

OPEN POSSIBILITIES.

Introduction to the OCP HPC SubProject's HPCM (High Performance Computing Module)

Nomenclature : Read "Processor" as CPU and/or Accelerator

SERVER

Re-Inventing HPC Architectures for a “Domain Specific Architecture” Computing World

Allan Cante, CEO, Nallasway

OPEN POSSIBILITIES.



OPEN
COMMUNITY®



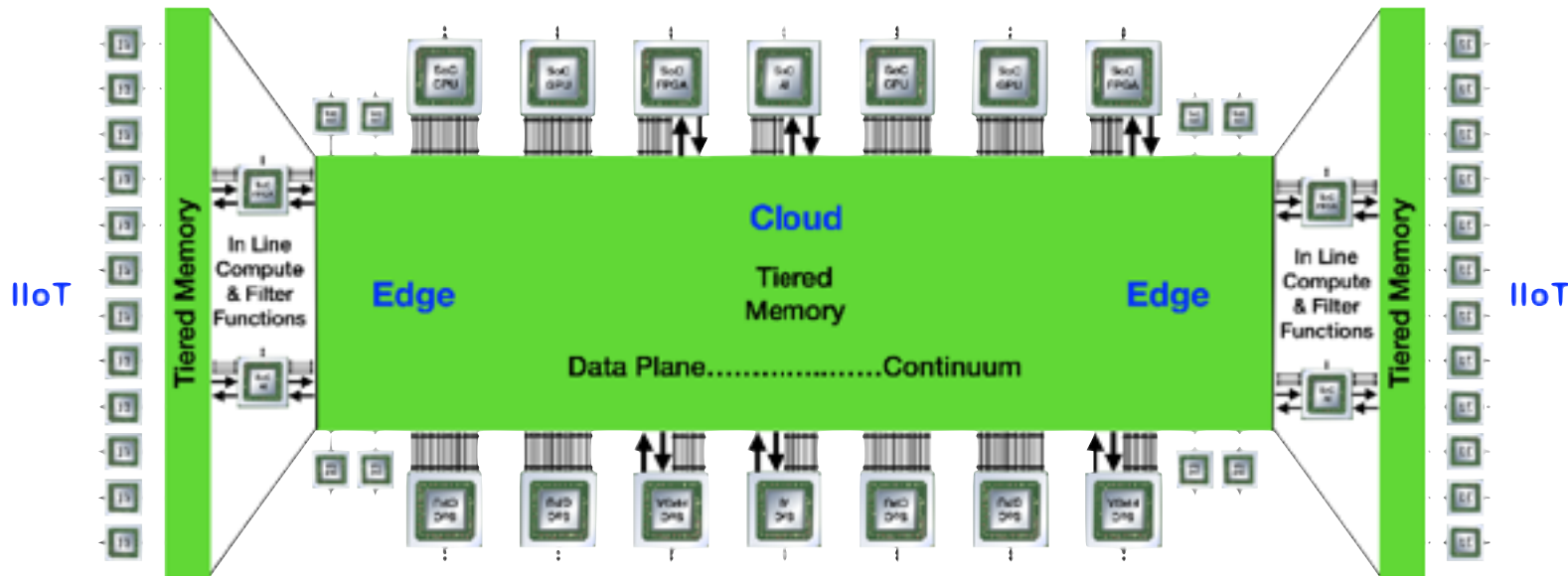
System Level Domain Specific Architecture Template



SERVER



HIGH
PERFORMANCE
COMPUTING



Sourced from Slide 11 of

[OpenCAPL - A Memory Centric Fabric in Data Centric World](#)
[Recording](#)

OPEN POSSIBILITIES.



System Level Domain Specific Architecture Template

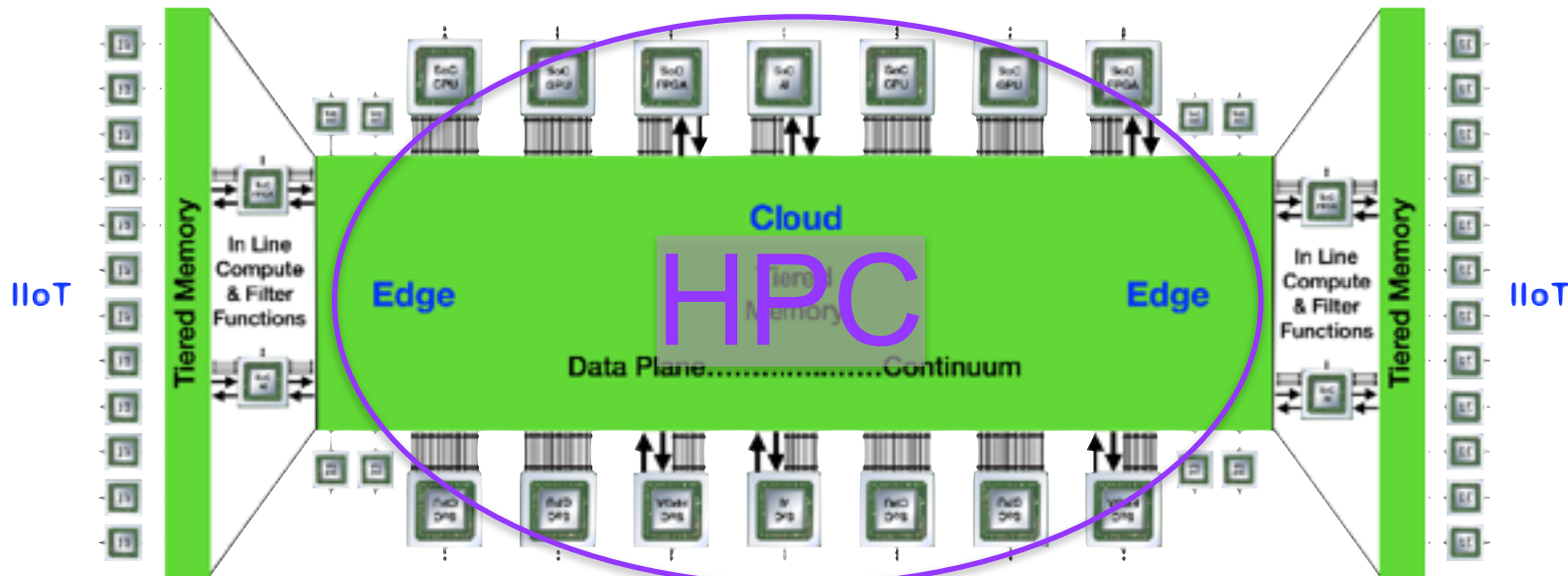


SERVER

- HPC is increasingly Data Bound & Less So Compute Bound



HIGH
PERFORMANCE
COMPUTING



Sourced from Slide 11 of

[OpenCAPL - A Memory Centric Fabric in Data Centric World](#)
[Recording](#)

OPEN POSSIBILITIES.



Heterogeneous w/ blurred Storage/Memory Boundaries

System attributes	ALCF Now	NERSC Now	OLCF Now	NERSC Pre-Exascale	ALCF Pre-Exascale	OLCF Exascale	ALCF Exascale
Name (Planned) Installation	Theta 2016	Cori 2016	Summit 2017-2018	Perlmutter (2020-2021)	Polaris (2021)	Frontier (2021-2022)	Aurora (2022-2023)
System peak	> 15.6 PF	> 30 PF	200 PF	> 120 PF	35 – 45 PF	> 1.5 EF	≥ 1 EF CF sustained
Peak Power (MW)	< 2.1	< 3.7	10	6	< 2	29	≤ 60
Total system memory	647 TB DOR4 + 70 TB HBM + 7.5 TB GPU memory	~1 PB DOR4 + High Bandwidth Memory (HBM) + 1.5 PB persistent memory	2.4 PB DOR4 + 3.4 PB HBM + 7.4 PB persistent memory	182 PB DOR4 + 240 TB HBM	> 250 TB	4.6 PB DOR4 + 4.6 PB HBM2e + 36 PB persistent memory	> 10 PB
Node performance (TF)	2.7 TF (KNL node) and 166.4 TF (GPU node)	> 3	43	> 70 (GPU) > 4 (CPU)	> 70 TF	TBD	> 130
Node processors	Intel Xeon Phi 7320 (64-core CPUs (KNL) and GPU nodes with 8 NVIDIA A100 GPUs coupled with 2 AMD EPYC 64-core CPUs	Intel Knights Landing many core CPUs Intel Haswell CPU in data partition	2 IBM Power9 CPUs + 8 Nvidia Volta GPUs	CPU only nodes: AMD EPYC Milan CPUs; CPU-GPU nodes: AMD EPYC Milan with NVIDIA A100 GPUs	1 CPU; 4 GPUs	1 HPC and AI optimized AMD EPYC GPU and 4 AMD Radeon Instinct GPUs	2 Intel Xeon Sapphire Rapids and 6 Xe Ponte Vecchio GPUs
System size (nodes)	4,392 KNL nodes and 24 DGX-A100 nodes	9,500 nodes 1,900 nodes in data partition	4508 nodes	> 1,500 (GPU) > 3,000 (CPU)	> 500	> 9,000 nodes	> 9,000 nodes
CPU-GPU Interconnect	NVLink on GPU nodes	N/A	NVLink Coherent memory across node	PCIe		AMD Infinity Fabric Coherent memory across the node	Unified memory architecture, RAMSO
Node-to-node Interconnect	Aries (KNL nodes) and HDR200 (GPU nodes)	Aries	Dual Rail EDR-88	HPE Slingshot NIC	HPE Slingshot NIC	HPE Slingshot	HPE Slingshot
File System	200 PB, 1.3 TB/s Lustre 10 PB, 210 GB/s Lustre	28 PB, 744 GB/s Lustre	250 PB, 2.5 TB/s GFTS	35 PB All Flash, Lustre	N/A	666 PB + 10 PB Flash performance tier, Lustre	≥ 230 PB, ≥ 25 TB/s DAOS



SERVER



HIGH
PERFORMANCE
COMPUTING



Office of
Science

Frontier: <https://www.olcf.rnl.gov/frontier/>
Aurora: <https://www.alcf.aal.gov/aurora>

ASCR Computing Upgrades At-a-Glance
November 24, 2020

OPEN POSSIBILITIES.



