HUNAN UNIVERSITY

计算机设计

学 生 姓 名 李 博 文

学 生 学 号 201708010602

专 业 班 级 智 能 1702

指 导 老 师 吴 强

论 文 题 目 Summit架 构 分 析

一、summit简介

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。预计将在2018年初提供给美国能源部橡树岭国家实验室，计算性能比原定指标提升四分之一以上。

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。

这台让美国重夺世界第一的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%。

为了给客户提供很高的I/O吞吐量，率很高，节点将使用Mellanox公司的双轨InfiniBand EDR连接以无阻塞胖树架构互联。

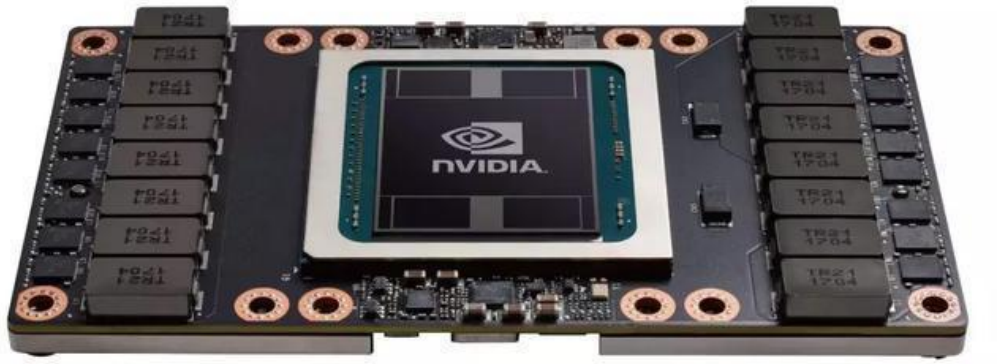
Summit超级计算机采用IBM Power9微处理器和NVIDIA Volta GPU进行数学协同处理。Summit的前身Titan超级计算机，拥有超过18000个节点，而Summit将有约3400个节点。每个节点将拥有至少500GB相干内存，以及800GB非易失性内存。

Summit超级计算机原定计算性能是150petaflops，交付性能达到200petaflops。中国的TaihuLight超级计算机性能指标是93 petaflops，峰值性能是124.5petaflops。IBM这款超级计算机交易据说价值3.25亿美元。

二、Summit架构

节点、机架和整体

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。



▲SXM2接口的Tesla V100

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。

计算节点

2CPU+6GPU

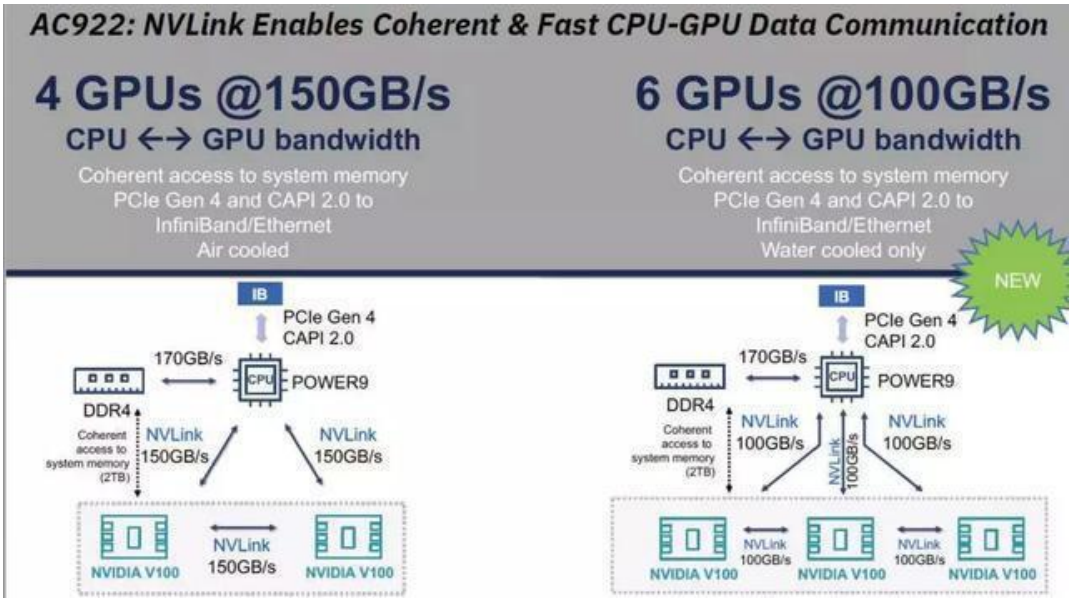
Summit采用的计算节点型号为Power System AC922，之前的研发代号为Witherspoon，后文我们将其简称为AC922，这是一种19英寸的2U机架式外壳。从内部布置来看，每个AC922内部有2个CPU插座，满足两颗Power 9处理器的需求。每颗处理器配备了3个GPU插槽，每个插槽使用一块GV100核心的计算卡。这样2颗处理器就可以搭配6颗GPU。

