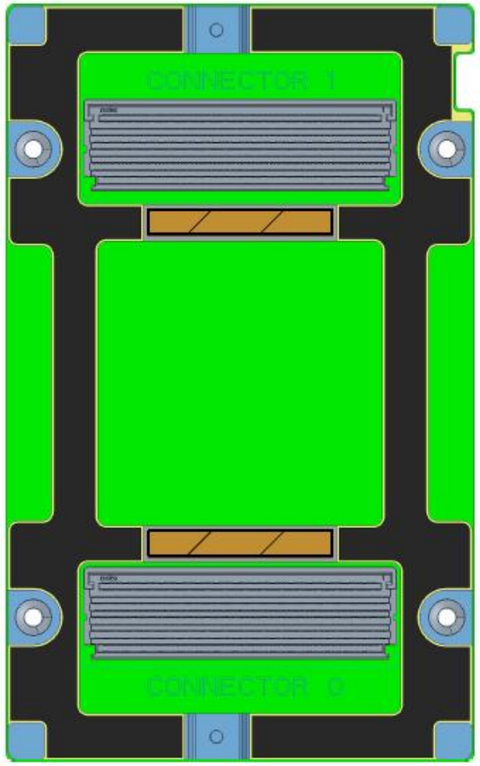
1. Provide a basic framework such that OAM from different vendors can be used in the same system.
2. Provide a full reference design such that redesign is minimal.

