

# 下一代HPC：内存为中心的开放组合式硬件

原创 唐僧 huangliang 企业存储技术 2022-01-17 08:59

收录于话题

#服务器 65 #HPC高性能计算 12

## 目录

- Exascale HPC的计算和存储挑战（让内存更贴近计算）
- OCP HPCM：高密度模块的功率和液冷需求
- Summit超算节点升级：POWER10 + Xilinx FPGA
- 百亿亿级超算Frontier：AMD EPYC4 + MI200 GPU + Xilinx

接前文：《[AMD EPYC Genoa '7004'：若DDR5换OMI，CXL.mem服务器内存池化](#)》

《[OMI串行内存\(续\)：当POWER10遇上OCP-HPC & OAM](#)》



今天的内容，来自我对NALLASWAY分享《HPC - The Next Generation Data (Memory) Centric HPC with Open Composable Hardware》的学习笔记。

## Exascale HPC的计算和存储挑战

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	> 120PF	35 – 45PF	>1.5 EF	≥ 1 EF DP sustained
Peak Power (MW)	< 2.1	< 3.7	10	6	< 2	29	≤ 60
Total system memory	847 TB DDR4 + 70 TB HBM + 7.5 TB GPU memory	~1 PB DDR4 + High Bandwidth Memory (HBM) + 1.5PB persistent memory	2.4 PB DDR4 + 0.4 PB HBM + 7.4 PB persistent memory	1.92 PB DDR4 + 240TB HBM	> 250 TB	4.6 PB DDR4 + 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 54-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 + 6 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 CPU 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	9,300 nodes 1,900 nodes in data partition	4608 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, RAMBO
Node-to-node interconnect	Aries (KNL nodes) and HDR200 (GPU nodes)	Aries	Dual Rail EDR-IB	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 GPFS	35 PB All Flash, Lustre	N/A	695 PB + 10 PB Flash performance tier, Lustre	≥ 230 PB, ≥ 25 TB/s DAOS



Office of Science

Frontier: <https://www.olcf.ornl.gov/frontier/>  
Aurora: <https://www.alcf.anl.gov/aurora>

ASCR Computing and Storage Science  
November 24, 2020

以上列出的几套HPC超级计算机都在美国，并且它们有共同的特点——Heterogeneous w/blurred Storage/Memory Boundaries（异构并且带有界限模糊的存储/内存组合），即除了传统DRAM系统内存、GPU内存（显存）之外，可能还有HBM高带宽内存、持久化内存。

这些HPC的拥有者，来自三家著名实验室——ALCF（阿贡国家实验室）、NERSC（国家能源研究科学计算中心，位于劳伦斯伯克利国家实验室）和OLCF（Oak Ridge National Laboratory，橡树岭国家实验室）。其中OLCF的Summit由IBM建造，当前排在Global Top500算力榜第2位；NERSC的Perlmutter由HPE（也就是收购的Cray）建造，使用了AMD EPYC + NVIDIA A100。

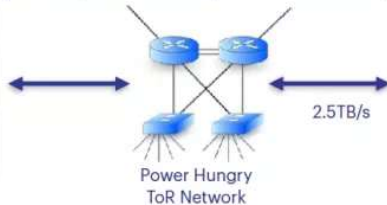
我在前文中提到过，未来的2套Exascale（百亿亿次级）HPC系统IBM都出局了。OLCF的Frontier将基于AMD EPYC CPU + Radeon Instinct GPU，存储的文件系统为Lustre；ALCF的Aurora则采用下一代Intel Xeon（Sapphire Rapids CPU + Xe Ponte Vecchio GPU），存储采用DAOS（分布式异步对象存储）。它们除了浮点算力将达到1 EFlops之外，整体功耗也比以前有较大提高，相比Summit的Peak Power 10 MW（兆瓦），Frontier和Aurora分别将达到29 MW和≤ 60 MW。

# 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\*



Power Hungry  
ToR Network



Summit HPC GPFS File System  
250 PetaBytes of Storage

2.5TB/s



SERVER



HIGH  
PERFORMANCE  
COMPUTING



OPEN POSSIBILITIES. \*Persistent Memory only used for Checkpoint Restarts

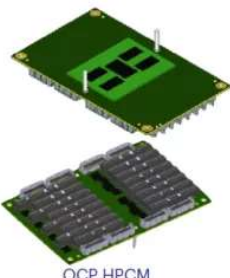
我们先从橡树岭的Summit来看看今天HPC的计算和存储挑战。从文件系统加载2.8PB数据到内存一次需要18分钟，通过计算机架向全部250PB文件系统推送一遍数据需要1.2天，这还是在GPFS文件系统存储总带宽达到2.5TB/s的情况下。因此，需要将计算、内存和存储更加“接近”（让内存更贴近计算）。

