# Summit架构分析

超级计算机作为人类顶尖技术的最佳代表，在全球各个领域都起着举足轻重的作用，一套优秀的超算能够极大地提高科研效率甚至推动一个行业的发展进步。我国近年来在超级计算机领域频频发力，推出了诸如天河系列、“神威太湖之光”等多款超级计算机，甚至长期独占鳌头笑傲全球。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](https://baike.baidu.com/item/%E6%A9%A1%E6%A0%91%E5%B2%AD%E5%9B%BD%E5%AE%B6%E5%AE%9E%E9%AA%8C%E5%AE%A4" \t "https://baike.baidu.com/item/Summit%E8%B6%85%E7%BA%A7%E8%AE%A1%E7%AE%97%E6%9C%BA/_blank)，计算性能比原定指标提升四分之一以上。

1. 工作原理

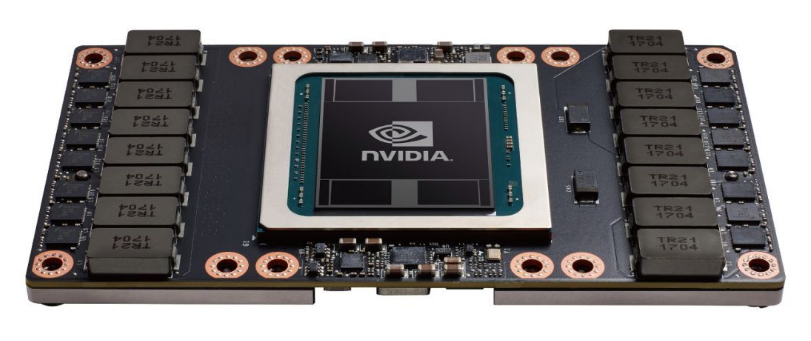
这台让美国重夺世界第一的Summit超算系统由4608台计算服务器组成，每个服务器包含两个22核Power9处理器（IBM生产）和6个Tesla V100图形处理单元加速器（NVIDIA生产）。Summit还拥有超过10PB的存储器，配以快速、高带宽的路径以实现有效的数据传输。

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。

二．节点、机架和整体

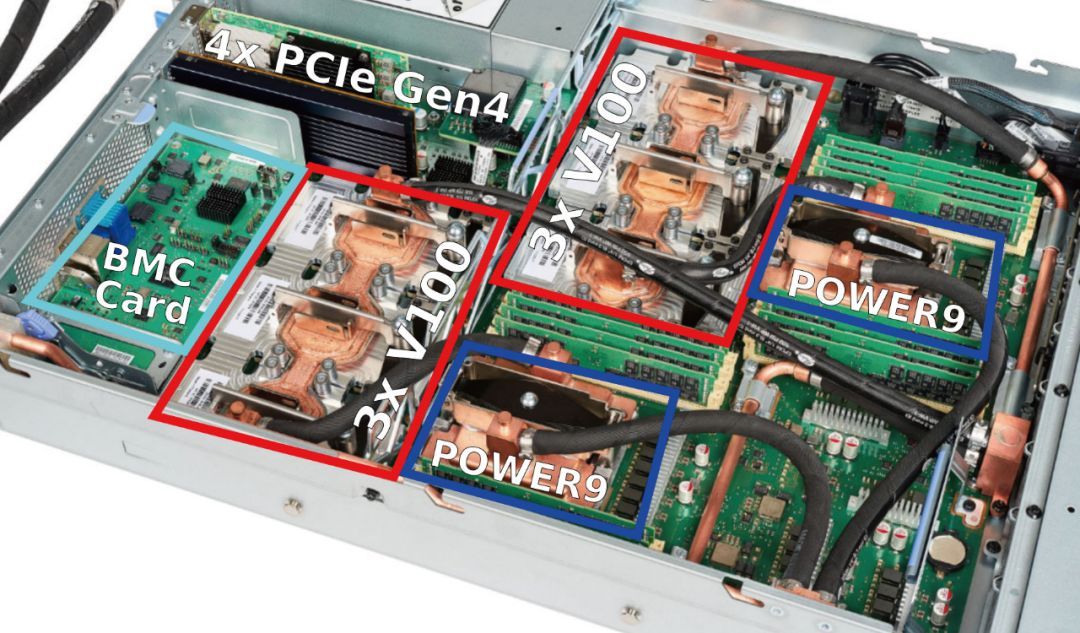
从硬件架构方面来看，Summit依旧采用的是异构方式，其主CPU来自于IBM Power 9，22核心，主频为3.07GHz，总计使用了103752颗，核心数量达到2282544个。GPU方面搭配了27648块英伟达Tesla V100计算卡，总内存为2736TB，操作系统为RHEL 7.4。从架构角度来看，Summit并没有在超算的底层技术上予以彻底革新，而是通过不断使用先进制程、扩大计算规模来获得更高的性能。