# Overview OAM: Mechanical/Thermal



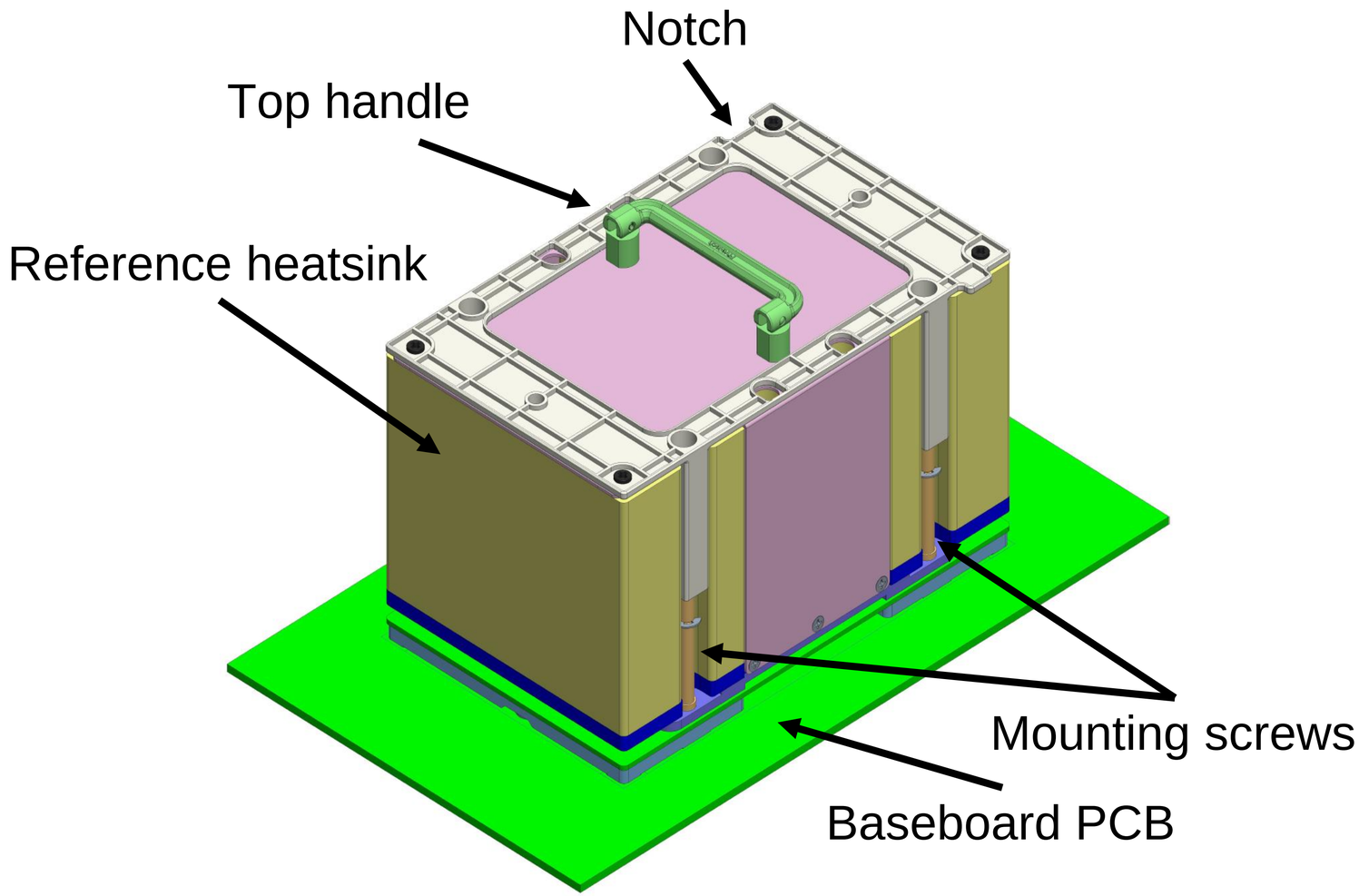



TOP VIEW



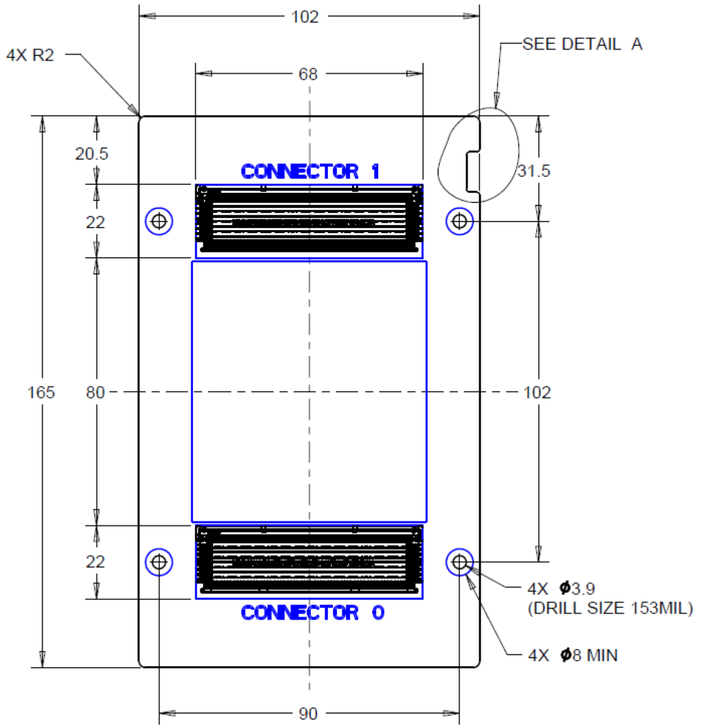

BOTTOM VIEW

# Overview OAM: Mechanical/Thermal

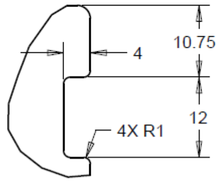


# Mech Requirements – OAM PCB

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




BOTTOM VIEW

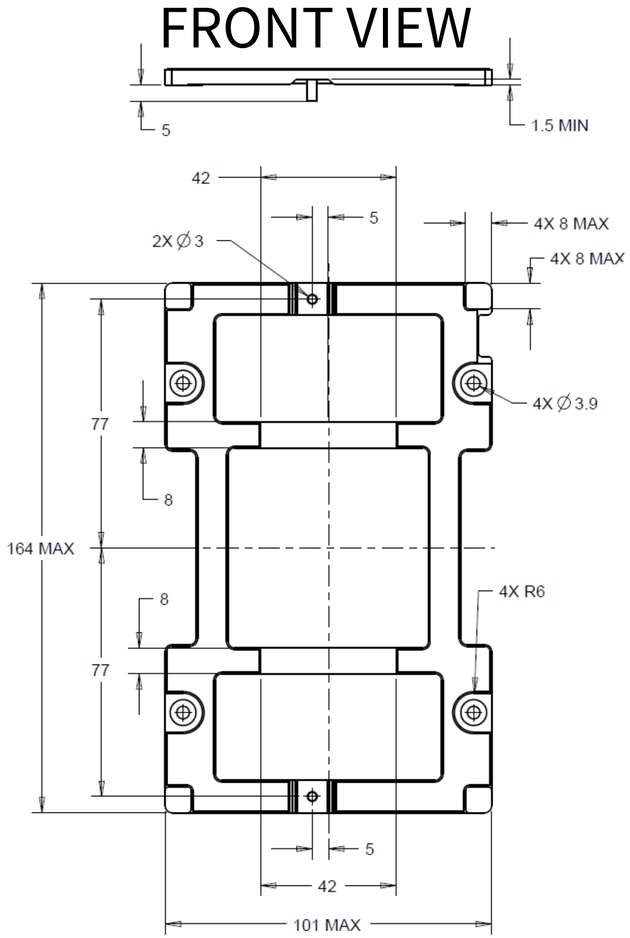


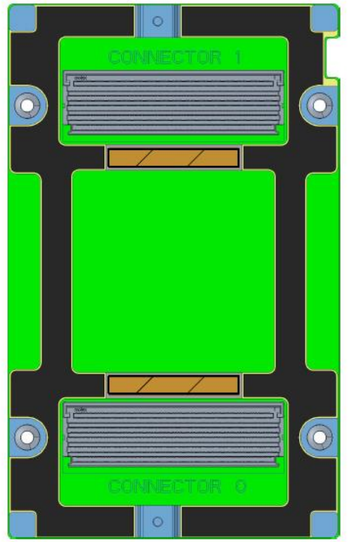

DETAIL A
NOTCH LOCATION

# Mech Requirements – OAM Bottom Stiffener

- 5±0.15mm stiffener as required by connector
