Summit 架构分析

201708010814-李宗鸿

一、 Summit 简介

Summit 计算机是目前（2019 年）运行速度最快的超级计算机，由美国能源部橡树岭国家实验室发布，其合作伙伴是 IBM 和 NVIDIA。

从最新的（2019 年 11 月 18 日）发布的世界前五百的超算排名数据上来看。其峰值浮点性能为 200.795 PFlop /秒，Linpack 浮点性能为 148.6PFlop /秒。功率为 10,096 kW。

RANKING

List	Rank	System	Vendor	Total Cores	Rmax (TFlops)	Rpeak (TFlops)	Power (kW)
11/2019	1	IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband	IBM	2,414,592	148,600.0	200,794.9	10,096.00

如图为超算排名提供的数据：

Home / DOE/SC/Oak Ridge Natio... / Summit - IBM Power System ...

Summit - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband

Site:	DOE/SC/Oak Ridge National Laboratory
System URL:	http://www.olcf.ornl.gov/olcf-resources/compute-systems/summit/
Manufacturer:	IBM
Cores:	2,414,592
Memory:	2,801,664 GB
Processor:	IBM POWER9 22C 3.07GHz
Interconnect:	Dual-rail Mellanox EDR Infiniband
Performance	
Linpack Performance (Rmax)	148,600 TFlop/s
Theoretical Peak (Rpeak)	200,795 TFlop/s
Nmax	16,473,600
HPCG [TFlop/s]	2,925.75
Power Consumption	
Power:	10,096.00 kW (Submitted)
Power Measurement Level:	3
Measured Cores:	2,397,824
Software	
Operating System:	RHEL 7.4
Compiler:	XLC, nvcc
Math Library:	ESSL, CUBLAS 9.2
MPI:	Spectrum MPI

Summit 共有 27648 块 Tesla V100 计算卡以及 9216 颗 IBM POWER9 22C 3.07GHz CPU，核心数量达到 2414592 个。NV 表示这些流处理器通过 NVLink2.0 进行连接。每一块 Nvidia Tesla V100 计算卡拥有 5120 颗流处理器，所以 Summit 一共拥有 141557760 个 CUDA。总内存为 2801664GB，使用 Dual-rail Mellanox EDR Infiniband 进行互联。操作系统为 RHEL 7.4。

从架构角度来看，Summit 并没有在超算的底层技术上予以彻底革新，而是通过不断使用先进制程、扩大计算规模来获得更高的性能。虽然扩大规模是提高超算效能的有效方式，但是为了将这样多的 CPU、GPU 和相关存储设备有效组合也是一件困难的事情。在这一点上，Summit 采用了多级结构。最基本的结构被称为计算节点，众多的计算节点组成了计算机架，多个计算机架再组成 Summit 超算本身。

二、 计算节点:

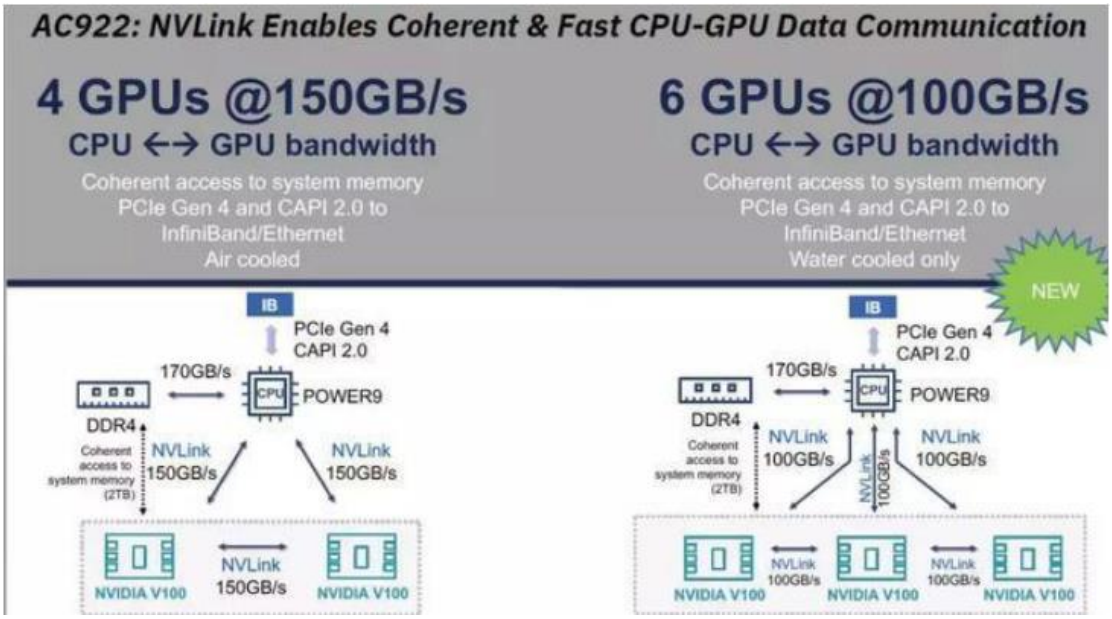
➢ 2CPU+6GPU 异构运算体系

Summit 采用的计算节点型号为 Power System AC922，之前的研发代号为 Witherspoon，这是一种 19 英寸的 2U 机架式外壳。从内部布置来看，每个 AC922 内部有 2 个 CPU 插座，满足两颗 Power 9 处理器的需求。每颗处理器配备了 3 个 GPU 插槽，每个插槽使用一块 GV100 核心的计算卡。这样 2 颗处理器就可以搭配 6 颗 GPU。

内存方面，每颗处理器设计了 8 通道内存，每个内存插槽可以使用 32GB DDR4 2666 内存，这样总计可以给每个 CPU 可以带来 256GB、107.7GB/s 的内存容量和带宽。GPU 方面，它没有使用了传统的 PCIe 插槽，而是采用了 SXM2 外形设计，每颗 GPU 配备 16GB 的 HBM2 内存，对每个 CPU-GPU 组而言，总计有 48GB 的 HBM2 显存和 2.7TBps 的带宽。

NV Link 2.0

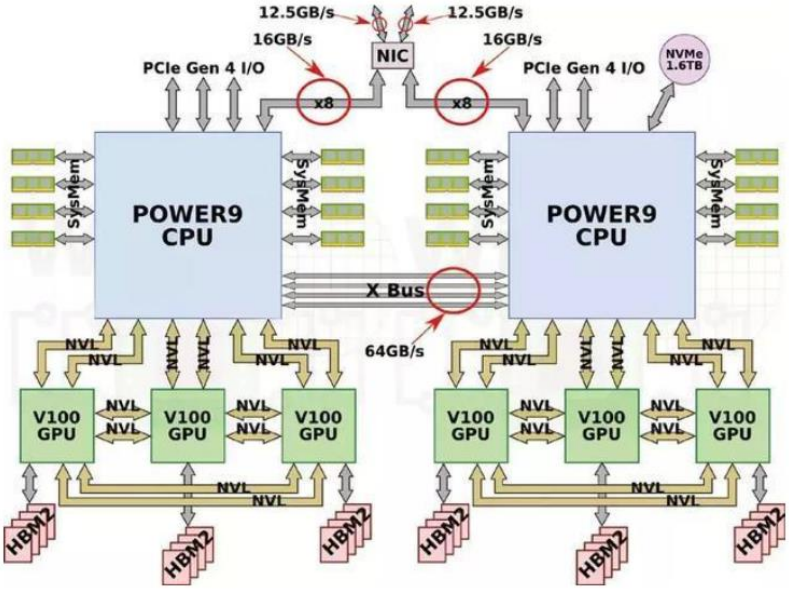
CPU 和 GPU 通过 NVLink 进行连接而不是使用 PCIe 总线(NVLink，是 NVIDIA 开发并推出的一种总线及其通信协议。NVLink 采用点对点结构、串列传输，用于中央处理器（CPU）与图形处理器（GPU）之间的连接，也可用于多个图形处理器之间的相互连接).