Today's HPC Compute & Storage Challenge



SERVER

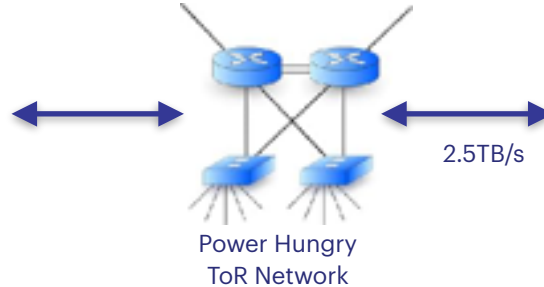
- CORAL Summit HPC Machine example
 - 18 Minutes to Load 2.8PB Memory from Filesystem once!
 - 1.2 Days to Push ALL 250PB Filesystem thru Compute Racks!
- Need to Bring Compute, Memory and Storage much closer



HIGH
PERFORMANCE
COMPUTING



Summit HPC Compute Racks
2.4 PetaBytes of DDR4
0.4 Petabytes of HBM2
7.4 PBytes of Persistent Memory*



Summit HPC GPFS File System
250 PetaBytes of Storage

OPEN POSSIBILITIES.

*Persistent Memory only used for Checkpoint Restarts

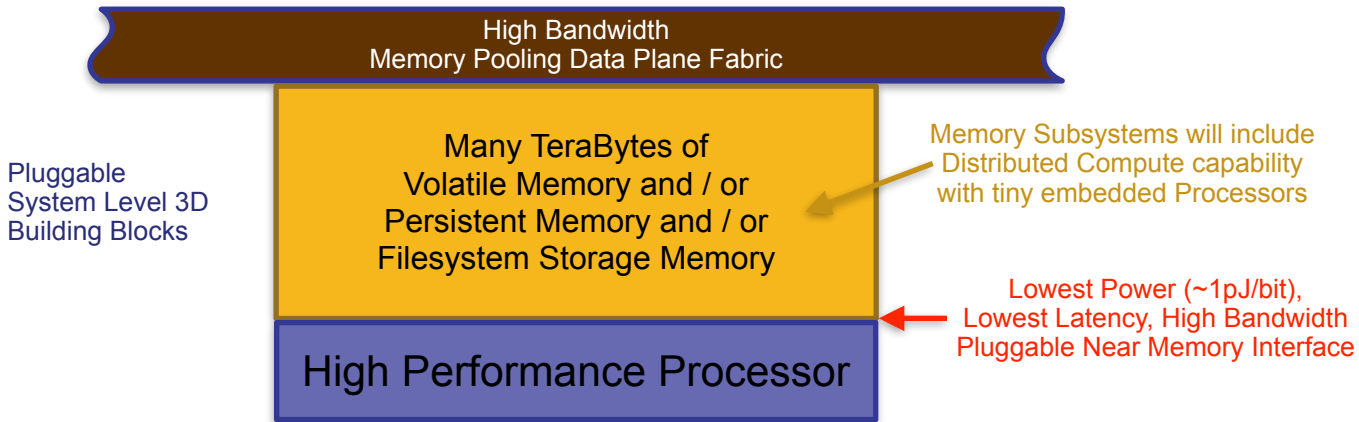


Data Centric HPC Solution - Abstract View



SERVER

- Tightly Couple Compute with ALL/ANY Memory Types
- Efficiently share Processors Near Memory with Other Processors



HIGH
PERFORMANCE
COMPUTING

OPEN POSSIBILITIES.



If Tesla can “Re-Invent” then why not OCP?



SERVER



HIGH
PERFORMANCE
COMPUTING



Image sourced from [Screenrant.com](https://www.screenrant.com)

OPEN POSSIBILITIES.



We need to Innovate across Silos!



Image sourced from [Stage 2 Planning Partners](#)

OPEN POSSIBILITIES.



SERVER



HIGH
PERFORMANCE
COMPUTING



Disaggregated Racks to Hyper-converged Chiplets



Software
Composable

Power Ignored
Rack Interconnect
>20pJ/bit

Poor Latency

Rack Volume
>53K Cubic Inches

Baseline Physical
Composability

Power Baseline
Node Interconnect
5-10pJ/bit

Baseline Latency

Node Volume >800
Cubic Inches

Expensive Physical
composability

Power Optimized
Chiplet Interconnect
<1pJ/bit

Optimal Latency

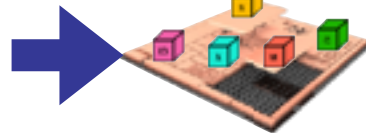
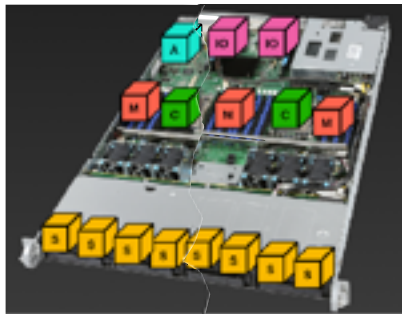
SIP Volume <1
Cubic Inch



SERVER



HIGH
PERFORMANCE
COMPUTING



OPEN POSSIBILITIES.



Disaggregated Racks to Hyper-converged Chiplets



Disaggregated Racks to Hyper-converged Chiplets



HPCM brings the two Together



Software
Composable

Power Ignored
Rack Interconnect
>20pJ/bit

Poor Latency

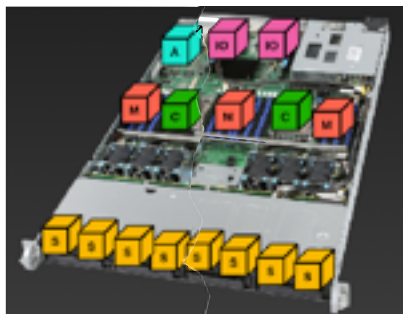
Rack Volume
>53K Cubic Inches

Baseline Physical
Composability

Power Baseline
Node Interconnect
5-10pJ/bit

Baseline Latency

Node Volume >800
Cubic Inches

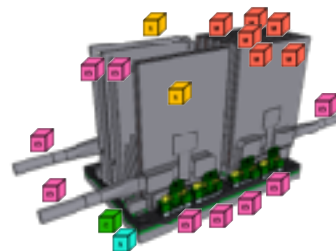


Software & Physical
Composability

Power Optimized
Flexible Chiplet
Interconnect 1-2pJ/bit

Optimal Latency

Module Volume
<150 Cubic Inches



Expensive Physical
composability

Power Optimized
Chiplet Interconnect
<1pJ/bit

Optimal Latency

SIP Volume <1
Cubic Inch



SERVER



HIGH
PERFORMANCE
COMPUTING

OPEN POSSIBILITIES.

OCF HPC Module, HPCM,
Populated with E3.S, NIC-3.0, & Cable IO



Overview of OCP HPC SubProject's HPCM (High Performance Computing Module)

Allan Cattle, CEO, Nallasway

OPEN POSSIBILITIES.



High Performance Computing Module, HPCM

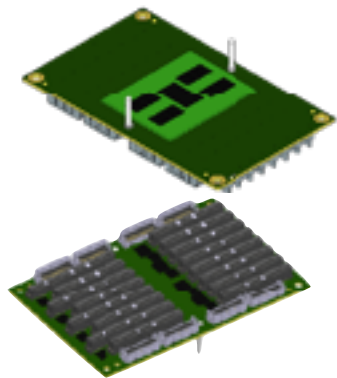


SERVER

- Modular, Flexible and Composable Module - Protocol Agnostic!
 - Memory, Storage & IO interchangeable depending on Application Need
 - Processor must use HBM or have Serially Attached Memory

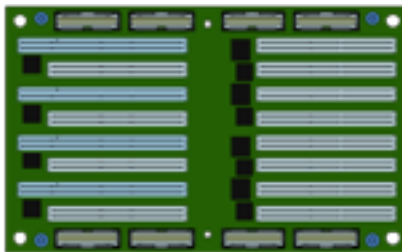


HIGH
PERFORMANCE
COMPUTING

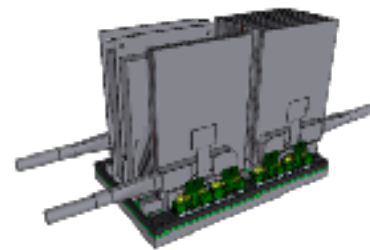


OCP HPCM

HPCM Standard
could Support
Today's Processors
e.g.
NVIDIA Ampere
Google TPU
IBM POWER10
Xilinx FPGAs
Intel FPGAs
Graphcore IPU
PCIe Switches
Ethernet Switches



HPCM Interconnect for all Processor / Switch types
16x EDSFF 4C/4C+ + 8x Nearstack x8 Connectors
Total of 320x Transceivers