- 3mm alignment pins that extend 10mm below OAM PCB surface
- Die spring (rectangular profile coil spring) to provide unmate force
- EMI gaskets for grounding to baseboard


SIDE VIEW, FULL ASSEMBLY
(BOTTOM STIFFENER IN BLUE)



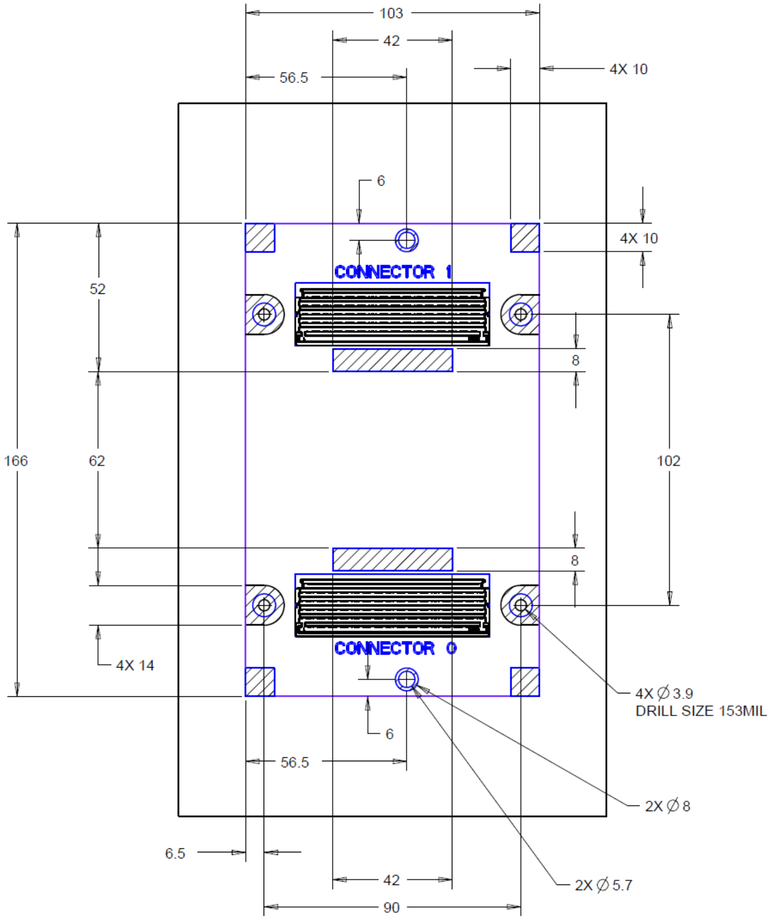




BOTTOM VIEW

# Mech Requirements – System Baseboard

- Component KOZ 103 x 166mm: 0mm height
- Cross-hatched locations: Grounding Pads
- EMI grounding pads located north and south of the connectors
- 4x Mounting Holes for M3.5 screws
- 2x SMT nuts used as alignment features