注：Summit集群一共还配备有7.4PB的持久化内存，用途只是为了Checkpoint重启恢复。

## OCP HPCM：高密度模块的功率和液冷需求

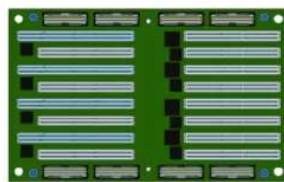
### High Performance Computing Module, HPCM

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

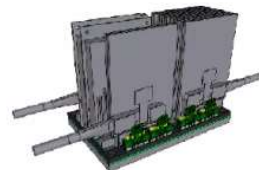


OCP HPCM

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 Modules,  
8x Nearstack x8 Cables



SERVER



HIGH  
PERFORMANCE  
COMPUTING



OPEN POSSIBILITIES.



由OCP提出的HPCM（高性能计算模块）我在前面2篇中已经给大家介绍过，它有一个要求——处理器必须使用HBM或者拥有串行连接的内存，因为这个架构无法使用传统的DIMM内存模组。

扩展阅读：《[CXL、GenZ、CCIX架构以及未来的PM、内存和SSD形态](#)》  
《[数据中心NVMe SSD和EDSFF前瞻：来自Intel、HPE、Dell& SNIA等](#)》

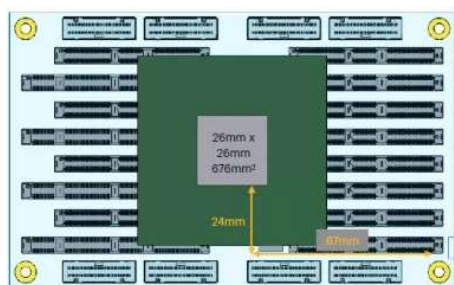
HPCM设计主要的好处是：模块化、灵活和可组合的模块——并且是协议无关的（也就是异构）。

在HPCM今天支持的处理器列表里，我们看到有NVIDIA Ampere、Google TPU、IBM POWER10、Xilinx和Intel的FPGA、GraphcoreIPU，以及PCIe和以太网Switch交换机。带有一大堆插槽的方向是它的背面（主处理器在正面），内存可以通过OMI串行接口连接，EDSFF（E3.S）也可以扩展内存、SSD或者SCM等，此外还有OCP NIC 3.0网卡。

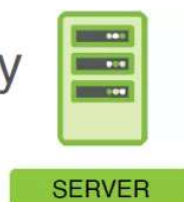
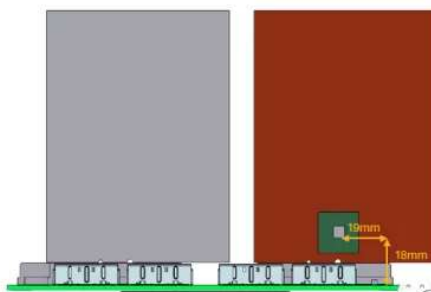
由于Sapphire Rapids将提供带有RAMBOCache（HBM）的版本，我理解如果Intel未来想支持HPCM应该不难？都说一流的公司做标准，EDSFF和CXL这两大Intel主导的标准，都已经用在HPCM中了。

## 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



OPEN POSSIBILITIES.