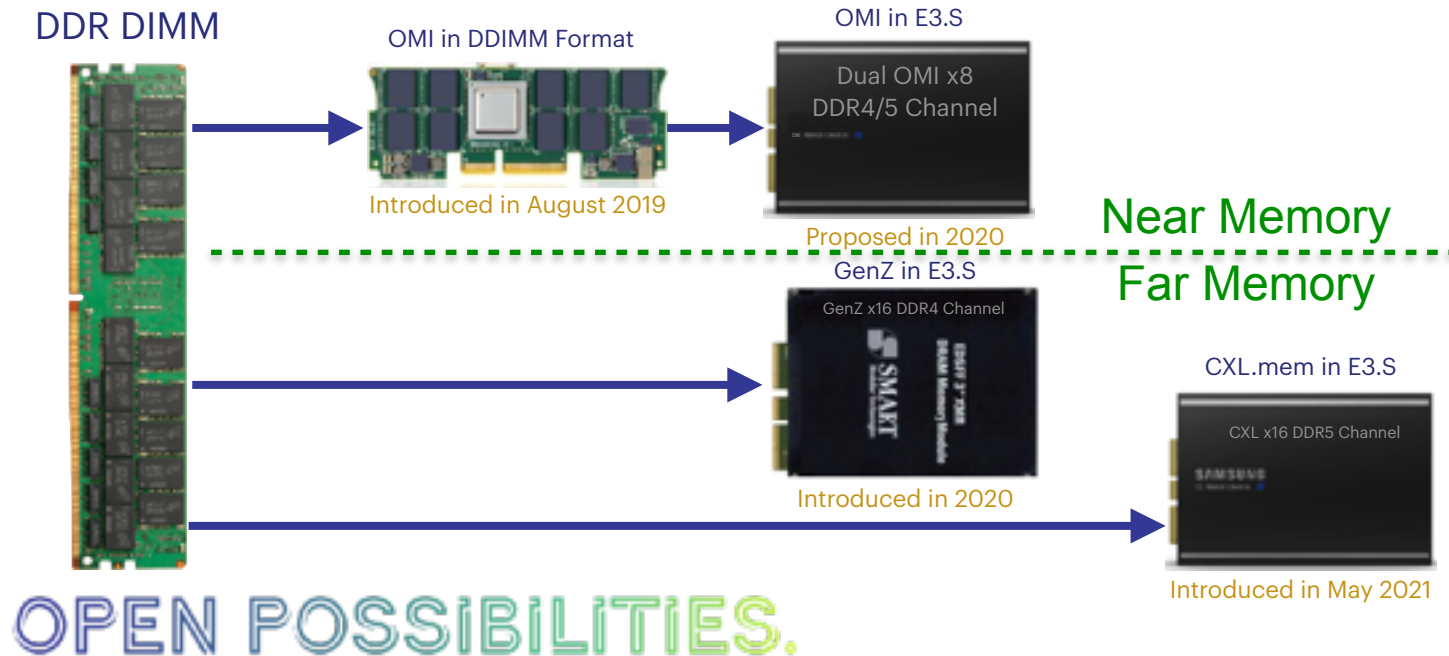
Example HPCM Bottom
View Populated with
8x E3.S Modules,
2x OCP NIC 3.0 Modules,
4x TA1002 4C Cables &
8x Nearstack x8 Cables

OPEN POSSIBILITIES.



Memory IO is finally going Serial!

- Making Memory Composable with EDSFF E3.S like Storage & IO



SERVER



HIGH
PERFORMANCE
COMPUTING



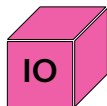
Modular Building Blocks Available Today



SERVER

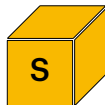
- Network, Memory, & IO use [Common EDSFF Interconnect](#)

OCP - NIC 3.0



Typically < 100W

SNIA - E1.S & E3.S



E1.S (25mm)

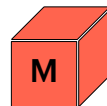


E3.S

Jedec - DDIMM



GenZ in E3.S



CXL.mem in E3.S

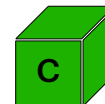


OMI in E3.S



200W to 1KW

OCP - OAM



HIGH
PERFORMANCE
COMPUTING

OPEN POSSIBILITIES.



Dense Modularity = Power Saving Opportunity

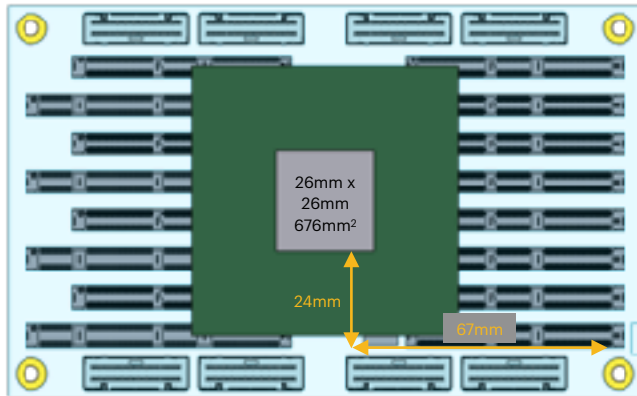
- Processor Die Bump to E3.S ASIC <5 Inches - Manhattan Distance
 - Opportunity to reduce PHY Channel to 5-10dB, 1-2pJ/bit
- Enabling Low Power



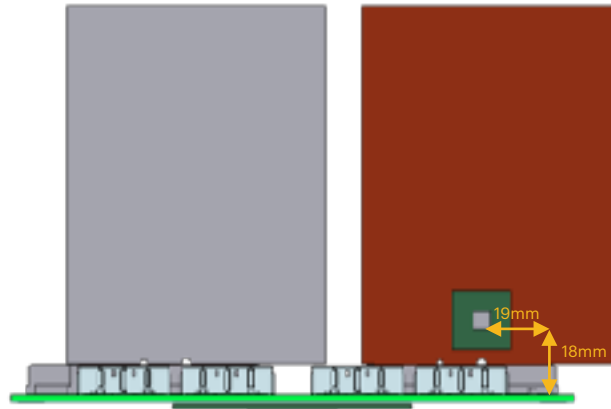
SERVER



HIGH
PERFORMANCE
COMPUTING



OPEN POSSIBILITIES.



Installing 8 HPCMs in OAI Chassis

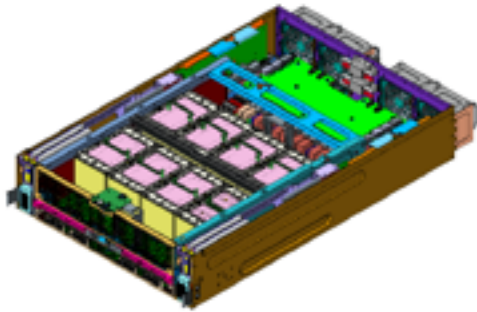


SERVER



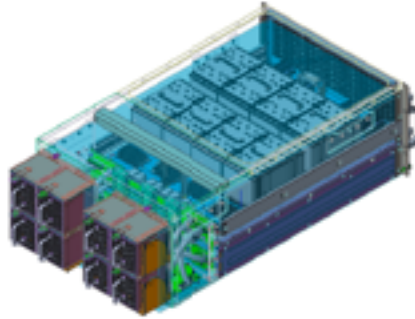
HIGH
PERFORMANCE
COMPUTING

Inspur 21" Co-Planar system



- 21 inch 30U, 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 x'6 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

OPEN POSSIBILITIES.



Re-Architect - Start with a Cold Plate

For High Wattage HPCM Modules

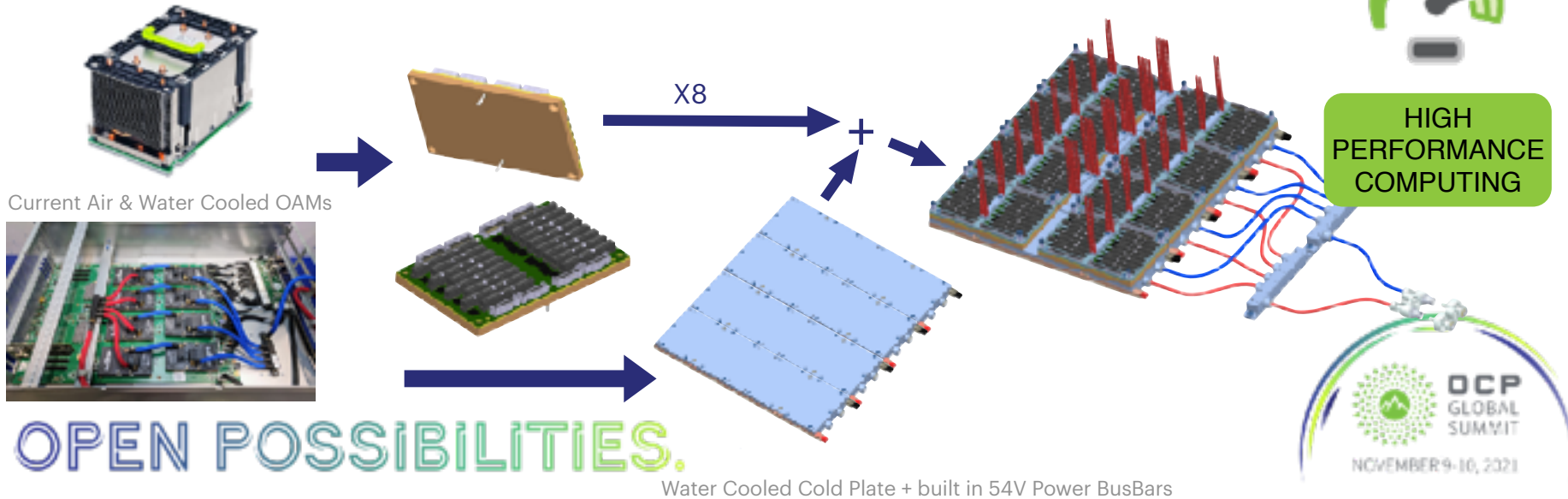


SERVER

- Capillary Heatspreader on module to dissipate die heat across module surface area
- Heatsinks are largest Mass, so make them the structure of the assembly
 - Integrate liquid cooling into the main cold plate

