湖南大學

HUNAN UNIVERSITY

计算机系统设计

选题名称：summit 架构分析

姓　　名：肖若愚

学　　号：201708010619

专业班级：物联 1702

一、Summit 介绍

Summit 超级电脑，实验室代号“OLCF-4”，是 IBM 为美国能源部旗下橡树岭国家实验室开发建造的超级电脑。机组于 2018 年 6 月 8 日落成，理论运算能达 200 PFLOPS（浮点运算速度每秒 20 亿亿次），超过峰值运算性能 125 PFLOPS 的神威·太湖之光，被认为有可能成为世界上最快的超级计算机。2018 年 6 月 25 日正式获 TOP500 认证为全球最快的超级电脑。

二、summit 设计

summit 一共有 4,608 个运算节点，每节点就是一台主机，每个节点内仍然使用与泰坦类似的 CPU+GPU 异质运算体系，由两颗 POWER9 CPU 以及六块 NVIDIA Tesla V100 运算加速卡组成，CPU 与 GPU 之间的连接采用的是英伟达（NVIDIA）开发的 NVLink 总线而非常见的 PCIe[8]，每个节点的 CPU 和 GPU 共享一共 512GiB 的一致性存储器（GPU 拥有的第二代高带宽存储器，加上 CPU 拥有的多通道 DDR4 存储器），CPU 和 GPU 可相互直接访问这个存储器空间以共享数据，另外还配备了容量高达 800GB 的非易失性随机存取存储器（NVRAM）作为突发性缓存或扩展存储器容量之用。

每个节点之间的连接采用的是双路 InfiniBand 互联，并使用非阻塞胖树拓扑（non-blocking fat-tree topology）交换结构，每路带宽为 200Gb/s。容量高达 250PB 的分布式存储系统也使用 InfiniBand 与运算节点连接。

本机组另建于新机房内，该机房占地有约两个网球场的面积（约 522 平方米），与橡树岭国家实验室已有的泰坦不同，泰坦使用大型空冷系统冷却，而高峰则是使用液冷系统，每分钟流量高达 4,000 加仑，4,608 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，几乎是泰坦的两倍。本机组仅 GPGPU 部分的双精度浮点数的运算性能就高达 215 PFLOPS；Tesla V100 内置有用于深度学习运算的 Tensor Core，因此每颗 GPGPU 也能提供约 125 TFLOPS 的混合精度浮点数性能，而全机组的更高达 3.3 EFLOPS（1 EFLOPS=1000 PFLOPS）。