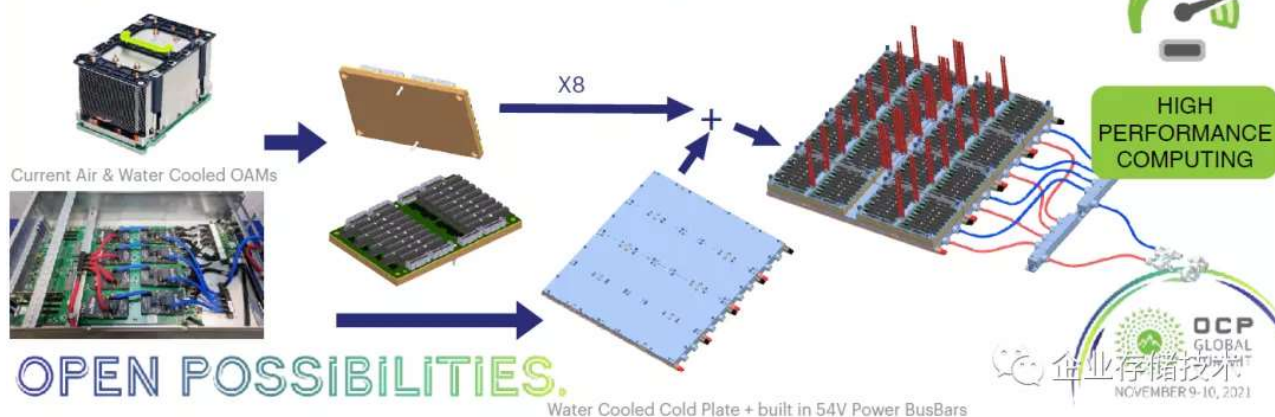


高密度模块还有个好处——减少导线长度带来的功率损失（电阻）。如上图，从处理器的Die引出到达E3.S模块上的ASIC芯片，只有不到5英寸。这样有机会降低PHY通道信号强度5-10dB，发热量1-2pJ/bit。

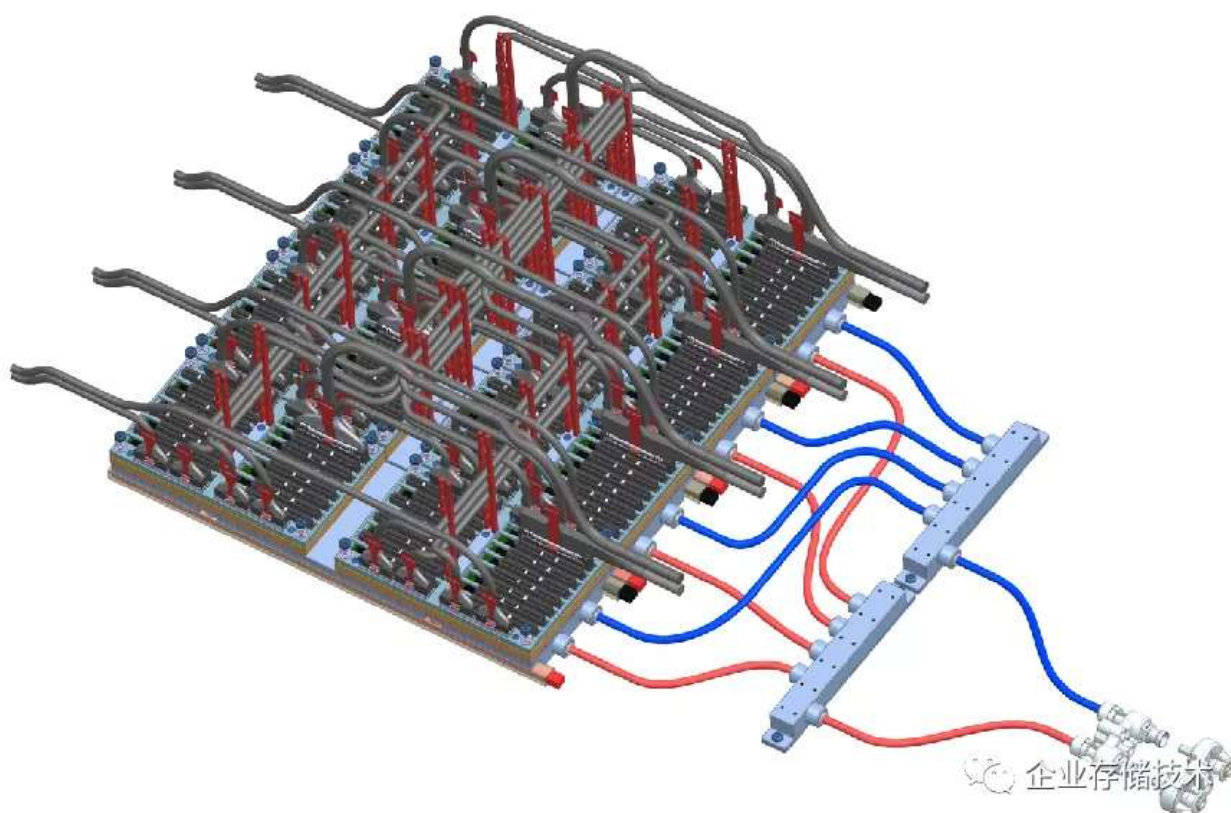
# Re-Architect - Start with a Cold Plate