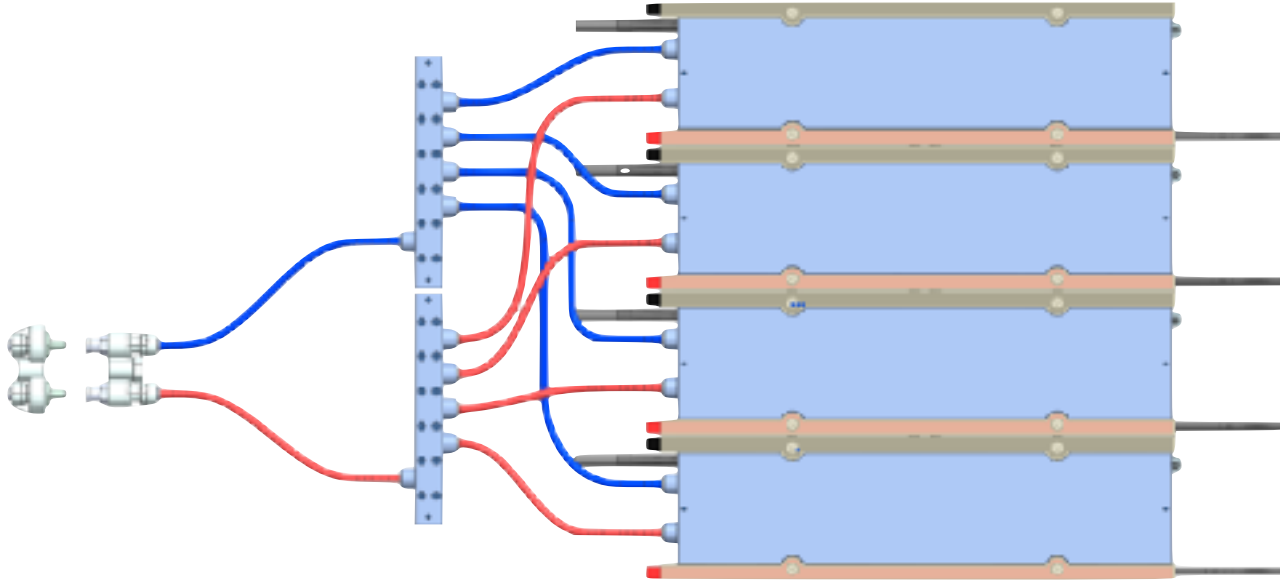


HIGH
PERFORMANCE
COMPUTING



Cold Plate from Backside

- 54V Power Bus Bars shown - Powering HPCMs



OPEN POSSIBILITIES.



SERVER



HIGH
PERFORMANCE
COMPUTING



Add Topology Cabling - No Retimers

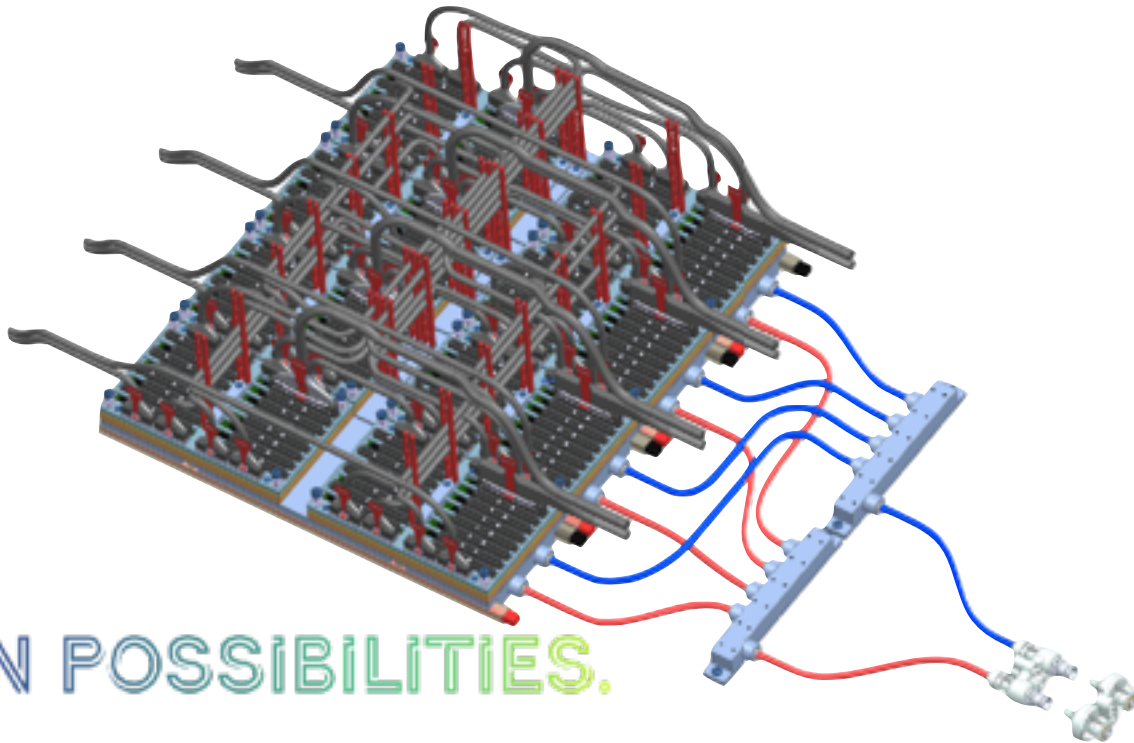
- Fully Connected Topology + Connections to HIB & QDD IO



SERVER



HIGH
PERFORMANCE
COMPUTING

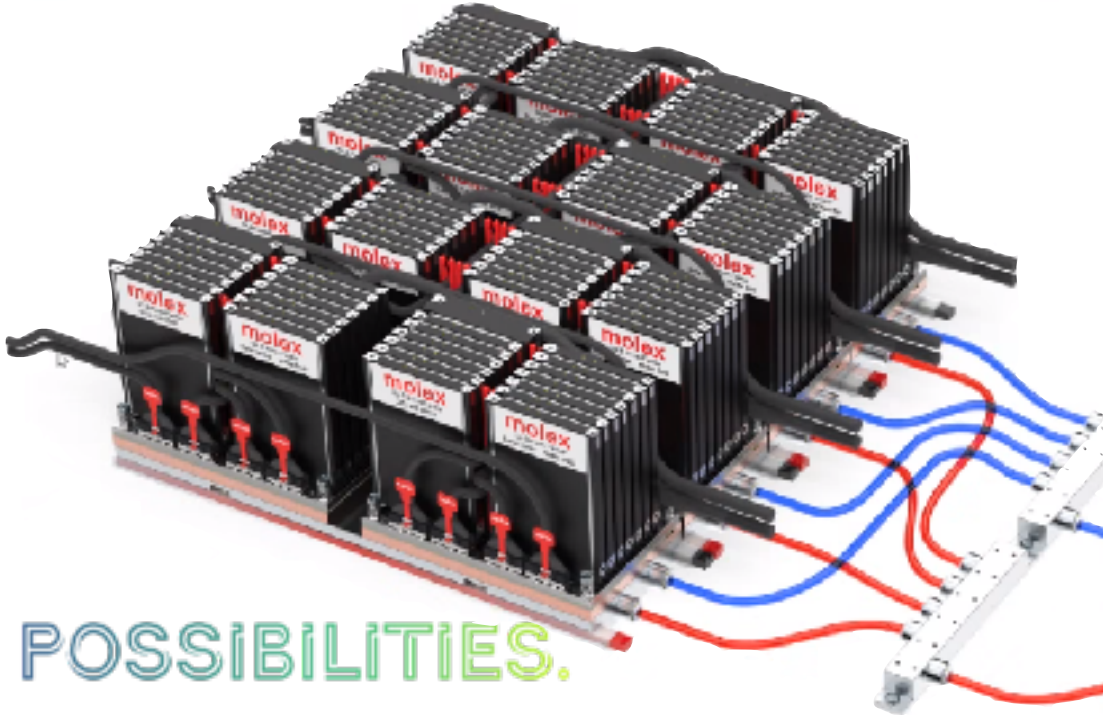


OPEN POSSIBILITIES.



Add E3.S and NIC 3.0 Modules

- Pluggable into OCP OAI Chassis



OPEN POSSIBILITIES.



SERVER



HIGH
PERFORMANCE
COMPUTING



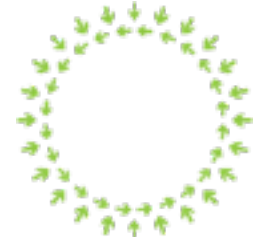
How HPCM provides Efficient & Flexible Interconnect to support increased Fabric Speeds

Allan Cattle, CEO, Nallasway

Tang Junyan, Mahesh Bohra, Dan Dreps, IBM

Bob Dillman & Gus Panella, Molex

OPEN POSSIBILITIES.

