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

Nomenclature: Read “Processor” as CPU and/or Accelerator
Re-Inventing HPC Architectures for a “Domain Specific Architecture” Computing World

Allan Cantle, CEO, Nallasway
System Level Domain Specific Architecture Template

Sourced from Slide 11 of
OpenCAPI - A Memory Centric Fabric in Data Centric World
Recording
HPC is increasingly Data Bound & Less So Compute Bound
Heterogeneous w/ blurred Storage/Memory Boundaries

<table>
<thead>
<tr>
<th>System attributes</th>
<th>ALCF Now</th>
<th>NERSC Now</th>
<th>OLCF New</th>
<th>NERSC Pre-Exascale</th>
<th>ALCF Pre-Exascale</th>
<th>OLCF Exascale</th>
<th>ALCF Exascale</th>
</tr>
</thead>
<tbody>
<tr>
<td>System peak</td>
<td>&gt; 15.5 PF</td>
<td>&gt; 32 PF</td>
<td>200 PF</td>
<td>&gt; 120 PF</td>
<td>35 – 45 PF</td>
<td>&gt; 1.5 EF</td>
<td>&gt; 1.5 EF</td>
</tr>
<tr>
<td>Peak Power (MW)</td>
<td>&lt; 2.1</td>
<td>&lt; 3.7</td>
<td>10</td>
<td>6</td>
<td>&lt; 2</td>
<td>29</td>
<td>≤ 90</td>
</tr>
<tr>
<td>Total system memory</td>
<td>847 TB DOR4 + 70 TB HBM + 7.5 TB GPU memory</td>
<td>&gt; 1 PB DOR4 + 92 PB DDR4 + 192 PB HBM Persistent memory</td>
<td>2.4 PB DDR4 + 0.4 PB HBM + 7.4 PB Persistent memory</td>
<td>&gt; 250 TB</td>
<td>&gt; 4.6 PB DDR4 + 4.6 PB HBM2e + 36 PB Persistent memory</td>
<td>&gt; 10 PB</td>
<td></td>
</tr>
<tr>
<td>Node performance (TF)</td>
<td>2.7 TF (XNL nodes) and 196.4 TF (GPU node)</td>
<td>&gt; 3</td>
<td>43</td>
<td>&gt; 70 (GPU)</td>
<td>&gt; 70 TF</td>
<td>TBD</td>
<td>&gt; 130</td>
</tr>
<tr>
<td>Node processors</td>
<td>Intel Xeon Phi 7320 64-core CPUs (KNL) and GPU nodes with 8 NVIDIA A100 GPUs</td>
<td>Intel Knights Landing many core GPUs</td>
<td>Intel Haswell CPU in data partition</td>
<td>2 IBM Power7 CPUs + 6 Nvidia Volta GPUs</td>
<td>CPU-only nodes: AMD EPYC Milan CPU, CPU-GPU nodes: NVIDIA EPYC Milan with NVIDIA A100 GPUs</td>
<td>1 CPU + 4 GPUs</td>
<td>2 Intel Xeon: Sapphire Rapids and 8 Xe Parla node-to-node GPUs</td>
</tr>
<tr>
<td>System size (nodes)</td>
<td>4,292 KNL nodes and 24 DGX-A100 nodes</td>
<td>9,101 nodes in data partition</td>
<td>4096 nodes</td>
<td>&gt; 1,300 nodes CPU, &gt; 3,000 nodes GPU</td>
<td>&gt; 500</td>
<td>&gt; 9,000 nodes</td>
<td>&gt; 9,000 nodes</td>
</tr>
<tr>
<td>CPU-GPU Interconnect</td>
<td>NVLINK on GPU nodes</td>
<td>N/A</td>
<td>NVLINK</td>
<td>NVLINK</td>
<td>PCIe</td>
<td>AMD Infinity Fabric, Coherent memory across node</td>
<td>Unified memory architecture, RAMBEO</td>
</tr>
<tr>
<td>Node-to-node Interconnect</td>
<td>Aries (KNL nodes) and HDR200 (GPU nodes)</td>
<td>Aries</td>
<td>Dual Rambus DDR-8</td>
<td>HPE Slingshot NIC</td>
<td>HPE Slingshot NIC</td>
<td>HPE Slingshot</td>
<td>HPE Slingshot</td>
</tr>
<tr>
<td>File System</td>
<td>200 PB, 1.3 TB/s Lustre, 10 PB, 210 GB/s Lustre</td>
<td>28 PB, 744 GB/s Lustre</td>
<td>250 PB, 2.5 TB/s GPFS</td>
<td>35 PB All Flash, Lustre</td>
<td>N/A</td>
<td>695 PB + 10 PB Flash performance tier: Lustre</td>
<td>≤ 230 PB, ≥ 26 TB/s DAOS</td>
</tr>
</tbody>
</table>

OPEN POSSIBILITIES.