TOP VIEW

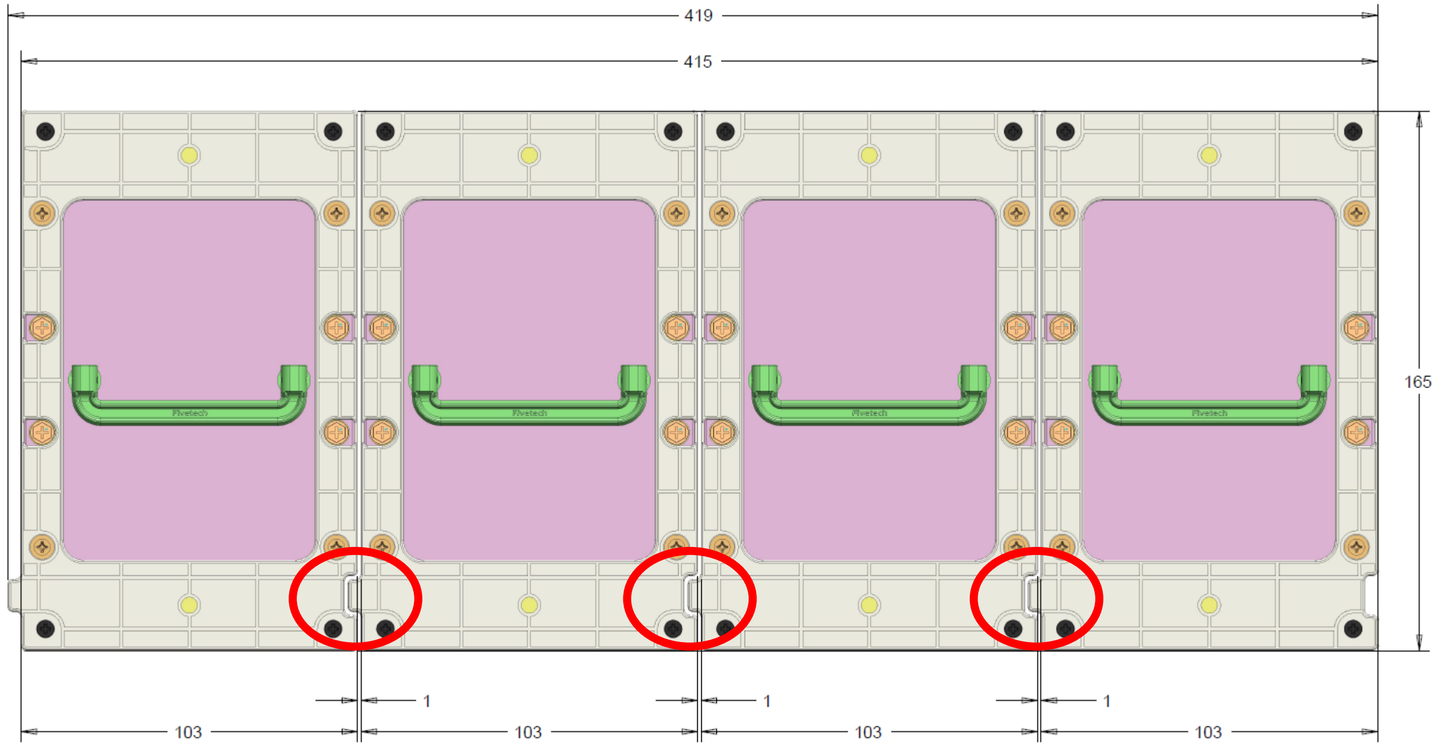


# Mech Recommendations – Alignment Features (1)

Notch provides orientation and keying (OPTIONAL, BUT RECOMMENDED)