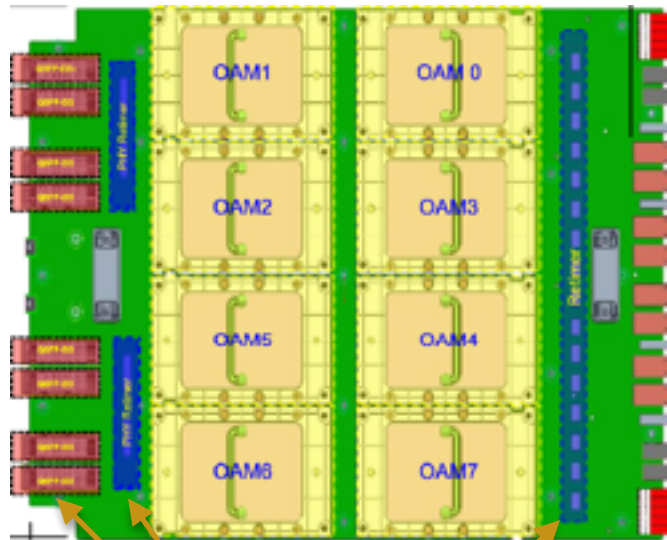


OPEN
COMMUNITY®



Challenges of Compute Interconnect

- Growing demand for Faster and wider Interconnect
 - IO increasing % of Total Power
 - More IO = More Complex PCBs
 - PCB Losses increase
 - Shorter traces
- Retimers increasingly required
 - Add latency, Power, cost, & consume real estate
 - Zero return on investment!



Retimers and Active Cables increasingly required as Fabric Speeds increase



SERVER



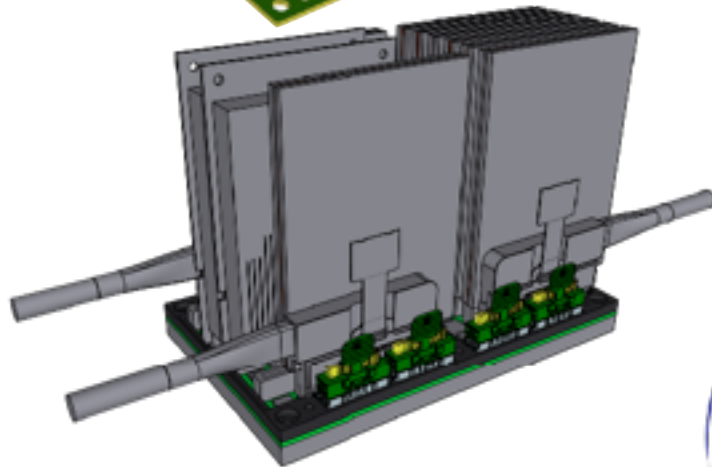
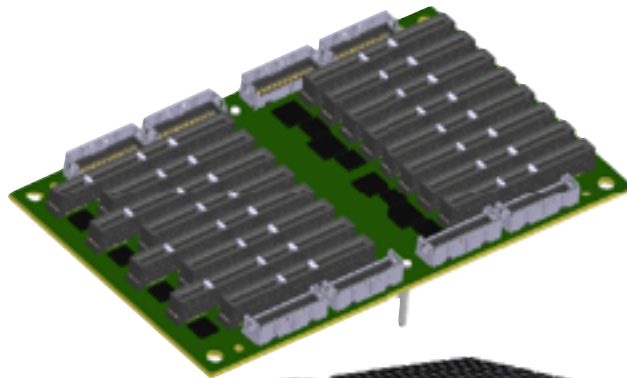
HIGH
PERFORMANCE
COMPUTING

OPEN POSSIBILITIES.



HPCM Interconnect Innovation

- HPCM Increases System Level Density
 - 3D Construction brings Compute, Media & IO Closer together
- Leverage TA-1002 Interconnect to support Media & IO Modules as well as Direct IO
- Leverage Nearstack-PCIe for motherboard-less cabled Fabric Topology Interconnect



SERVER



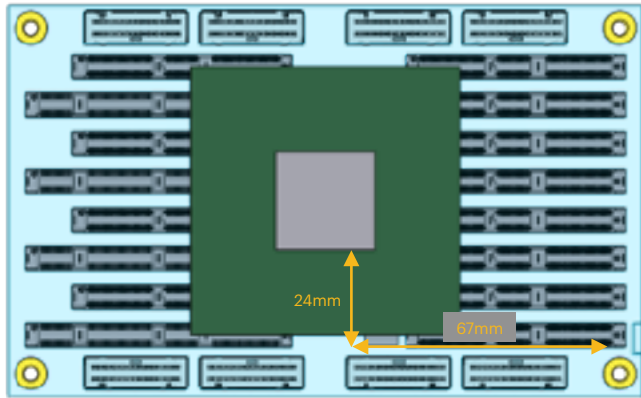
HIGH
PERFORMANCE
COMPUTING

OPEN POSSIBILITIES.

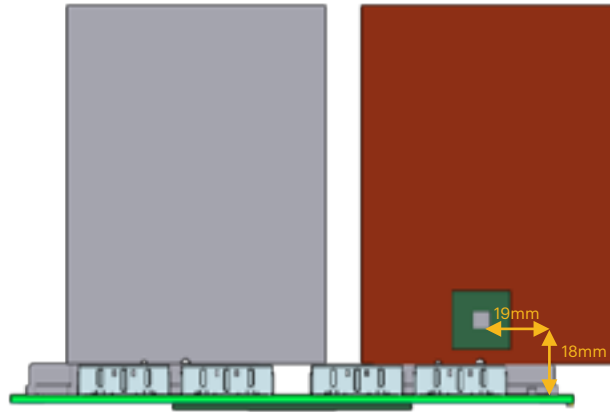


HPCM Processor to Media/IO Module

- Processor to Media / IO Module Manhattan Distance
 - 128mm (<5 Inches) worst case
 - ~10dB Channel with opportunity to reduce IO Power
- Possible further improvement using HPCM as Processor Substrate



OPEN POSSIBILITIES.



SERVER



HIGH
PERFORMANCE
COMPUTING



HPCM Processor to Module Interconnect

With Packaged Processor and Controller Chips



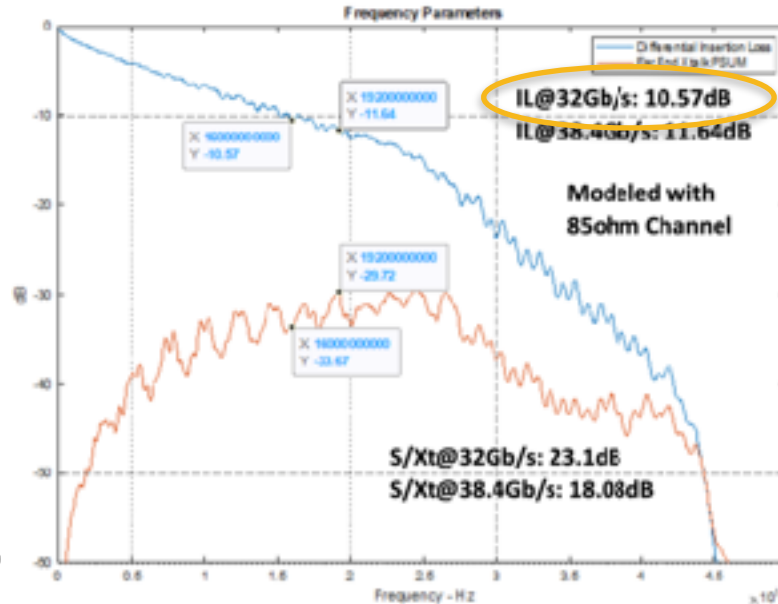
SERVER



HIGH
PERFORMANCE
COMPUTING

Channel based on OMI Module:

- GL102 pkg wiring (30mm)
 - Module Via S12
 - 24mm Pin Area Wiring
 - Meg6 Open Area (67mm)
 - DIMM PCB Via S12
 - DIMM Conn (C2)
 - Meg6 Open Area (37mm)
 - E3.S Controller Package
- (Nominal PCB and PKG corner)



OPEN POSSIBILITIES.

Courtesy of IBM



Insertion Loss allocation Table - Conservative

With Packaged Processor and Controller Chips

Channel Section	Loss @ 32Gb/s	Comments
GL102 package wiring (30mm)	2.8dB	
Module Via S12	1dB	Assume 1.6mm via length & 15mil back drilled stub
Meg6 - 24mm Pin Area Wiring	1.2dB	Conservative assumption with 30mm package wiring & 24mm PCB zig-zag wiring under package
Meg6 - Open Area PCB Trace (67mm)	2.6dB	
DIMM PCB Via S12	0.9dB	Assume 1.6mm via length & 15mil back drilled stub
DIMM Conn (C2)	1dB	
Meg6 - DIMM Open Area (37mm)	1.4dB	
E3.S Controller Package	0.4dB	
Total Channel	11.3dB	Measured channel difference due to impedance discrepancies & behavior

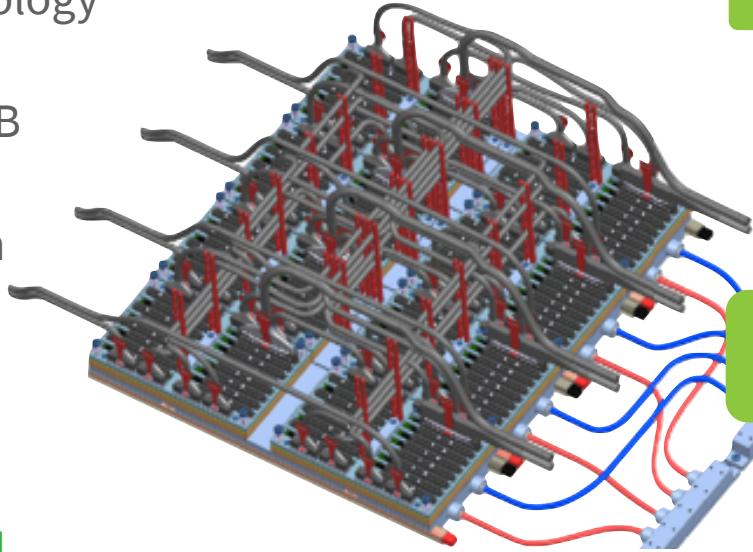
Insertion Loss allocation Table - Conservative