## For High Wattage HPCM Modules

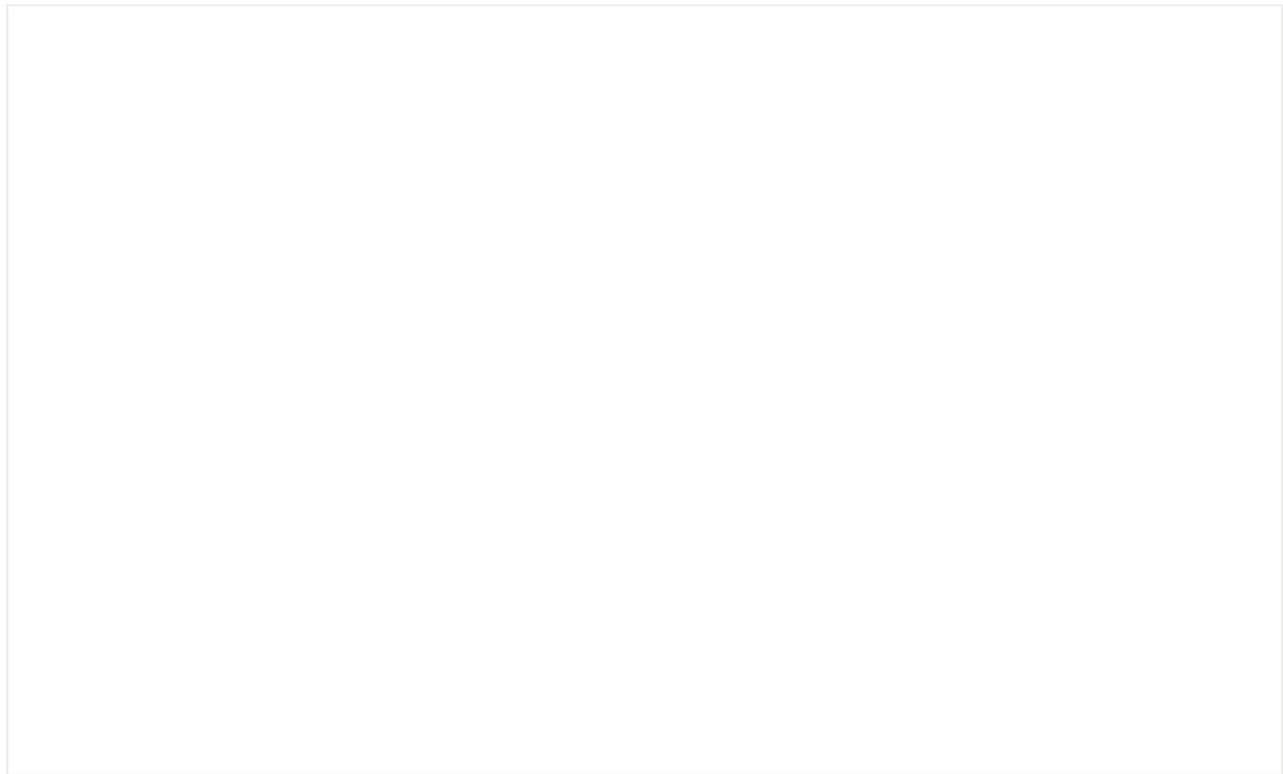
- Capillary Heatspreader on module to dissipate die heat across module surface area
- Heatsinks are largest Mass, so make them the structure of the assembly
- Integrate liquid cooling into the main cold plate