Alignment: ±1mm

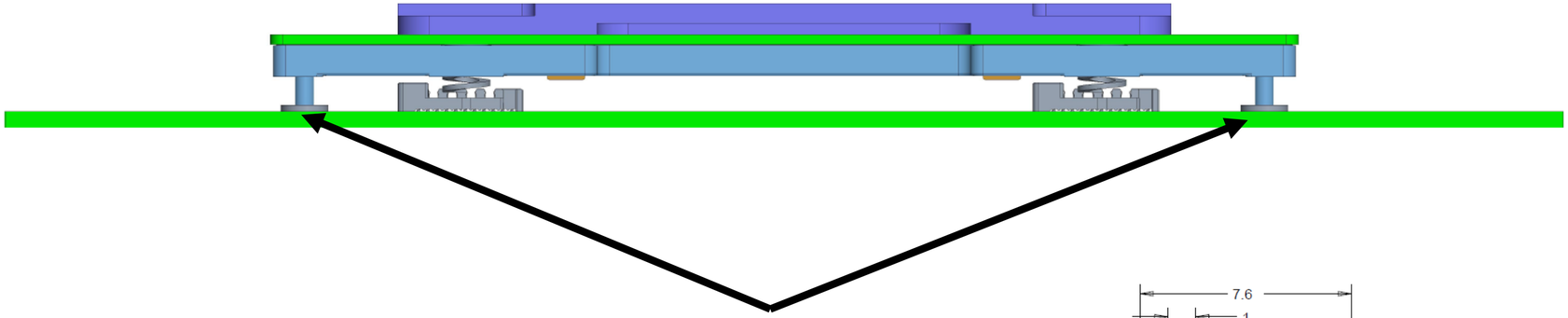
TOP VIEW



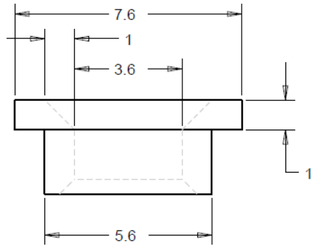
# Mech Recommendations – Alignment Features (2)