Derived with Bare Die Processor and Controller Chips

Channel Section	Loss @ 32Gb/s	Comments
Module Via S12	1dB	Assume 1.6mm via length & 15mil back drilled stub
Meg6 - 24mm Open Area Wiring	1dB	
Meg6 - Open Area PCB Trace (67mm)	2.6dB	
DIMM PCB Via S12	0.9dB	Assume 1.6mm via length & 15mil back drilled stub
DIMM Conn (C2)	1dB	
Meg6 - DIMM Open Area (37mm)	1.4dB	
Total Channel	7.9dB	Empirical estimate only

OAI Node Fully Connected Topology

- Module to Module Interconnect Topology
- Assume 112G PAM4 Fabric Speed
- HPCM Loss, w/ packaged Proc ~8.5dB
- Longest Cable - 12 Inches ~ 5.8dB
 - 34awg Cable Loss = 0.37 dB/inch
 - Connector Loss = 0.7dB/ea (typ)
- Total Channel Loss Estimate
 - Proc to Proc = $8.5 + 5.8 + 8.5$
 - ~22.8dB
 - 7.2dB Spare on a 30dB 112G channel



SERVER



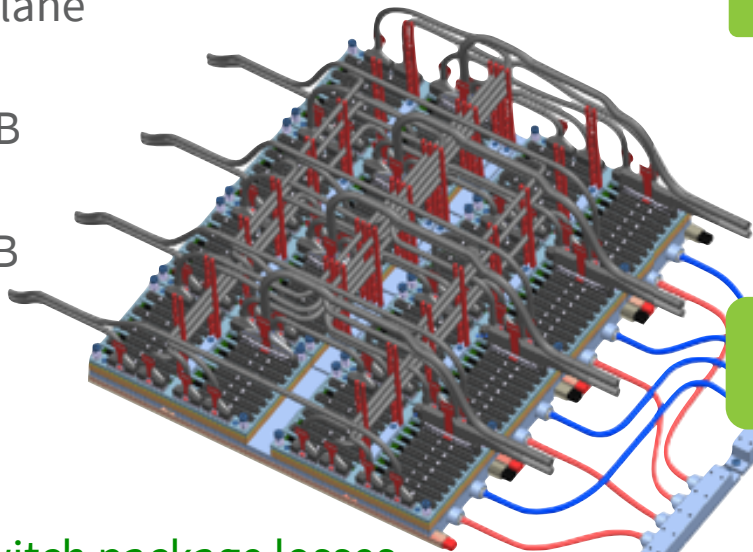
HIGH
PERFORMANCE
COMPUTING

OPEN POSSIBILITIES.



OAI Node HPCM to HIB Interconnect

- HPCM TA-1002 to HIB Examax Backplane
- 64G PAM4 CXL/PCIe G6 Fabric Speed
- HPCM Loss, w/ packaged Proc ~8.5dB
- Longest Cable - ~12 Inches
 - TA-1002 Loss + PCB fingers ~ 2 dB
 - 34awg Cable loss = 0.37 dB/in
 - Backplane Loss ~ 0.7 dB
- Total Proc to HIB Backplane Loss
 - $\sim 8.5 + 2 + 4.4 + 0.7 = 15.6\text{dB}$
 - 14.4dB spare for HIB PCB and Switch package losses
 - Retimerless compared to UBB implementations



SERVER



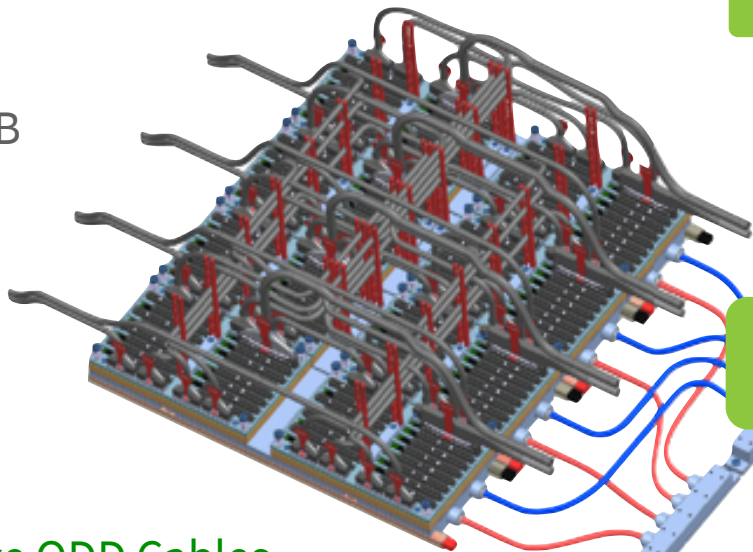
HIGH
PERFORMANCE
COMPUTING

OPEN POSSIBILITIES.



OAI Node HPCM to QDD Interconnect

- HPCM Nearstack to QDD Fabric IO
- Assume 112G PAM4 Fabric Speed
- HPCM Loss, w/ packaged Proc ~8.5dB
- Longest Cable - ~17 Inches
 - Connector Loss ~ 0.7 dB
 - 34awg Cable Loss ~ 0.37 dB/in
 - QDD Loss ~ 2.5 dB
 - Total Proc to QDD Loss
 - $\sim 8.5 + 0.7 + 6.3 + 2.5 = 18\text{dB}$
 - 12dB spare - May Support Passive QDD Cables
 - Retimerless compared to UBB implementations



SERVER



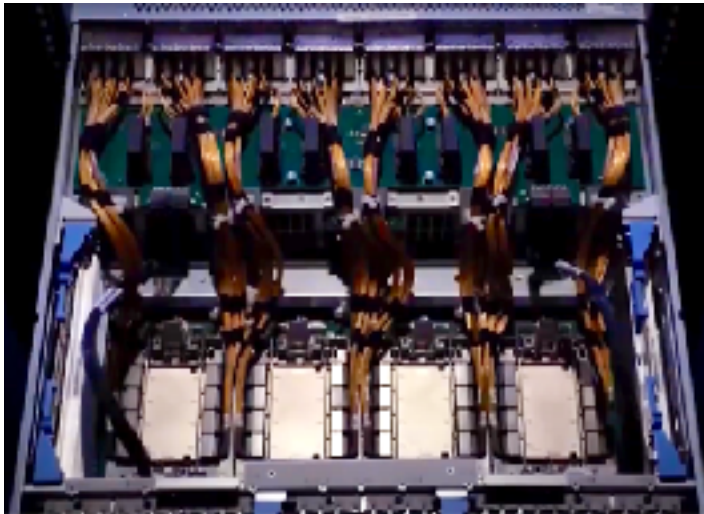
HIGH
PERFORMANCE
COMPUTING

OPEN POSSIBILITIES.



Cabled Solutions are reliable

- IBM's High Reliability E1080 Server



OPEN POSSIBILITIES.



SERVER



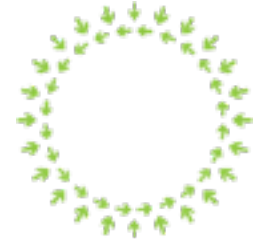
HIGH
PERFORMANCE
COMPUTING



How HPCM's Thermal Management Cold Plate solution turns traditional approaches on their head

Chris Chapman, Boyd Corporation
Bob Dillman, Molex
Allan Cante, Nallasway

OPEN POSSIBILITIES.