三、summit 硬件架构

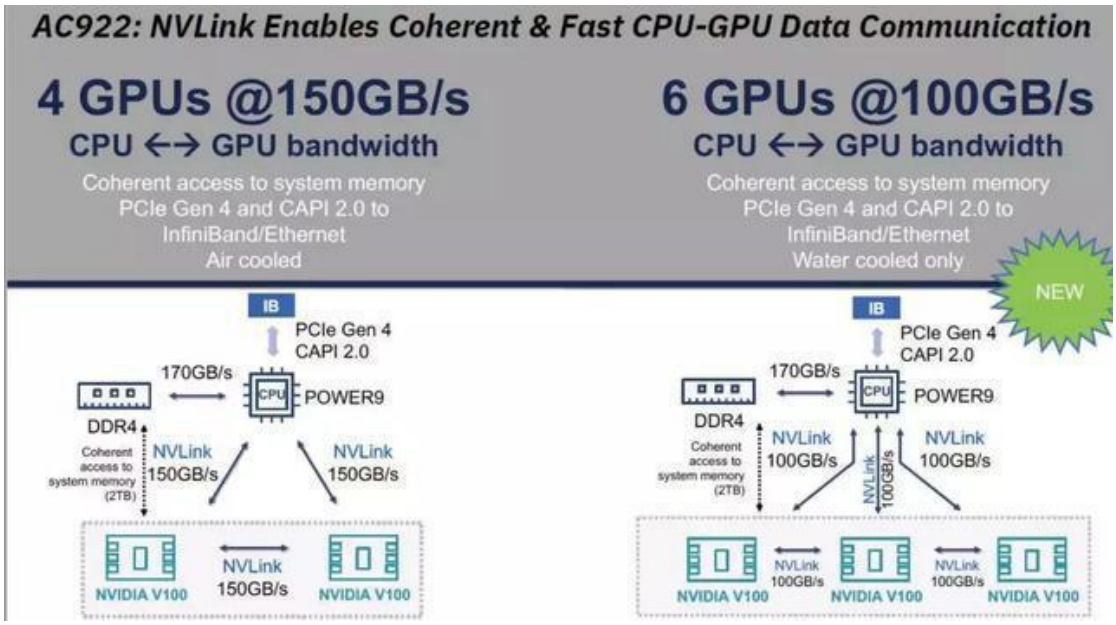
从硬件架构方面来看，Summit 依旧采用的是异构方式，其主 CPU 来自于 IBM Power 9，22 核心，主频为 3.07GHz，总计使用了 103752 颗，核心数量达到 2282544 个。GPU 方面搭配了 27648 块英伟达 Tesla V100 计算卡，总内存为 2736TB，操作系统为 RHEL 7.4。从架构角度来看，Summit 并没有在超算的底层技术上予以彻底革新，而是通过不断使用先进制程、扩大计算规模来获得更高的性能。
虽然扩大规模是提高超算效能的有效方式，但是为了将这样多的 CPU、GPU 和相关存储设备有效组合也是一件困难的事情。在这一点上，Summit 采用了多级结构。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。

四、summit 节点分析

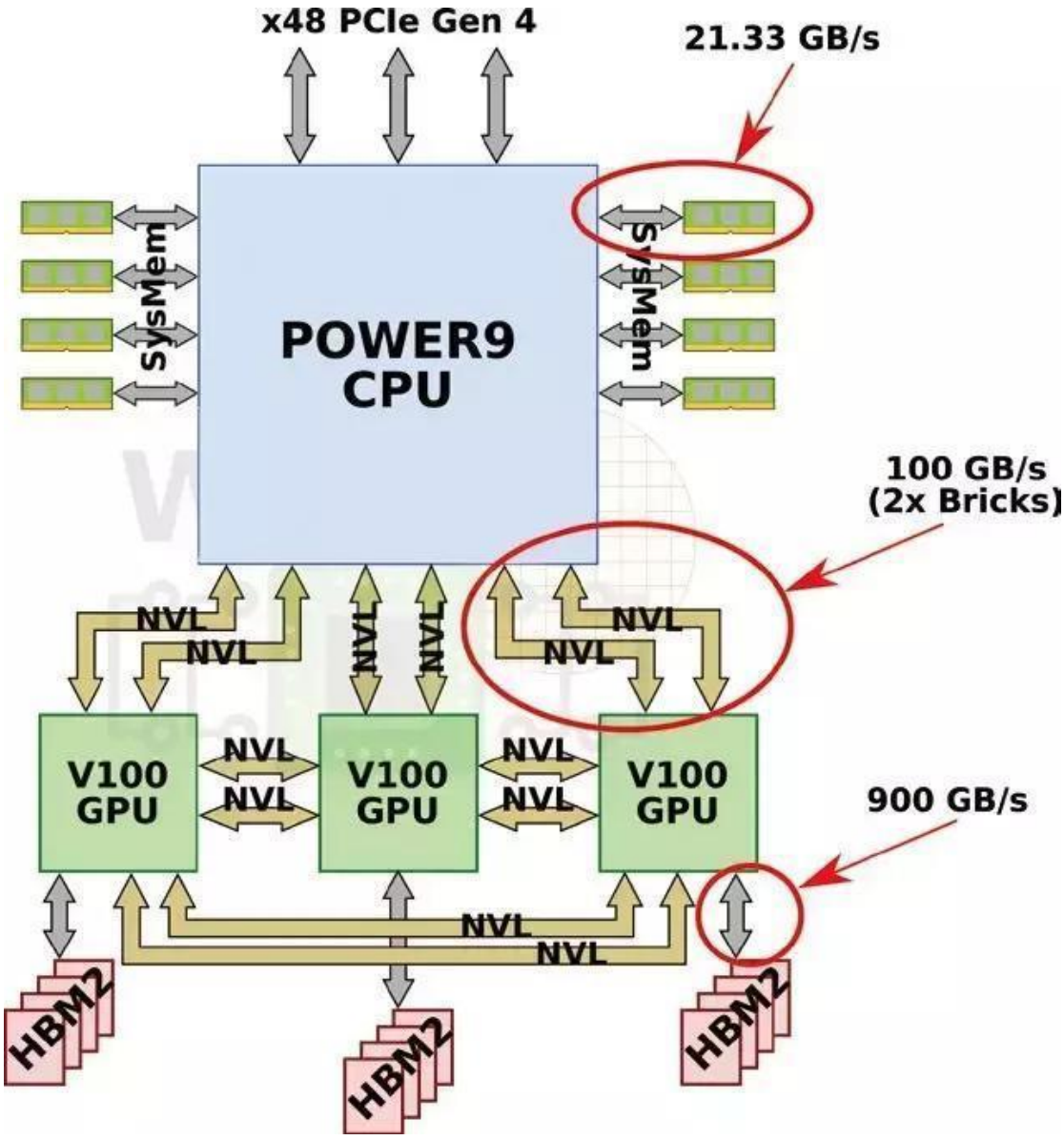
Summit 采用的计算节点型号为 Power System AC922，这是一种 19 英寸的 2U 机架式外壳。从内部布置来看，每个 AC922 内部有 2 个 CPU 插座，满足两颗 Power 9 处理器的需求。每颗处理器配备了 3 个 GPU 插槽，每个插槽使用一块 GV100 核心的计算卡。这样 2 颗处理器就可以搭配 6 颗 GPU。