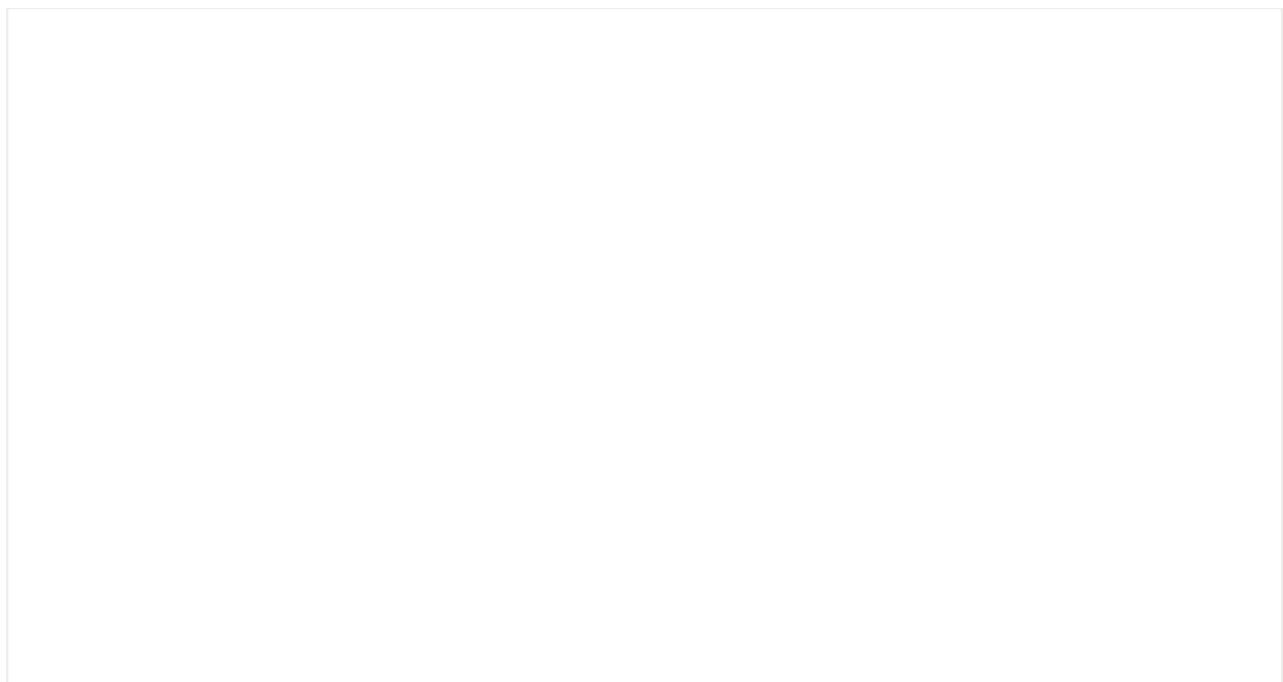
高功率密度的HPCM模块，需要重新设计冷板式液冷架构。为了节约占地空间、提高PUE能源效率，液冷用于高性能计算集群早已不新鲜，比如我在几年前写过的国产超算神威太湖之光也是液冷散热的。上图可以看到HPCM用于处理器一面的巨大散热片，以及蓝色和红色的冷热水管/水路设计。



在之前的文章中我有一点可能搞错了——HPCM模块背面的高密度I/O模组，单纯风冷散热估计也不太容易搞定。从上面列出的水管走向布局就可以看出这个。



上图应该是molex的散热仿真系统（模拟在OCP OAI机箱内），一共排列了4x4=16个HPCM模块。当朝上的“背面”都插满E3.S和NIC 3.0模块，这个密度不是一般的存储服务器或者大内存系统可比吧。