Alignment pins on bottom stiffener: ±0.3mm

SIDE VIEW



SMT nuts on baseboard



# Mech Recommendations – Alignment Features (3)

Molex Mirror Mezz Connector Gatherability: 0.76mm

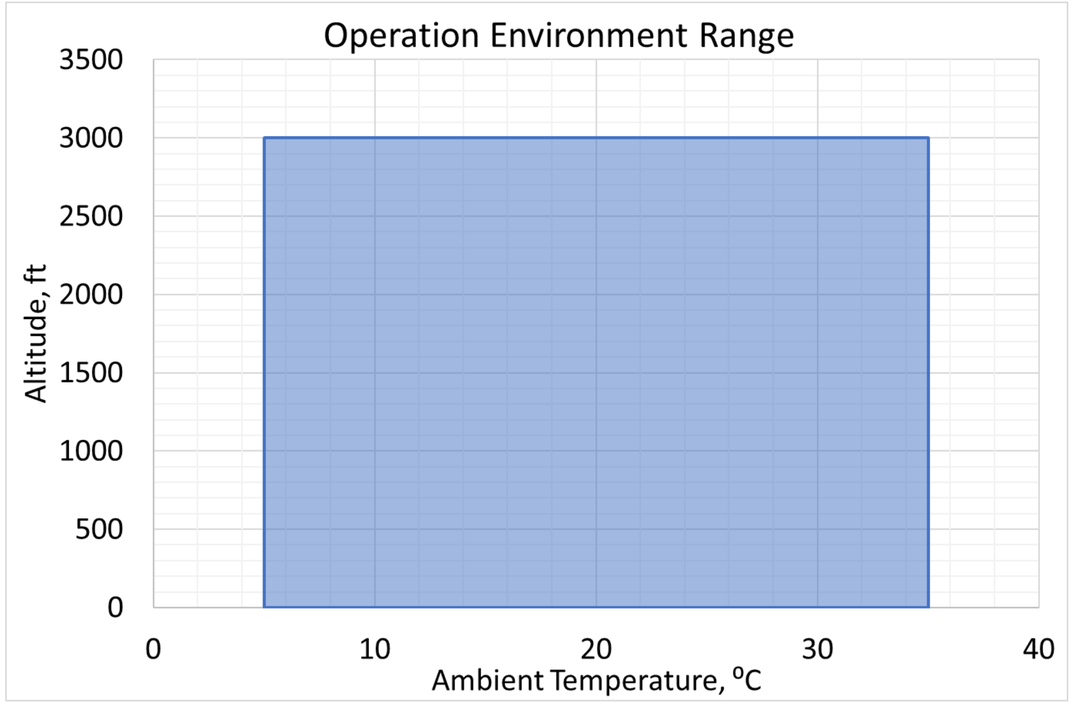
SIDE VIEW

# Mech Recommendations – HS Reference Design

- Heatsink reference design shown for 3U air cooled system
  - Facebook booth to examine full chassis platform
- Top handle to accommodate handling for tight pitch and large weight (max 2kg)
- Long M3.5 mounting screw design for easy serviceability
- Only one replaceable heatsink assembly for the module
- Other heatsink parts and TIMs should not need replacement over the module lifetime

# Thermal Requirements – Operation Environment

- Ambient Temp: 5°C to 35°C
  - Approach Temp: 5°C to 48°C
- Altitude: sea level to 3000ft
- Humidity: 20% to 90%
- Cold boot temp limit: TBD
- Storage temp: -20°C to 85°C




- No ambient temp compensation/de-rating for altitude

# Summary

Server

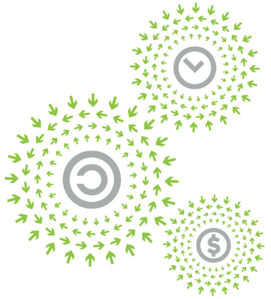
HPC

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

Specifications

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

- **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



OPEN PLATINUM™

# Call to Action

Server

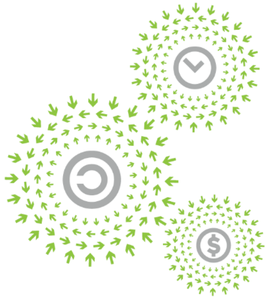
HPC

We invite you to join the OAI 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

Specifications



OPEN PLATINUM™

# OAM Infrastructure Talk at Server Track

SERVER

Refer to our OAM Infrastructure Talk at the Server Track to gain a system-view for an interoperable infrastructure.

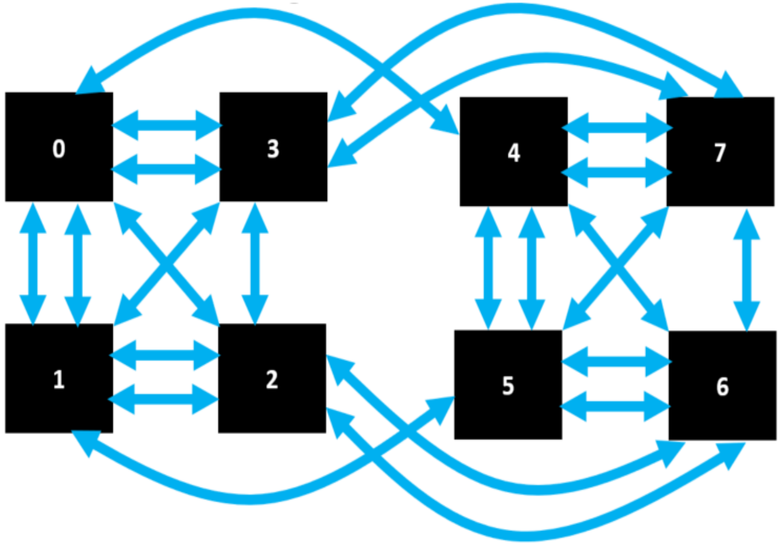
https://2019ocpglobalsummit.sched.com/event/Jikl/ocp-accelerator-module-oam-system-an-open-accelerator-infrastructure-project

# 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 general 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.
- Tiffany Jin is a Mechanical Engineer for data center 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.
- Cheng Chen is a Thermal Engineer for hardware design at Facebook. He leads the thermal design of AI/ML training platforms including Big Basin, and general purpose 2S compute platforms. His studies focus on energy-efficient cooling strategies for high power platforms, and thermal roadmap for AI training modules.
- 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.