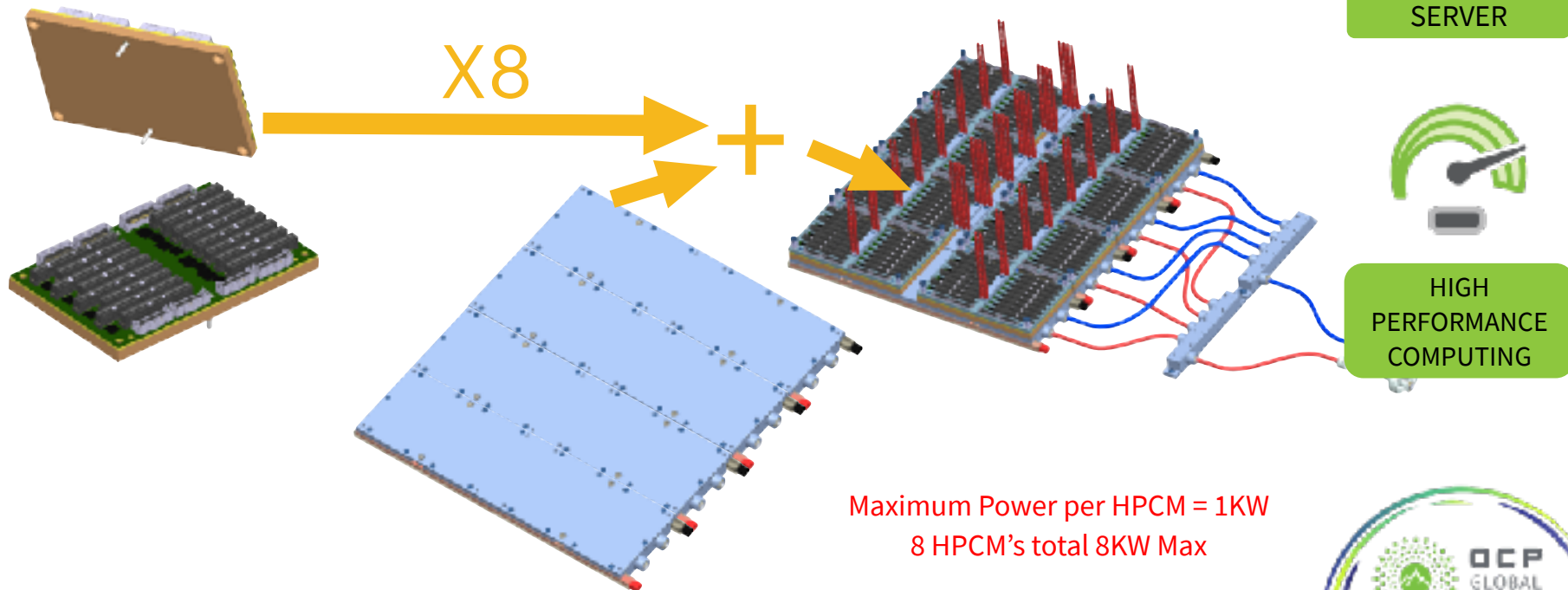


OPEN
COMMUNITY®



HPCM proposed Thermal Solution

Cooling 8x HPCM Module Processors



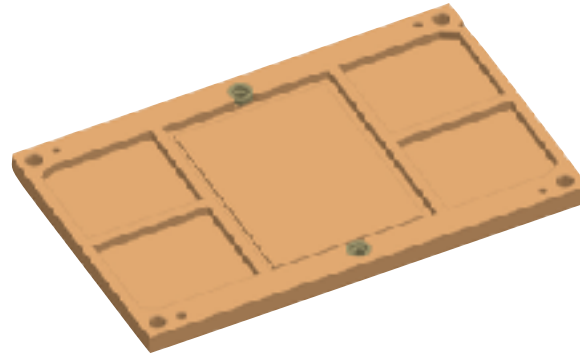
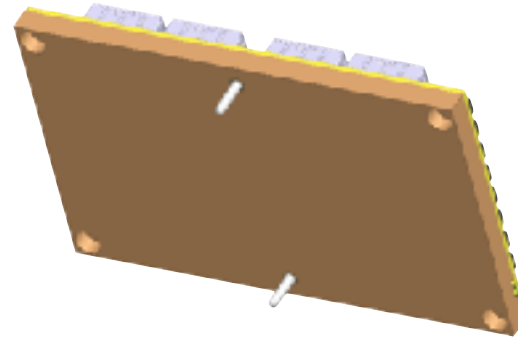
Maximum Power per HPCM = 1KW
8 HPCM's total 8KW Max

OPEN POSSIBILITIES.



HPCM proposed Thermal Solution

- Thermal Heat Spreader
- Required to Normalize Different HPCM Modules for mating to the main cold plate
- Cavities in Heat Spreaders required for surrounding components, primarily PSUs
- Necessitates 2 Thermal Interfaces
 - Silicon to Heat Spreader
 - Heat Spreader to Cold Plate



SERVER



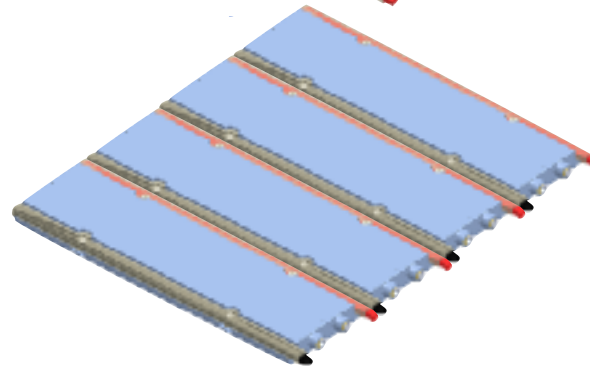
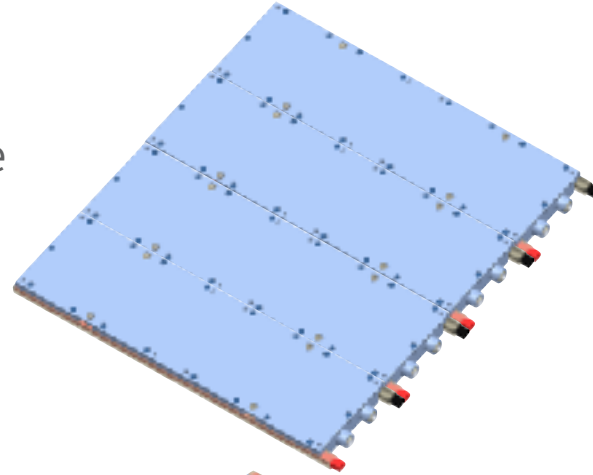
HIGH
PERFORMANCE
COMPUTING

OPEN POSSIBILITIES.



HPCM proposed Thermal Solution

- Water Cooled Cold Plate
- Provides HPCM Mechanical Infrastructure
- Cold water to each HPCM site



SERVER



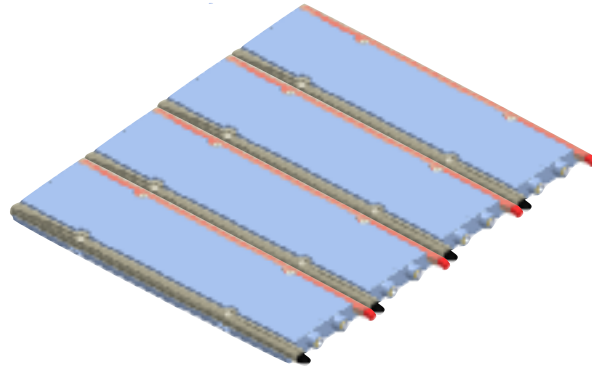
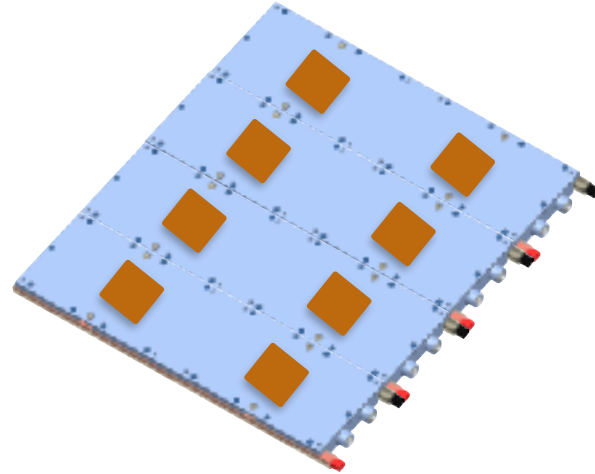
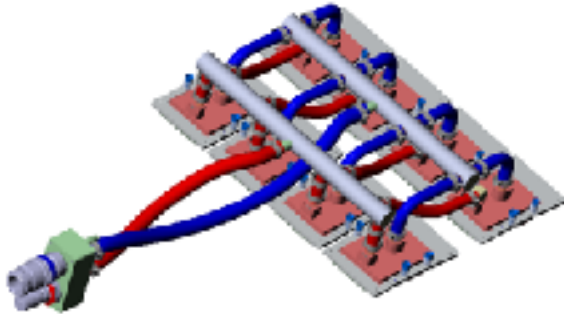
HIGH
PERFORMANCE
COMPUTING

OPEN POSSIBILITIES.



Cold Plate Feasibility

- A single cold plate concept that delivers power to OAM modules utilizing 8 meso-channel “cooling cores” should perform similarly to a cooling loop array if each of the 8 OAM interfaces are independently fastened to the cold plate



SERVER



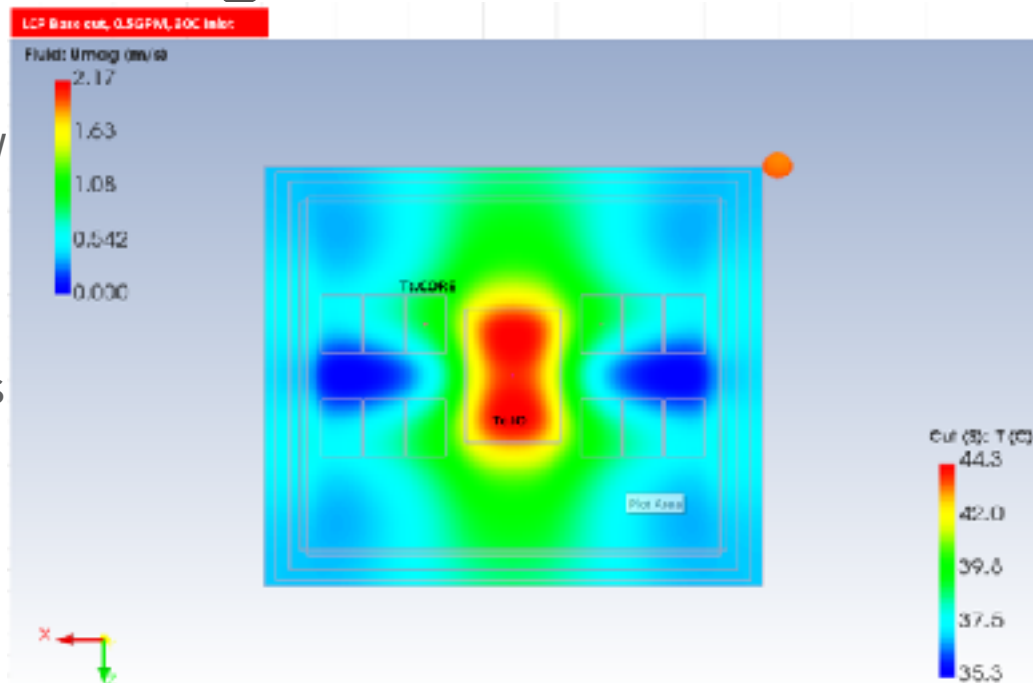
HIGH
PERFORMANCE
COMPUTING

OPEN POSSIBILITIES.