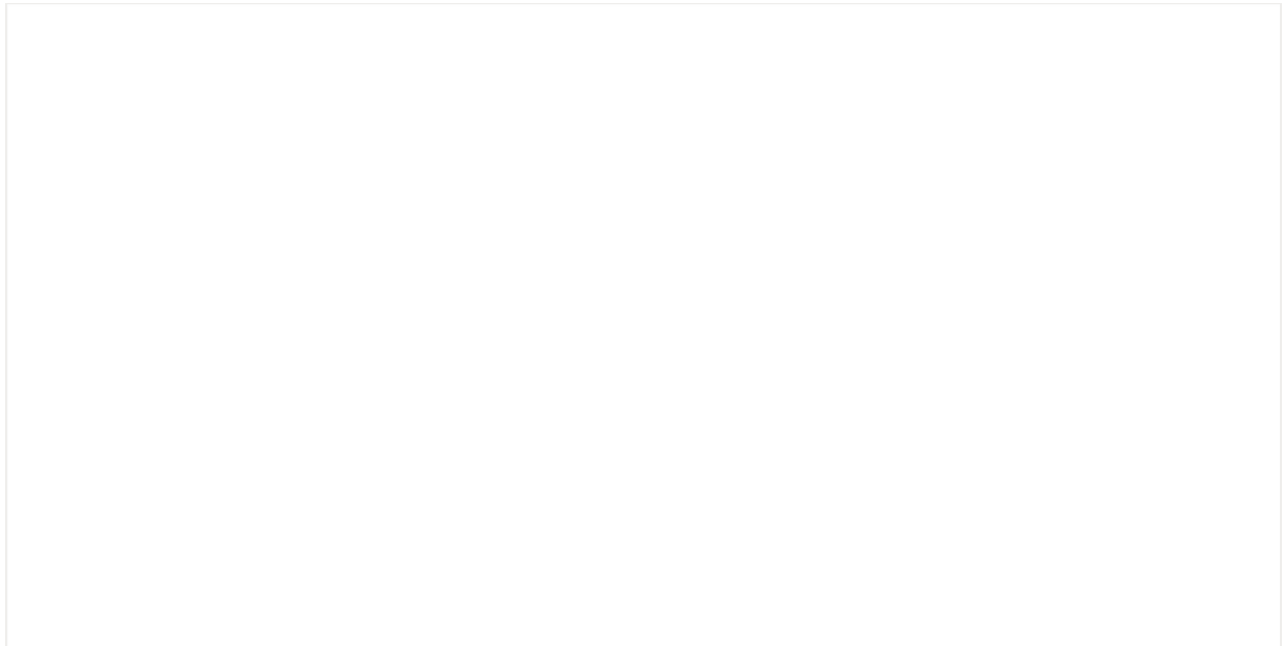
这里提到的POC/Demo，就是开发一组小型HPCM模块可以/需要包含的部分：

- 1个CPU——带有串行内存——Power10 HPCM（注：AMDEPYC4还没有正式发布）
- 1个GPU——带有HBM板载内存——AMD MI200 HPCM
- 1个FPGA——支持所有IO类型——Xilinx Versal VP1802 HPCM

- 1个Switch——PCIe/CXL或者以太网——Microchip或Broadcom?
- 1个AI处理器——带有HBM或者串行内存?

大家注意到共同特点了吧？前面我也提到过，在各种HPCM当中，除了用于数据交换的Switch模块之外，所有处理器都需要本地/近端内存。

### Summit超算节点升级：POWER10+ Xilinx FPGA




如果我没理解错的话，橡树岭的Summit应该也会有节点上的升级。（欢迎大家指出文中错误，在下面留言）



Summit应该是2018年完工的，曾经有2年排在HPC Top500榜首吧。上图中绿色的连线代表NVIDIA V100 GPU的NVLINK——GPU之间互连很好理解，而从Power9 CPU连接GPU的NVLINK通道其实[我以前也写过](#)（这部分NVLINK也可以配置为OpenCAPI）。





上图是基于POWER10和FPGA的Summit节点，AI/ML大数据型配置——可达56TB内存@8TB/s带宽。

在一个OAI机箱内，有2个IBM POWER 10单芯片CPU HPCM模块，和6个Xilinx VP1802 FPGA HPCM模块。CPU和FPGA之间主要通过OpenCAPI连接，另外也有PCIe Gen5；在FPGA之间有一种x4 56G Generic（绿色）和另一种32G Generic（紫红色）互连，加载一起算是做到了点对点网状连接吧。

内存部分，每颗POWER10的HPCM上可以安装4TB OMI内存；每颗Xilinx FPGA则可以配置8TB OMI内存。如果不差钱，这个异构系统的内存可以全部由POWER来管理调度。