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。

继续进一步深入 AC922 的话，其主要的技术难题在于 CPU 和 GPU 之间的连接。传统的英特尔体系中，CPU 和 GPU 之间的连接采用的是 PCIe 总线，带宽稍显不足。但是在 Summit 上，由于 IBM Power 9 处理器的加入，因此可以使用更强大的 NVLink 来取代 PCIe 总线。



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。

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。

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。

节点性能表

	按插座计算		以节点计算	
处理器	POWER9	V100	POWER9	V100
数量	1	3	2	6
FLOPS(单精度)	1.081 TFLOPS (22 × 49.12 GFLOPs)	47.1 TFLOPS (3 × 15.7 TFLOPs)	2.161 TFLOPS (2 × 22 × 49.12 GFLOPs)	94.2 TFLOPs (6 × 15.7 TFLOPs)
FLOPS (双精度)	540.3 GFLOPs (22 × 24.56 GFLOPs)	23.4 TFLOPS (3 × 7.8 TFLOPs)	1.081 TFLOPS (2 × 22 × 24.56 GFLOPs)	46.8 TFLOPS (6 × 7.8 TFLOPs)
AI FLOPS	-	375 TFLOPS (3 × 125 TFLOPs)	-	750 TFLOPS (6 × 125 TFLOPs)
内存	256 GiB (DRR4) 8 × 32 GiB	48 GiB (HBM2) 3 × 16 GiB	512 GiB (DRR4) 16 × 32 GiB	96 GiB (HBM2) 6 × 16 GiB
带宽	170.7 GB/s (8 × 21.33 GB/s)	900 GB/s/GPU	341.33 GB/s (16 × 21.33 GB/s)	900 GB/s/GPU