▲Summit的一个计算节点，以及其内部设备

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。

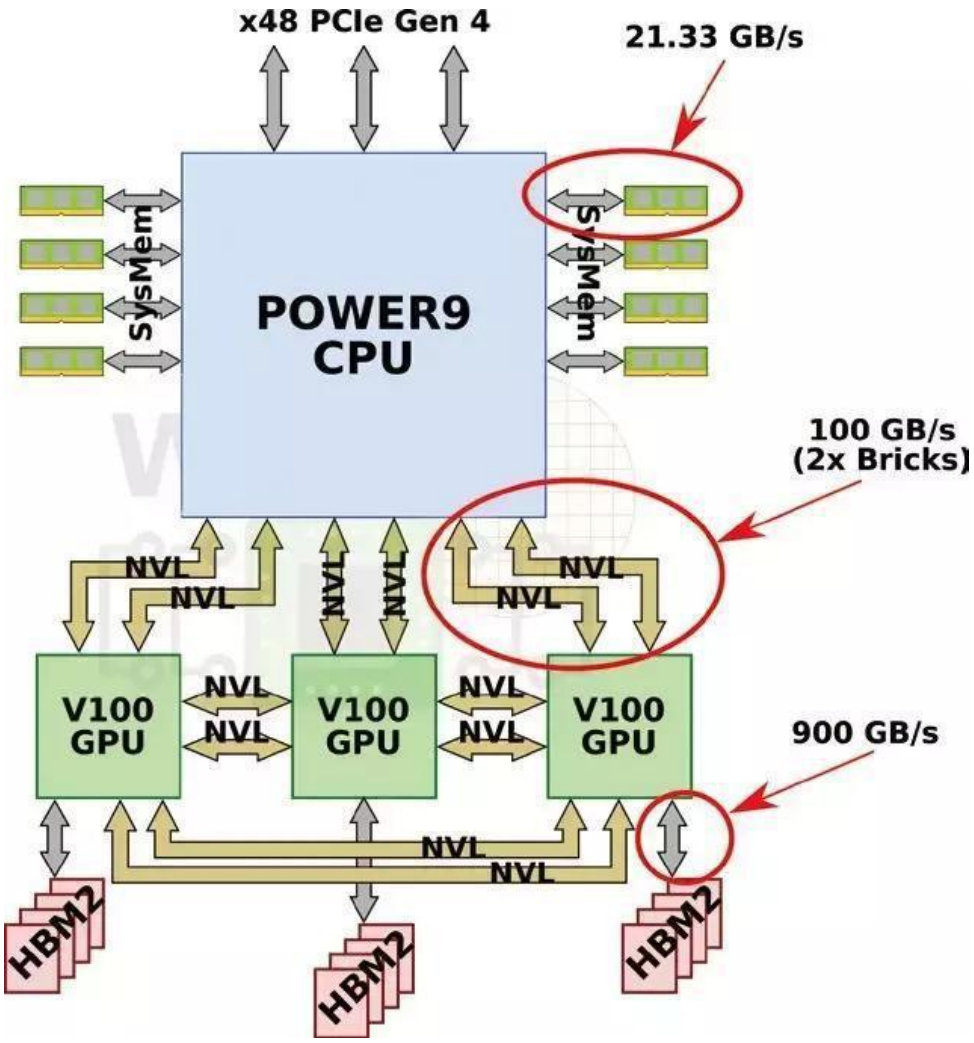
风生水起的NVLink 2.0

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。本刊在之前的文章中也曾深入分析过NVLink的相关技术，在这里就不再赘述。



▲NVLink 2.0在民用市场无法施展拳脚，但是在超算市场可谓风生水起，图为IBM展示的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。