Frontier: https://www.alcf.anl.gov/foundier/
Aurora: https://www.alcf.anl.gov/aurora

ASCR Computing Upgrades At-a-Glance
November 24, 2020
Today’s HPC Compute & Storage Challenge

- 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

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

Power Hungry ToR Network
- 2.5TB/s

*Persistent Memory only used for Checkpoint Restarts
Data Centric HPC Solution - Abstract View

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

High Bandwidth Memory Pooling Data Plane Fabric

Many TeraBytes of Volatile Memory and / or Persistent Memory and / or Filesystem Storage Memory

Pluggable System Level 3D Building Blocks

High Performance Processor

Memory Subsystems will include Distributed Compute capability with tiny embedded Processors

Lowest Power (~1pJ/bit), Lowest Latency, High Bandwidth Pluggable Near Memory Interface

OPEN POSSIBILITIES.
If Tesla can “Re-Invent” then why not OCP?
We need to Innovate across Silos!
Disaggregated Racks to Hyper-converged Chiplets

Software Composable
Power Ignored
Rack Interconnect >20\text{pJ}/\text{bit}
Poor Latency
Rack Volume >53K Cubic Inches

Baseline Physical Composability
Power Baseline
Node Interconnect 5-10\text{pJ}/\text{bit}
Baseline Latency
Node Volume >800 Cubic Inches

Expensive Physical composability
Power Optimized
Chiplet Interconnect <1\text{pJ}/\text{bit}
Optimal Latency
SIP Volume <1 Cubic Inch

Power Ignored
Rack Interconnect >20\text{pJ}/\text{bit}
Poor Latency
Rack Volume >53K Cubic Inches

Baseline Latency
Node Volume >800 Cubic Inches

Power Optimized
Chiplet Interconnect <1\text{pJ}/\text{bit}
Optimal Latency
SIP Volume <1 Cubic Inch

OPEN POSSIBILITIES.
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

- Power Optimized Chiplet Interconnect <1pJ/bit

- Node Volume
- Rack Volume >53K Cubic Inches

- Server
- High Performance Computing

Open Possibilities.
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

Power Optimized
Chiplet Interconnect <1pJ/bit

Node Volume <1 Cubic Inch

Rack Volume >53K Cubic Inches

Server

High Performance Computing

Open Possibilities.
HPCM brings the two Together

Software Composable

Baseline Physical Composability

Power Ignored Rack Interconnect >20pJ/bit

Power Optimized Chiplet Interconnect 1-2pJ/bit

Power Baseline Node Interconnect 5-10pJ/bit

Optimal Latency

Baseline Latency

Node Volume >800 Cubic Inches

Module Volume <150 Cubic Inches

Power Optimized Flexible Chiplet Interconnect <1pJ/bit

Rack Volume >53K Cubic Inches

SIP Volume <1 Cubic Inch

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

OPEN POSSIBILITIES.
Overview of OCP HPC SubProject’s HPCM (High Performance Computing Module)

Allan Cantle, CEO, Nallasway
High Performance Computing Module, HPCM

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

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
Memory IO is finally going Serial!

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

**DDR DIMM**
- OMI in DDIMM Format
  - Introduced in August 2019
- OMI in E3.S
  - Proposed in 2020
  - GenZ in E3.S
  - Introduced in 2020

**Near Memory**
- Dual OMI x8 DDR4/5 Channel
- GenZ x16 DDR4 Channel
- CXL.mem in E3.S
  - Introduced in May 2021

**Far Memory**
- CXL x16 DDR5 Channel

**OPEN POSSIBILITIES.**
Modular Building Blocks Available Today

- Network, Memory, & IO use **Common EDSFF Interconnect**

Typically < 100W

- **SNIA - E1.S & E3.S**
- **Jedec - DDIMM**
- **GenZ in E3.S**
- **OMI in E3.S**

200W to 1KW

- **OCP - OAM**
- **CXL.mem in E3.S**

**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
Installing 8 HPCMs in OAI Chassis

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

**Hyve Design Solutions 19" Stacked System**
- 19 inch 6RU, 30 inch (762mm) depth
- 8"OAMs
- UBB: Combined FC+ 6 port HCM Topology
- 4*PCIE Gen3x16 slots for host uplink
- 12*PCIE Gen3 x16 slots for flexible IO expansion
  (PCIE interface will be revised to Gen4 in next release.)