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。

三．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。

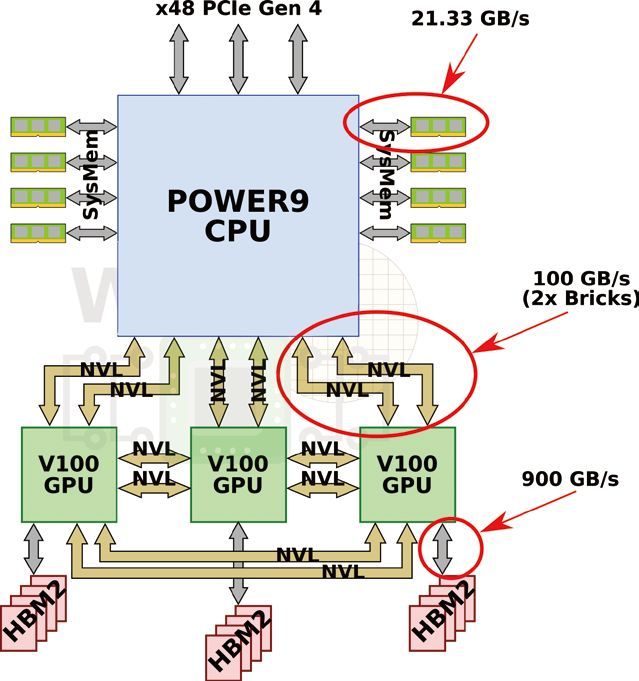


Summit的一个计算节点，以及其内部设备：内存方面，每颗处理器设计了8通道内存，每个内存插槽可以使用32GB DDR4 2666内存，这样总计可以给每个CPU可以带来256GB、107.7GB/s的内存容量和带宽。GPU方面，它没有使用了传统的PCIe插槽，而是采用了SXM2外形设计，每颗GPU配备16GB的HBM2内存，对每个CPU-GPU组而言，总计有48GB的HBM2显存和2.7TBps的带宽。

