湖南大學

HUNAN UNIVERSITY

计算机设计

学生姓名	周珍冉
学生学号	201708010610
专业班级	智能 1702
指导老师	吴强
完成日期	2019.12.11

Summit 架构分析

一、 Summit

Summit 超级计算机是 IBM 计划研发的一款超级计算机，其计算性能超过中国 TaihuLight 超级计算机。2018 年 11 月 12 日，新一期全球超级计算机 500 强榜单在美国达拉斯发布，美国超级计算机“Summit”蝉联冠军。2019 年 11 月 18 日，全球超级计算机 500 强榜单发布，美国超级计算机“Summit”以每秒 14.86 亿亿次的浮点运算速度再次登顶。

Summit 超算系统由 4608 台计算服务器组成，每个服务器包含两个 22 核 Power9 处理器（IBM 生产）和 6 个 Tesla V100 图形处理单元加速器（NVIDIA 生产）。Summit 还拥有超过 10PB 的存储器，配以快速、高带宽的路径以实现有效的数据传输。

凭借每秒高达 20 亿亿次(200PFlops)的浮点运算速度峰值，Summit 的威力将是 ORNL 之前排名第一的系统 Titan 的 8 倍，相当于普通笔记本电脑运算速度的 100 万倍，比之前位于榜首的中国超级计算机“神威·太湖之光”峰值性能（每秒 12.5 亿亿次）快约 60%。

Summit 超级计算机采用 IBM Power9 微处理器和 NVIDIA Volta GPU 进行数学协同处理。Summit 的前身 Titan 超级计算机，拥有超过 18000 个节点，而 Summit 将有约 3400 个节点。每个节点将拥有至少 500GB 相干内存，以及 800GB 非易失性内存。

二、 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。

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。

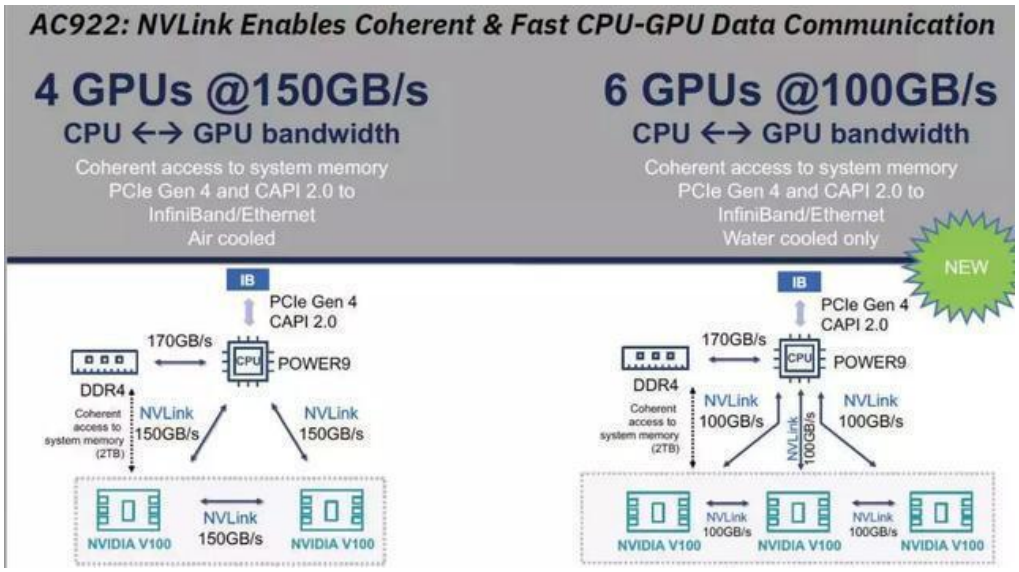
计算节点——2CPU+6GPU

Summit 采用的计算节点型号为 Power System AC922，之前的研发代号为 Witherspoon，简称为 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，总计给每个 CPU 可以带来 256GB、107.7GB/s 的内存容量和带宽。GPU 方面，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。

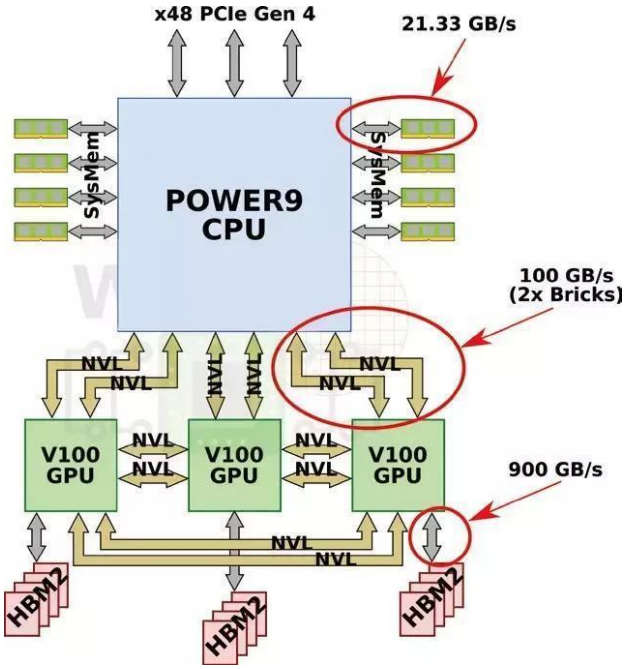
进一步深入 AC922，其主要的技术难题在于 CPU 和 GPU 之间的连接。传统的英特尔体系中，CPU 和 GPU 之间的连接采用的是 PCIe 总线，带宽稍显不足。但是在 Summit 上，由于 IBM Power 9 处理器的加入，因此可以使用更强大的 NVLink 来取代 PCIe 总线。