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

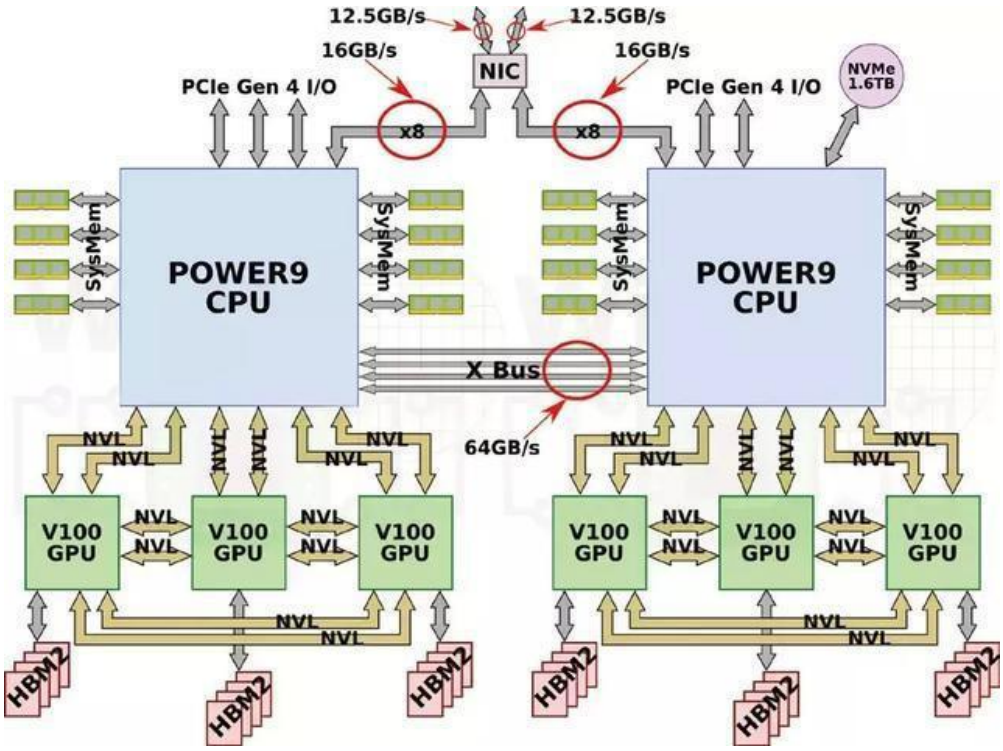
和CPU类似，GV100 GPU也有6个NVLink 2.0通道，同样也分为3组，其中一组连接CPU，另外2组连接其他两颗GPU。和CPU-GPU之间的链接一样，GPU与GPU之间的连接带宽也是100GB/s。

CPU之间的通讯

X总线登场

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。



▲国外WikiChip机构制作的Summit内部CPU间通讯结构示意图

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

请注意，这里的数值和最终公开的数据存在一些差异，其主要原因是公开数据的性能只包含GPU部分，这也是大多数浮点密集型应用可以实现的最高性能。当然，如果包含CPU的话，Summit本身的峰值性能将超越220PFlops。

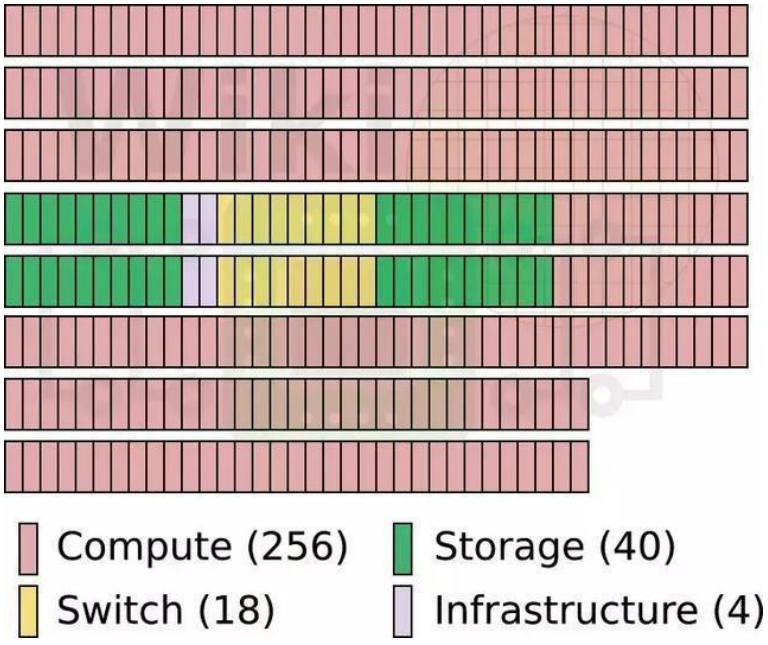
Summit的性能		
Summit峰值性能		
处理器	CPU	GPU
型号	POWER9	V100
数量	9,216 /2 × 18 × 256	27,648/ 6 × 18 ×256
峰值FLOPS	9.96 PF	215.7 PF
峰值AI FLOPS	N/A	3.456 EF

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

除了CPU和GPU外，每个节点都配备了1.6TB的NVMe SSD和一个Mellanox Infiniband EDR网络接口。

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。

▲一个开放的机架有18个计算节点，开关在中部和顶部

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。

目前业内报告的Summit系统性能依旧偏向保守，当然，最好性能并不是最有意义的，实际的负载性能最为重要。橡树岭国家实验室在初步测试Summit针对基因组数据的性能时，达到了1.88 exaops的混合精度性能，这个测试主要是用的是GV100的张量核心矩阵乘法，这也是迄今为止报告的最高性能。