单颗 Power 9 处理器有 3 组共 6 个 NVLink 通道，每组 2 个通道。由于 Power 9 处理器的 NVLink 版本是 2.0，因此其单通道速度已经提升至 25GT/s，2 个通道可以在 CPU 和 GPU 之间实现双向 100GB/s 的带宽，此外，Power 9 还额外提供了 48 个 PCIe 4.0 通道。和 CPU 类似，GV100 GPU 也有 6 个 NVLink 2.0 通道，同样也分为 3 组，其中一组连接 CPU，另外 2 组连接其他两颗 GPU。和 CPU-GPU 之间的链接一样，GPU 与 GPU 之间的连接带宽也是 100GB/s。

CPU 之间的通讯

除了 CPU 和 GPU、GPU 之间的通讯外，由于每个 AC922 上拥有 2 个 CPU 插槽，因此 CPU 之间的通讯也很重要。Summit 的每个节点上，CPU 之间的通讯依靠的是 IBM 自家的 X 总线。X 总线是一个 4byte 的 16GT/s 链路，可以提供 64GB/s 的双向带宽，能够基本满足两颗处理器之间通讯的需求。



另外在 CPU 的对外通讯方面，每一个节点拥有 4 组向外的 PCIe 4.0 通道，包括两组 x16（支

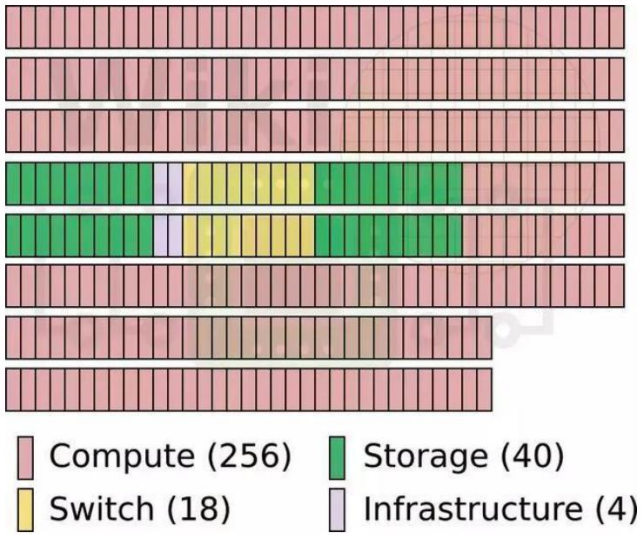
持 CAPI），一组 x8（支持 CAPI）和一组 x4。其中 2 组 x16 通道分别来自于两颗 CPU，x8 通道可以从一颗 CPU 中配置，另一颗 CPU 可以配置 x4 通道。其他剩余的 PCIe 4.0 通道就用于各种 I/O 接口，包括 PEX、USB、BMC 和 1Gbps 网络等。

完整的节点性能

Summit 的一个完整节点拥有 2 颗 22 核心的 Power 9 处理器，总计 44 颗物理核心。每颗 Power 9 处理器的物理核心支持同时执行 2 个矢量单精度运算。换句话说，每颗核心可以在每个周期执行 16 次单精度浮点运算。在 3.07GHz 时，每颗 CPU 核心的峰值性能可达 49.12GFlops。一个节点的 CPU 双精度峰值性能略低于 1.1TFlops，GPU 的峰值性能大约是 47TFlops。

机架和系统

机架是由计算节点组成的并行计算单元，Summit 的每个机架中安置了 18 个计算节点和 Mellanox IB EDR 交换器。每个节点都配备了双通道的 Mellanox InfiniBand ConnectX5 网卡，支持双向 100Gbps 带宽。节点的网卡直接通过插槽连接至 CPU，带宽为 12.5GBx2—实际上每个节点的网络都是由 2 颗 CPU 分出的 PCIe 4.0 x8 通道合并而成，PCI-E 4.0 x8 的带宽为 16GB/s，合并后的网卡可以为每颗 CPU 提供 12.5GB/s 的网络直连带宽，这样做可以最大限度地降低瓶颈。



▲国外WikiChip机构制作的Summit的系统结构布局图。

由于一个机架有 18 个计算节点，因此总计有 9TB 的 DDR4 内存和另外 1.7TB 的 HBM2 内存，总计内存容量高达 10.7TB。一个机架的最大功率为 59kW，峰值计算能力包括 CPU 的话是 846TFlops，只计算 GPU 的话是 775TFlops。在机架之后就是整个 Summit 系统了。完整的 Summit 系统拥有 256 个机架，18 个交换机架，40 个存储机架和 4 个基础架构机架。完整的 Summit 系统拥有 2.53PB 的 DDR4 内存、475TB 的 HBM2 内存和 7.37PB 的 NVMe SSD 存储空间。

Summit 使用液冷系统，每分钟流量高达 4000 加仑，4608 台主机连同液冷系统的整机组全速运行时的功率就高达一千五百万瓦。