Evaluate “Cooling Core”

- Initial CFD analyzed the new form factor required
- Similar performance was obtained compared to a traditional OAM module cold plate



SERVER



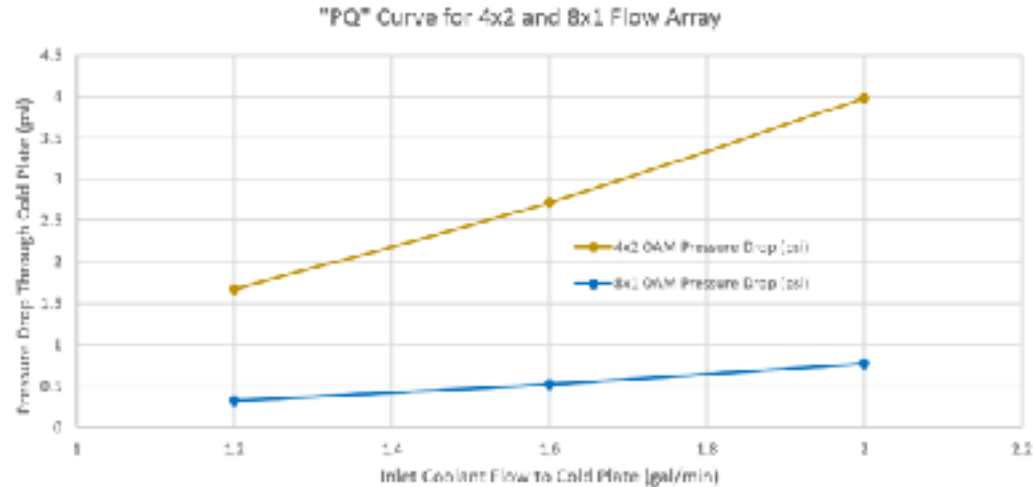
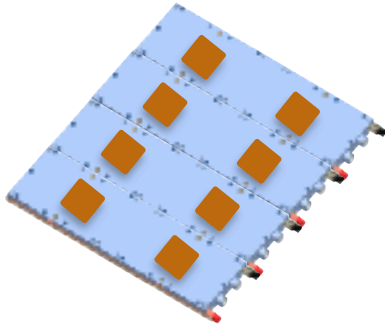
HIGH
PERFORMANCE
COMPUTING

OPEN POSSIBILITIES.



4x2 and 8x1 Flow Network

- Two flow network models were developed for the cold plate assembly
- The all parallel 8x1 array shows the lower pressure drop as shown in the 'PQ' curve



OPEN POSSIBILITIES.



SERVER

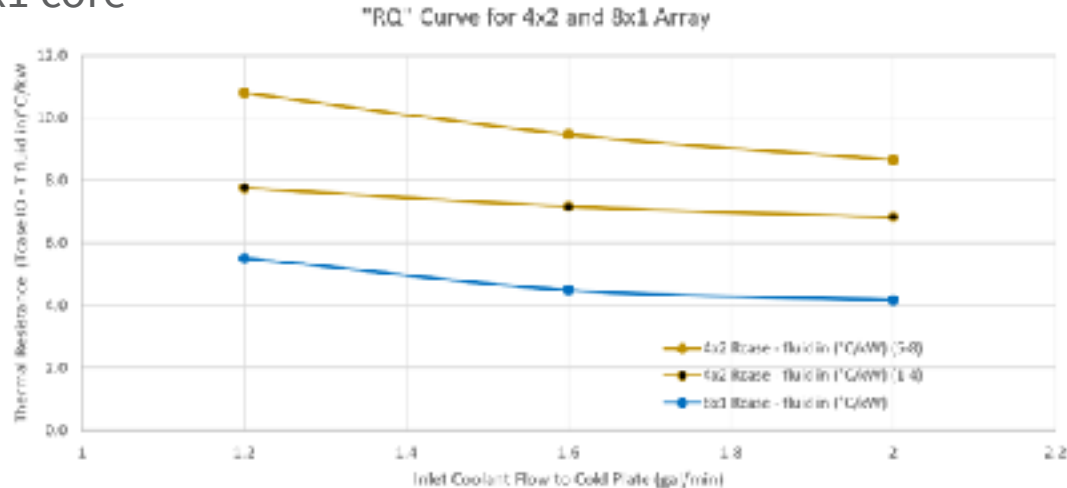
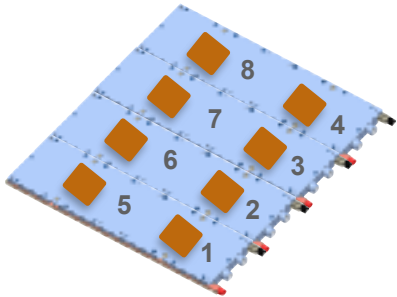


HIGH
PERFORMANCE
COMPUTING



Cold Plate Thermal Performance

- The thermal resistance “RQ” curves are shown and the 4x2 array is split into two curves; one for parallel cores 1-4 and another for 5-8 which are in series in order
- The 8x1 resistance is lower however the cores 1-4 in the 4x2 will run cooler than any 8x1 core



SERVER



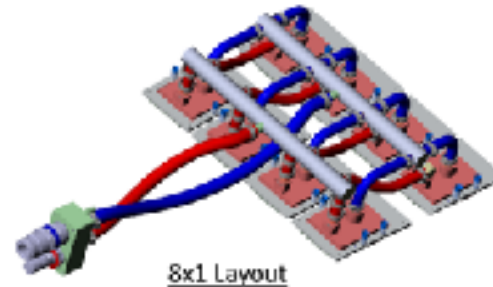
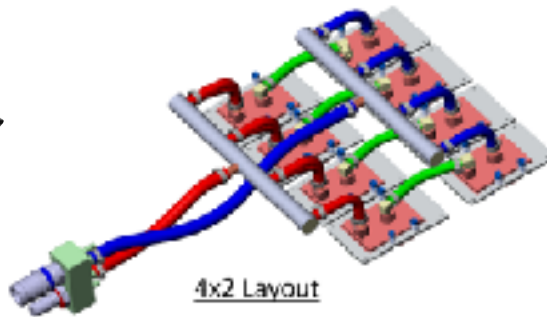
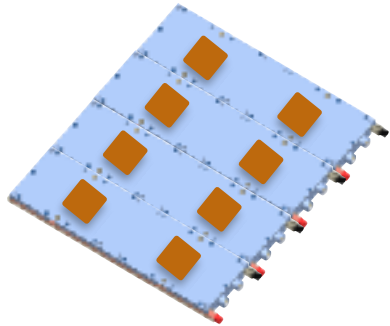
HIGH
PERFORMANCE
COMPUTING

OPEN POSSIBILITIES.



Summary

- Initial study indicates that a cold plate with meso-channel cooling cores will achieve the necessary cooling required as compared to conventional cooling loops
- Further study is recommended as additional electro-mechanical and packaging features can be incorporated into the cold plate as we now understand the keep out area necessary for cooling



SERVER



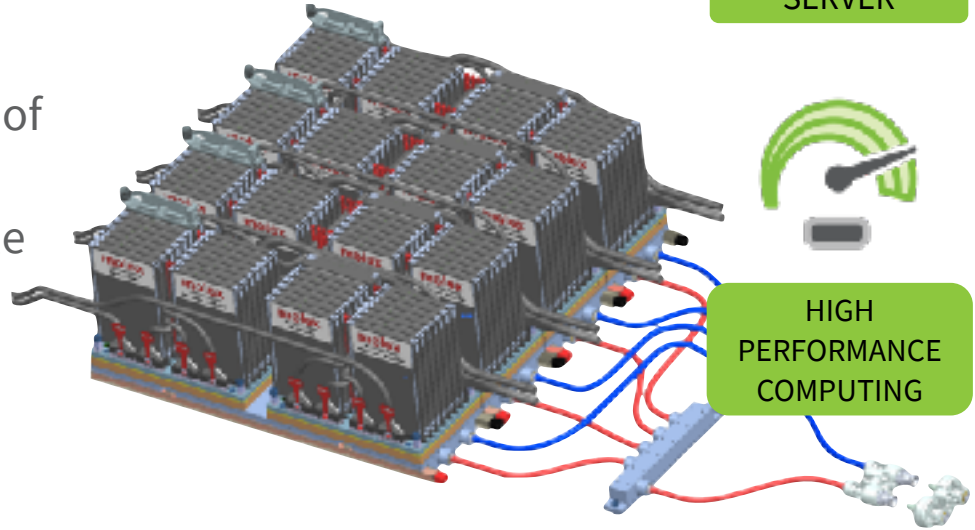
HIGH
PERFORMANCE
COMPUTING

OPEN POSSIBILITIES.



Air Cooling 128x E3.S Modules

- Up to 128x E3.S Modules @ 25W each
- Maximum Total Power 3.2KW
- Proposed airflow from bottom to top of E3.S modules
- Large Cooling surface area per module
- Baffling and Managing Airflow challenge



SERVER



HIGH
PERFORMANCE
COMPUTING

OPEN POSSIBILITIES.



Call to Action

- Please help bring HPCM to reality by Joining the OCP HPC Sub Project
- We are also seeking Funding in order to build PoCs to prove out Concepts
- Where to find additional information (URL links)

Project Wiki with latest HPC Charter and Meeting Recordings : <http://www.opencompute.org/wiki/HPC>

Mailing list: <https://ocp-all.groups.io/g/OCP-HPC>

Meeting Calendar : <https://www.opencompute.org/projects/high-performance-computing-incubation>

OPEN POSSIBILITIES.



Thank you!
Any Questions?