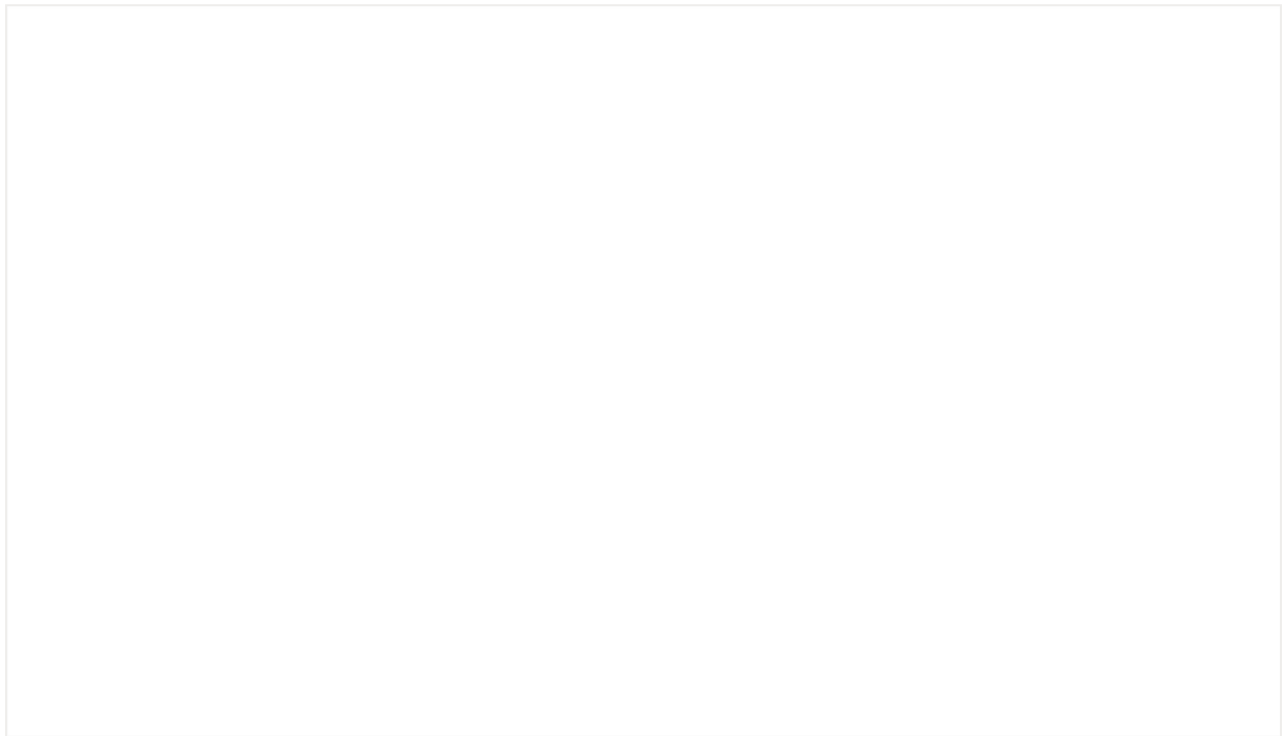


我分享的这份资料中，也有POWER10的HPCM示意图，由于[前文](#)中列出过就不重复了，下面重点看Xilinx VP1802。

如上图，蓝色的连接代表FPGA的PCIe G5通道（硬IP），而黄色的则是可配置的软IP——可以是以太网（最高112G/txr）、OMI或者PCIe/CXL等。

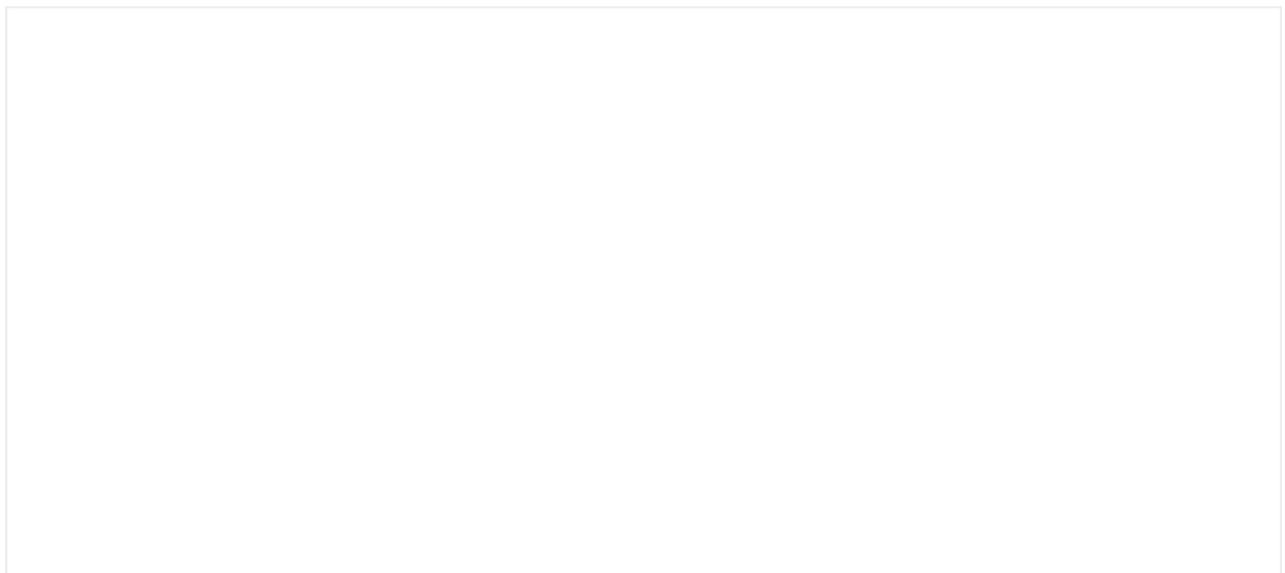


具体到Summit节点中Xilinx HPCM的用法，FPGA的软IP I/O通道被配置成了32x4=128 OMI Lanes——用来连接最大512GB的双OMI通道DDR5内存模组，也就是8TB Max Near Memory。