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

**Open Possibilities.**
Re-Architect - Start with a Cold Plate
For High Wattage HPCM Modules

- 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
Cold Plate from Backside

- 54V Power Bus Bars shown - Powering HPCMs
Add Topology Cabling - No Retimers

- Fully Connected Topology + Connections to HIB & QDD IO
Add E3.S and NIC 3.0 Modules

- Pluggable into OCP OAI Chassis
How HPCM provides Efficient & Flexible Interconnect to support increased Fabric Speeds

Allan Cantle, CEO, Nallasway
Tang Junyan, Mahesh Bohra, Dan Dreps, IBM
Bob Dillman & Gus Panella, Molex
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!
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
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
HPCM Processor to Module Interconnect
With Packaged Processor and Controller Chips

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)

Courtesy of IBM
## Insertion Loss allocation Table - Conservative

With Packaged Processor and Controller Chips

<table>
<thead>
<tr>
<th>Channel Section</th>
<th>Loss @ 32Gb/s</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>GL102 package wiring (30mm)</td>
<td>2.8dB</td>
<td></td>
</tr>
<tr>
<td>Module Via S12</td>
<td>1dB</td>
<td>Assume 1.6mm via length &amp; 15mil back drilled stub</td>
</tr>
<tr>
<td>Meg6 - 24mm Pin Area Wiring</td>
<td>1.2dB</td>
<td>Conservative assumption with 30mm package wiring &amp; 24mm PCB zig-zag wiring under package</td>
</tr>
<tr>
<td>Meg6 - Open Area PCB Trace (67mm)</td>
<td>2.6dB</td>
<td></td>
</tr>
<tr>
<td>DIMM PCB Via S12</td>
<td>0.9dB</td>
<td>Assume 1.6mm via length &amp; 15mil back drilled stub</td>
</tr>
<tr>
<td>DIMM Conn (C2)</td>
<td>1dB</td>
<td></td>
</tr>
<tr>
<td>Meg6 - DIMM Open Area (37mm)</td>
<td>1.4dB</td>
<td></td>
</tr>
<tr>
<td>E3.S Controller Package</td>
<td>0.4dB</td>
<td></td>
</tr>
<tr>
<td><strong>Total Channel</strong></td>
<td><strong>11.3dB</strong></td>
<td>Measured channel difference due to impedance discrepancies &amp; behavior</td>
</tr>
</tbody>
</table>
Insertion Loss allocation Table - Conservative
Derived with Bare Die Processor and Controller Chips

<table>
<thead>
<tr>
<th>Channel Section</th>
<th>Loss @ 32Gb/s</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Module Via S12</td>
<td>1dB</td>
<td>Assume 1.6mm via length &amp; 15mil back drilled stub</td>
</tr>
<tr>
<td>Meg6 - 24mm Open Area Wiring</td>
<td>1dB</td>
<td></td>
</tr>
<tr>
<td>Meg6 - Open Area PCB Trace (67mm)</td>
<td>2.6dB</td>
<td></td>
</tr>
<tr>
<td>DIMM PCB Via S12</td>
<td>0.9dB</td>
<td>Assume 1.6mm via length &amp; 15mil back drilled stub</td>
</tr>
<tr>
<td>DIMM Conn (C2)</td>
<td>1dB</td>
<td></td>
</tr>
<tr>
<td>Meg6 - DIMM Open Area (37mm)</td>
<td>1.4dB</td>
<td></td>
</tr>
<tr>
<td>Total Channel</td>
<td>7.9dB</td>
<td>Empirical estimate only</td>
</tr>
</tbody>
</table>
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
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
  - ~ 8.5 + 2 + 4.4 + 0.7 = 15.6dB
  - 14.4dB spare for HIB PCB and Switch package losses
- Retimerless compared to UBB implementations

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
  - ~ 8.5 + 0.7 + 6.3 + 2.5 = 18dB
- 12dB spare - May Support Passive QDD Cables
- Retimerless compared to UBB implementations
Cabled Solutions are reliable

- IBM’s High Reliability E1080 Server
How HPCM’s Thermal Management Cold Plate solution turns traditional approaches on their head

Chris Chapman, Boyd Corporation
Bob Dillman, Molex
Allan Cantle, Nallasway
HPCM proposed Thermal Solution

Cooling 8x HPCM Module Processors

Maximum Power per HPCM = 1KW
8 HPCM’s total 8KW Max
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
HPCM proposed Thermal Solution

- Water Cooled Cold Plate
- Provides HPCM Mechanical Infrastructure
- Cold water to each HPCM site
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
Evaluate “Cooling Core”

• Initial CFD analyzed the new form factor required
• Similar performance was obtained compared to a traditional OAM module cold plate
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
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.
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.
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
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
Thank you!
Any Questions?