四．NVLink 2.0

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：

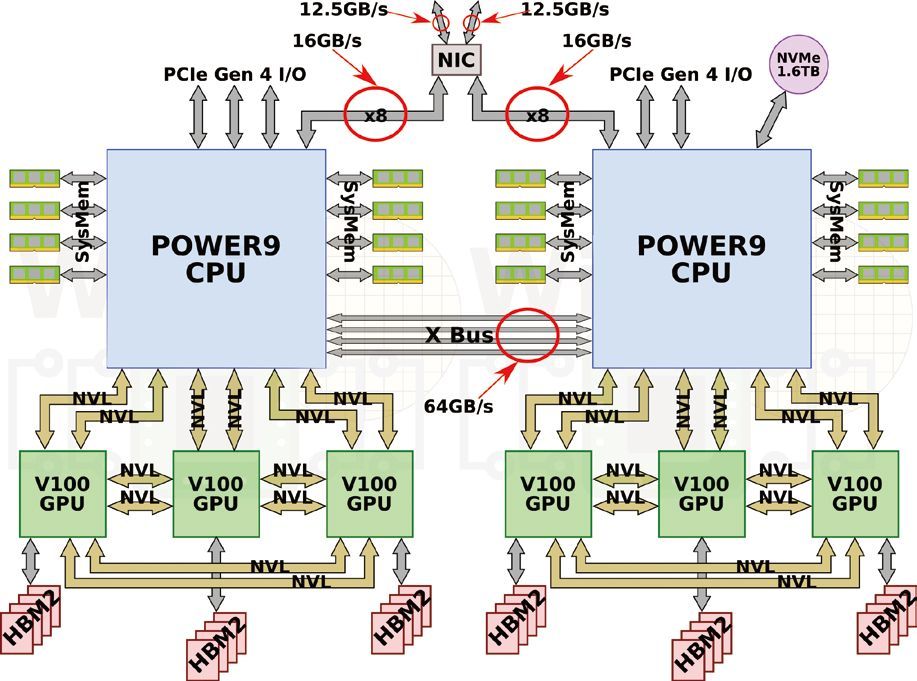


和CPU类似，GV100 GPU也有6个NVLink 2.0通道，同样也分为3组，其中一组连接CPU，另外2组连接其他两颗GPU。和CPU-GPU之间的链接一样，GPU与GPU之间的连接带宽也是100GB/s。

五．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。

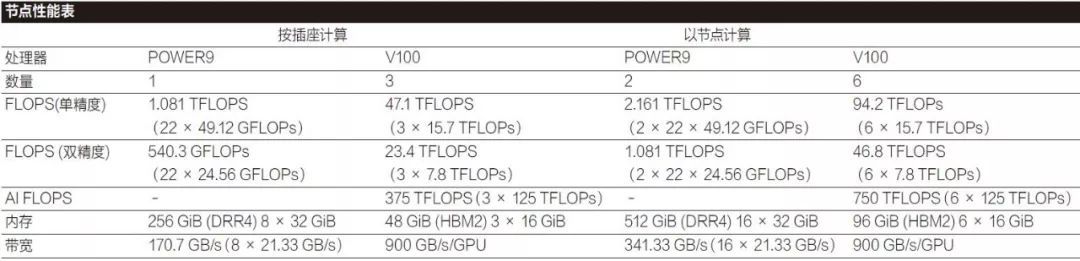
国外WikiChip机构制作的Summit内部CPU间通讯结构示意图：

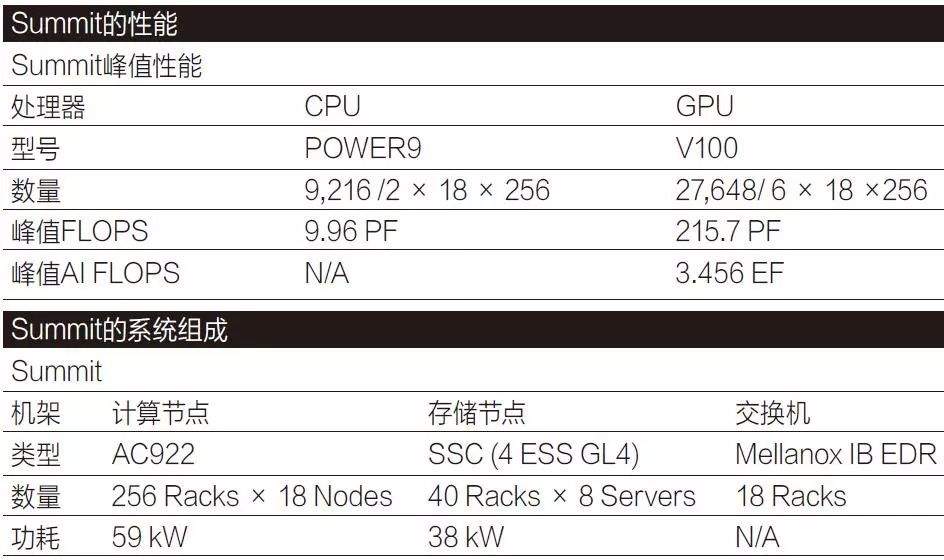


另外在CPU的对外通讯方面，每一个节点拥有4组向外的PCIe 4.0通道，包括两组x16（支持CAPI），一组x8（支持CAPI）和一组x4。其中2组x16通道分别来自于两颗CPU，x8通道可以从一颗CPU中配置，另一颗CPU可以配置x4通道。其他剩余的PCIe 4.0通道就用于各种I/O接口，包括PEX、USB、BMC和1Gbps网络等。

六．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。





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。

1. 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。

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。

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。

1. 性能数据

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。