NVLink 2.0 连接方案：



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。

WikiChip 机构制作的 Summit 内部 NVLink 2.0 连接示意图：

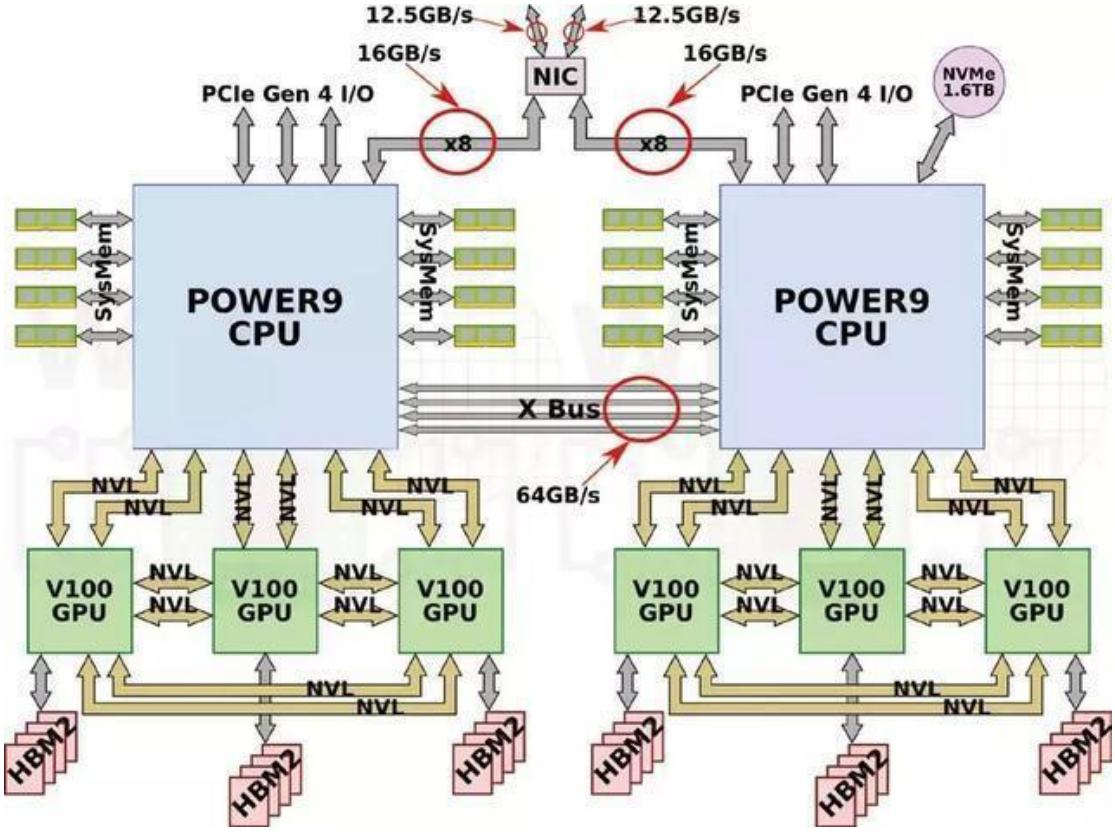


和 CPU 类似，GV100 GPU 也有 6 个 NVLink 2.0 通道，同样也分为 3 组，其中一组连接 CPU，另外 2 组连接其他两颗 GPU。和 CPU-GPU 之间的链接一样，

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。

WikiChip 机构制作的 Summit 内部 CPU 间通讯结构示意图：



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。

三、 节点性能情况

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。每个节点都配备了 1.6TB 的 NVMe SSD 和一个 Mellanox Infiniband EDR 网络接口。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。