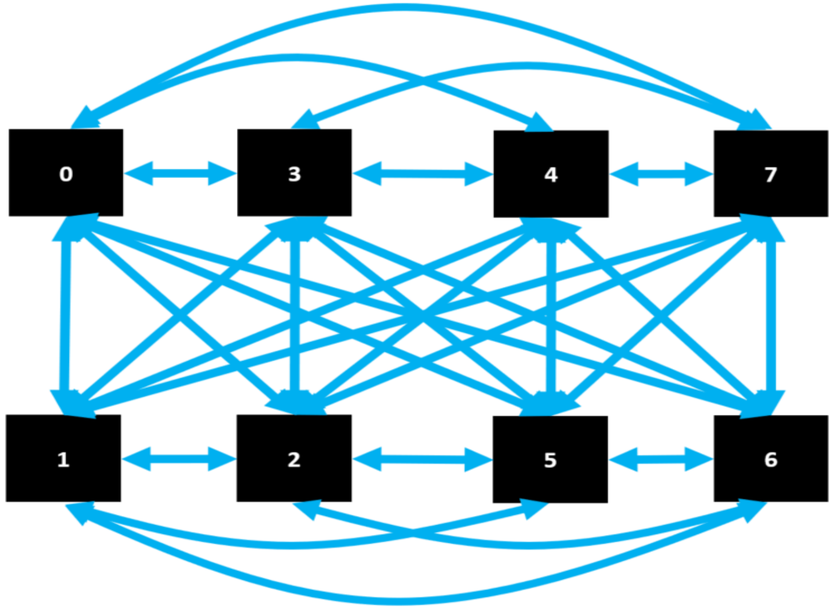

Open. Together.
OCP Global Summit | March 14–15, 2019
OCP
SUMMIT

# Interconnect Topology Examples




8 modules with 6 links per module

Hybrid Mesh Cube




8 modules with 7 links per module

Fully Connected

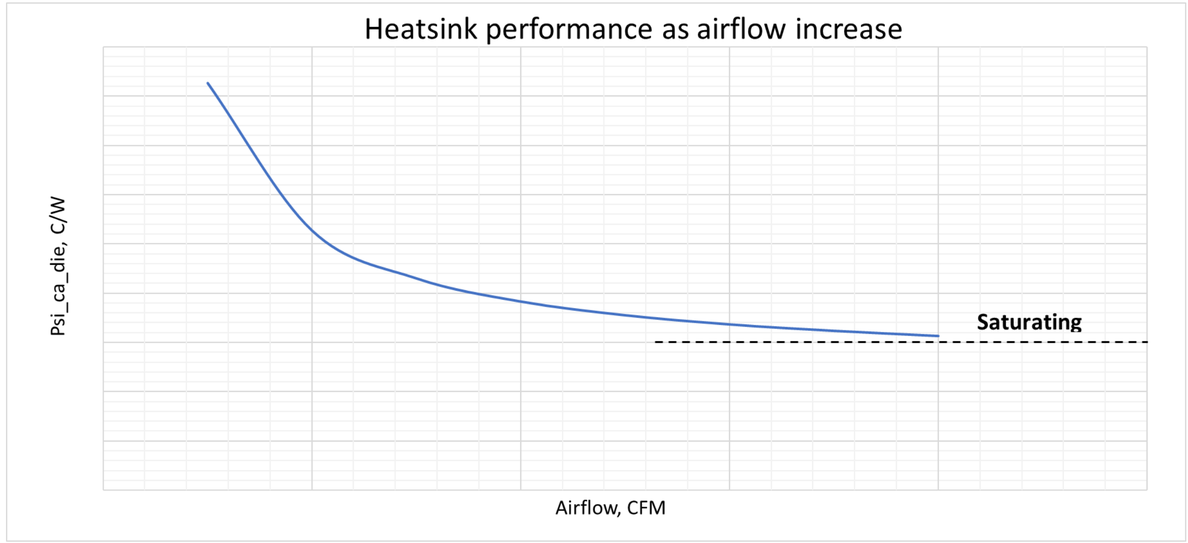
# Thermal Recommendation – Module Height

- Cooling Limit

  Air cooling is capable of supporting module power up to 440W, beyond which advanced cooling is probably needed.