Summit的性能

Summit峰值性能

处理器	CPU	GPU
型号	POWER9	V100
数量	9,216 /2 × 18 × 256	27,648/ 6 × 18 ×256
峰值FLOPS	9.96 PF	215.7 PF
峰值AI FLOPS	N/A	3.456 EF

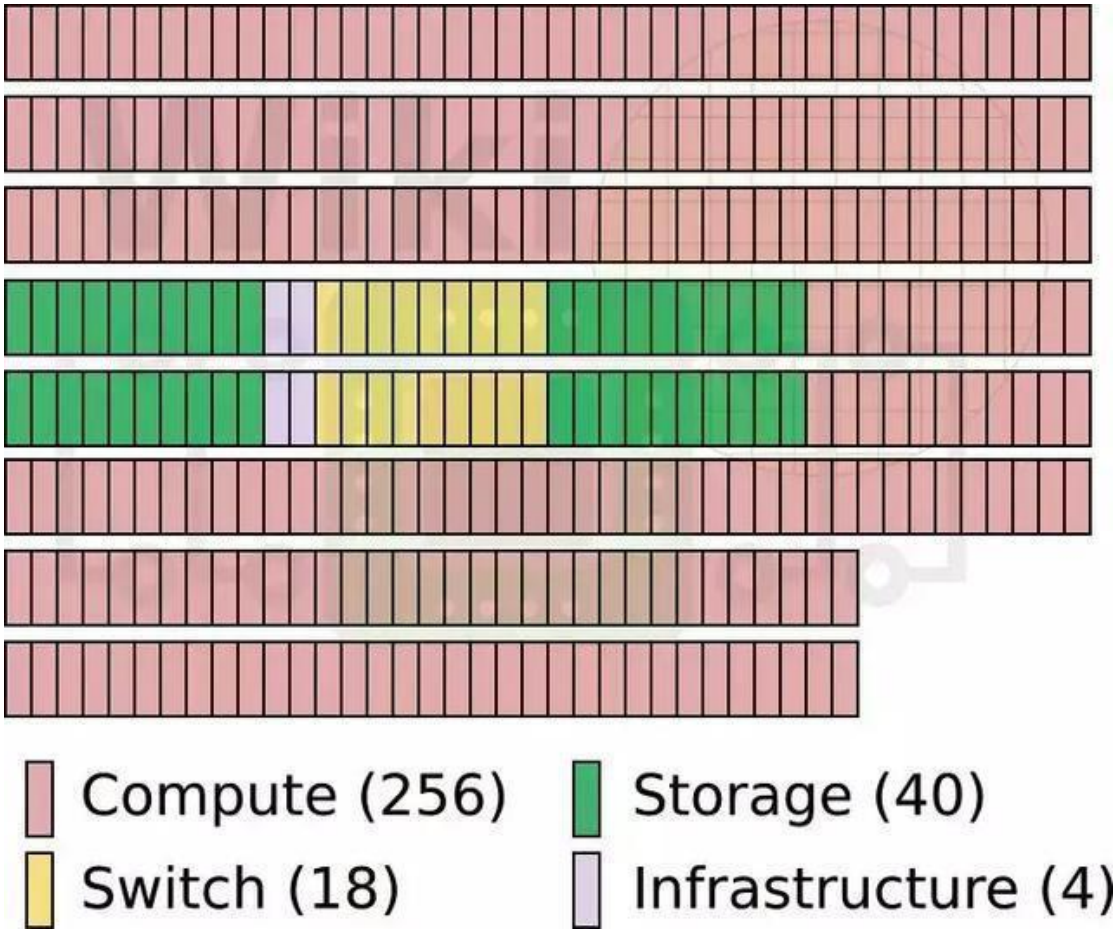
Summit的系统组成

Summit

机架	计算节点	存储节点	交换机
类型	AC922	SSC (4 ESS GL4)	Mellanox IB EDR
数量	256 Racks × 18 Nodes	40 Racks × 8 Servers	18 Racks
功耗	59 kW	38 kW	N/A

五、summit 系统与架构

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。



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。