上图为512GB DDR4 双OMI E3.S内存模组的参考（DDR5还没量产吗？）。2个Microchip OMI Buffer芯片提供双x8或者x4通道，双面总共72个颗粒具备ECC支持。

### 百亿亿级超算Frontier: AMDEPYC4 + MI200 GPU + Xilinx FPGA



最后，我们再来看看橡树岭下一代的Frontier Exascale超算。AMD的基本架构图可能有些朋友见过了，结合下面的资料，我判断上图中刀片容纳的是2个单AMD EPYC CPU的服务器节点。



看下这张图，大家明白AMD为什么要收购Xilinx了吧：)


这里列出带有GPU和FPGA的Frontier节点，同样是针对AI/ML大数据的配置。“OCP Server”刀片机箱中是AMD EPYC Genoa CPU，与OAI机箱之间通过PCIe Gen5/CXL/xGMI直通/交换连接。

具体到OAI机箱内部，4颗AMD MI200GPU HPCM之间是AMD自己的xGMI (Infinity Fabric) 互连，GPU与右边4个Xilinx FPGA HPCM之间是否有更多连接还不确定？FPGA在这里最大的作用可能是扩展内存——具体到AI/ML计算到底由FPGA还是GPU来做我也说不准？FPGA HPCM上配置的OMI内存倒是可以通过CXL共享给CPU一致性访问。

思考题：AMD EPYC4为什么不放到OAI机箱的HPCM模块上？（[我曾经写过该CPU支持OMI的可能](#)）

答案：按Frontier节点的设计，AMD CPU本地内存使用相对廉价的传统DDR5而不是OMI，这样就无法支持HPCM模块（原因我在上文中反复提过了），不过好处应该是性价比相对POWER好一点。





可能是由于AMD MI200 GPU公开的资料还有限，上面像是一张临时修改出来的图——从OAM模块迁移到HPCM。已经确定的4条I/O就是连接到CPU和另外3颗GPU的x16 Infinity Fabric（也可定义为PCIe）。

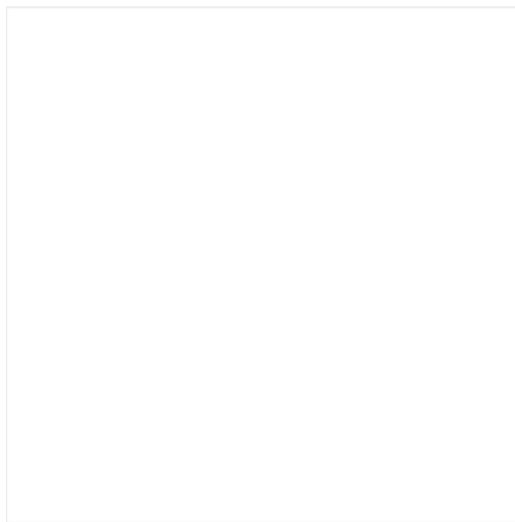
就写到这里吧，希望大家看了有点用：)

参 考 资 料 <https://pan.baidu.com/s/1StGWB5UqmKypcE2YOAqIOg?pwd=t4sg>  
提取码：t4sg

扩展阅读：《[企业存储技术](#)》[文章分类索引（微信公众号专辑）](#)》

*注：本文只代表作者个人观点，与任何组织机构无关，如有错误和不足之处欢迎在留言中批评指正。如果您想在这个公众号上分享自己的技术干货，也欢迎联系我：)*

尊重知识，转载时请保留全文，并包括本行及如下二维码。感谢您的阅读和支持！  
《企业存储技术》微信公众号：**HL\_Storage**



长按二维码可直接识别关注

历史文章汇总：<http://www.toutiao.com/c/user/5821930387/>  
<http://www.zhihu.com/column/huangliang>

点击下方“阅读原文”，查看更多历史文章



唐僧 huangliang

“ 欢迎转发，评论！ ”

喜欢作者

收录于话题 #服务器 65

下一篇 · DDR5支持On-die ECC，但为什么服务器内存价格会提高？

阅读原文 阅读 2443 文章已于2022/01/17修改

分享

收藏

赞 14

在看 9

写下你的留言