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

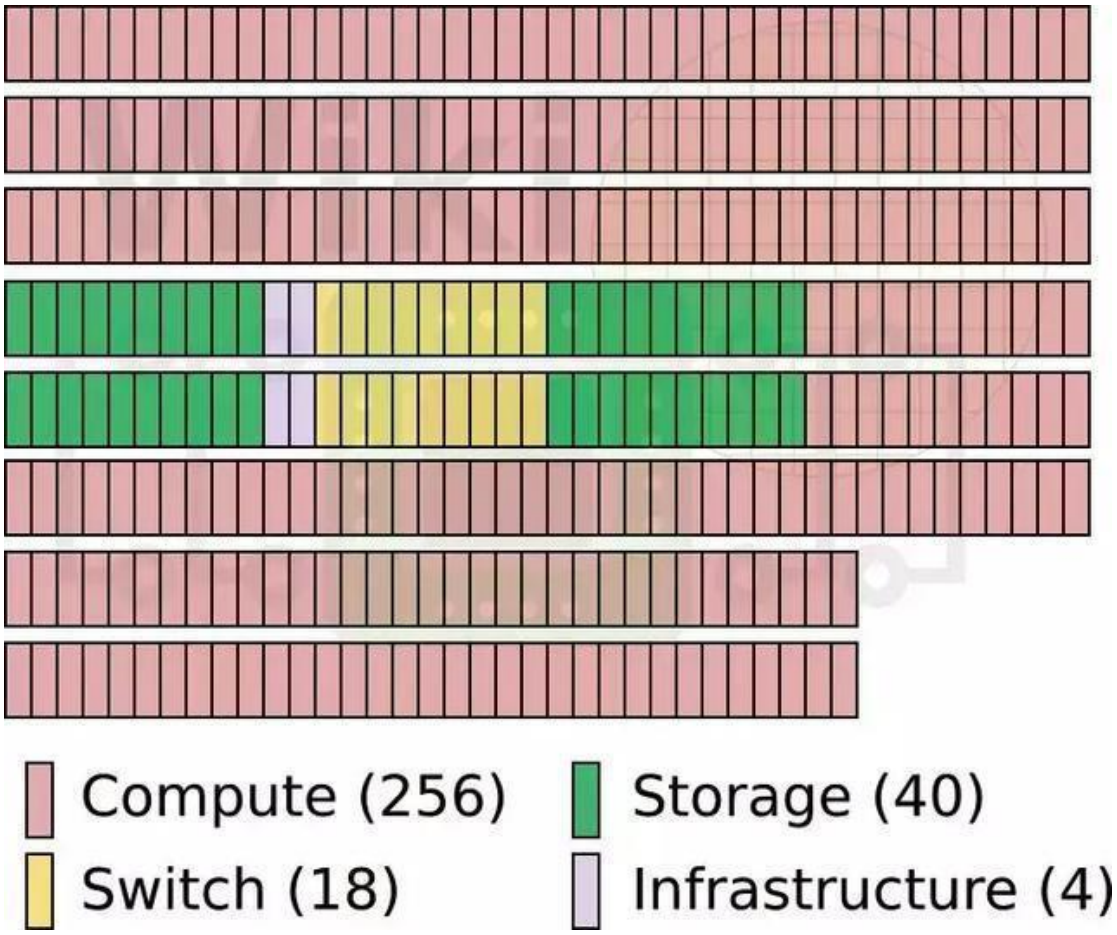
四、 Summit 机架和系统

1. 机架

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。

WikiChip 机构制作的 Summit 的系统结构布局图：



2. 系统

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。

五、 异构超算

A. 同构计算

同构计算是使用相同类型指令集和体系架构的计算单元组成系统的计算方

式。同构超算只单纯使用一种处理器，日本超算“京”只采用的处理器是富士通制造的 SPARC64 VIIIfx，神威蓝光只采用了 8704 片申威 1600，Mira 和 Sequoia（Sequoia 采用了约 160 万个 A2 处理核心），就只采用了 PowerPC A2 处理器，这些都没有采用 GPU 或众核芯片等加速器。日本的京，IBM 的 Mira 和 Sequoia 和中国的神威蓝光都是同构超算的代表。

B. 异构计算

异构计算主要是指使用不同类型指令集和体系架构的计算单元组成系统的计算方式。常见的计算单元类别包括 CPU、GPU 等协处理器、DSP、ASIC、FPGA 等。异构计算是一种并行和分布式计算，它或是用能同时支持 simd 方式和 mimd 方式的单个独立计算机，或是用由高速网络互连的一组独立计算机来完成计算任务。具体来说，异构计算是在运算中既使用处理器，又使用 GPU 或众核芯片等加速器。以美国泰坦和中国天河 2 号为例，泰坦有 18688 个运算节点，每个运算节点由 1 个 16 核心 AMD Opteron 6274 处理器和 1 个 NVIDIA Tesla K20 加速器组成，共计 299008 个运算核心；天河 2 号有 16000 个计算节点，每个节点由 2 片 Intel 的 E5 2692 和 3 片 Xeon PHI 组成，共使用了 32000 片 Intel 的 E5 2692 和 48000 片 Xeon PHI；天河 1A 使用了 14336 片 Intel Xeon X5670 处理器和 7168 片 NVIDIA Tesla M2050 高性能计算卡。除了泰坦和天河 2 号和天河 1A 之外，曙光 6000 也是采用了异构计算架构。

六、 总结心得

Summit 采用异构方式，这是一种并行和分布式计算，既使用处理器，又使用 GPU 或众核芯片等加速器。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。每个机架中安置了 18 个计算节点和 Mellanox IB EDR 交换器。一个完整的计算节点拥有 2 颗 22 核心的 Power 9 处理器，总计 44 颗物理核心，配备了 1.6TB 的 NVMe SSD 和一个 Mellanox Infiniband EDR 网络接口。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。