- Module Height

  To support representative liquid cooling solution (open loop), Max height from bottom of module to top of die: 13mm




# Thermal Requirement – Sensor Report

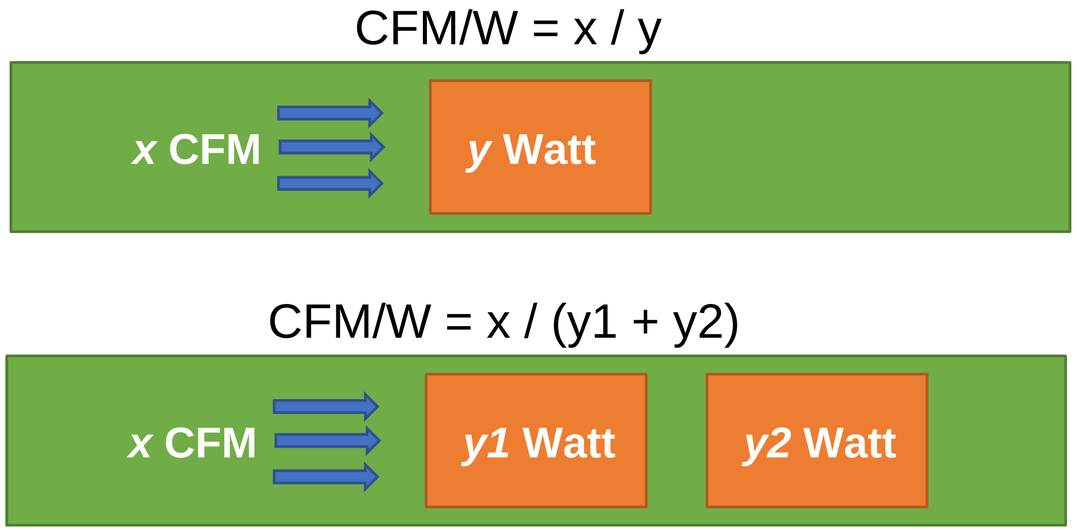
- ASIC and Memory temperature sensor readings will be reported to support fan speed control and hardware or software throttling.

Recommendation: before ASIC or Memory reach their throttling temperatures, the remaining components on module should be able to maintain within their temp limits.

# Thermal Recommendation – Airflow Budget

OAM should be capable of operating with full performance at or below a target CFM/W of 0.145, with ambient temp up to 30C at sea level.

- For a single OAM being shadowed by other components, the calculation uses the module power and airflow through its heatsink.
- For an OAM shadowing other components, the power calculation is the sum of Mezz card and upstream components.



# Thermal Requirement – Module Info

Following info will be provided for each product:

- ASIC & Memory (HBM or DIMM) junction temp limit
- ASIC & Memory junction to case/surface correlations
- Connector temp limit
- ASIC & Memory nominal operation temp range
- Pressure limit on die

# Thermal Recommendation – TIM (for die)

- Minimum Thermal Conductivity: 4 W/m*K
- Maximum Particle Size: TBD
- Operation Pressure range: TBD

# Thermal Recommendation – Reference Heatsink

A reference heatsink will be provided for each product, including:

- Mechanical & Thermal model
- Heatsink performance curves
  - Correlation of Airflow → Thermal resistance & Pressure drop
  - Correlation of Inlet temp & Power level → Airflow requirement → Pressure drop

# Thermal Recommendation – Heatsink installation

- Screw tightening
  - Screw head type: Philips #2
  - Tightening pattern: diagonal
  - Tightening stages: multi stage (TBD)
  - Torque: TBD
- Mounting pressure
  - Min/Max static pressures on die
  - Max dynamic pressure on die

# Thermal Recommendation – FRU Height

- FRU Height
  - For a representative air cooled FRU (module + heatsink), FRU height is: 99.3mm + 13.6mm (handle). A taller heatsink would deliver minimal cooling improvement.
  - Recommended max FRU height to fit within 3OU system: 121mm
  - Recommended max FRU height to fit within 4RU system: 155mm