Summit 架构分析

一、概述

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。

从现实情况来看，除了我们国家，美国在超算领域的实力依旧不可小觑。在2018 年的 6 月，美国能源部在橡树岭国家实验室正式宣布了全新的超级计算机——Summit。

2018 年 6 月 25 日，TOP500 组织发布了第 51 届全球超级计算机排行榜。在这个榜单中，来自于美国橡树岭国家实验室，受美国能源部资助的 Summit 暂居超级计算机榜首。

二、Summit 超级计算机架构分析

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。

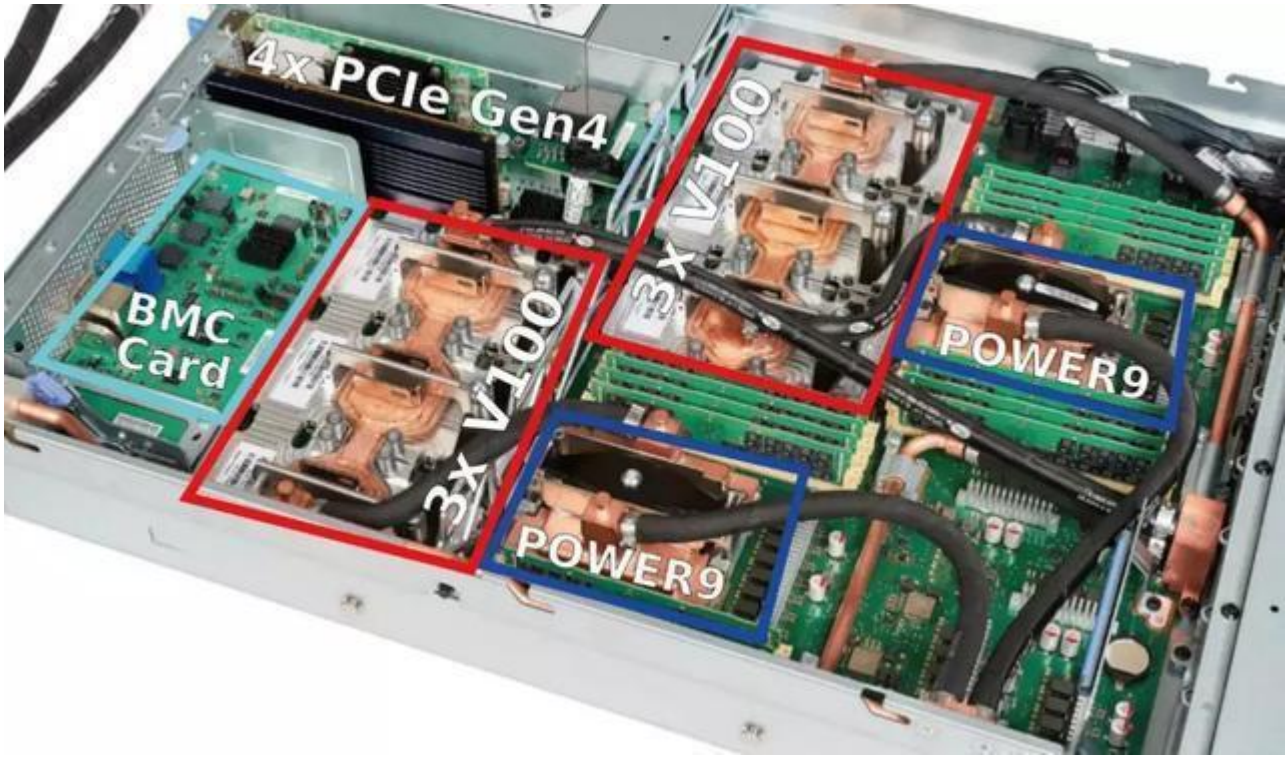
▲SXM2 接口的 Tesla V100。

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。



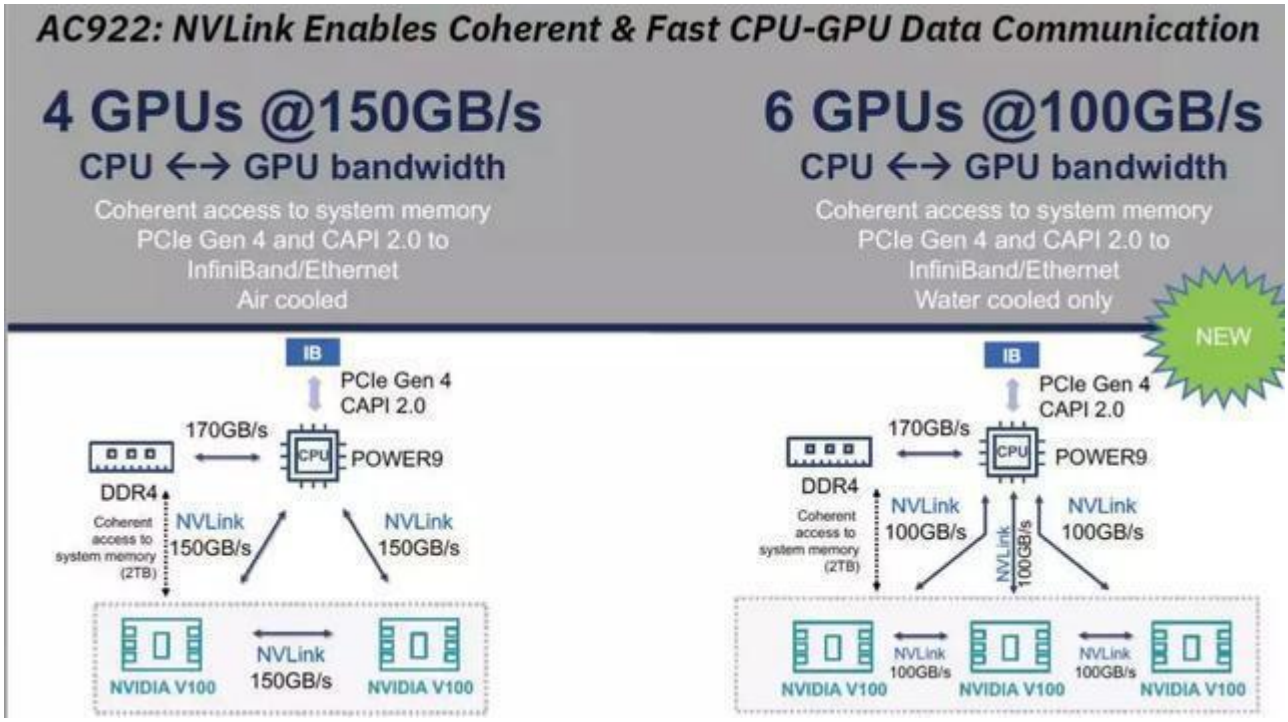
▲Summit 的一个计算节点，以及其内部设备。

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。

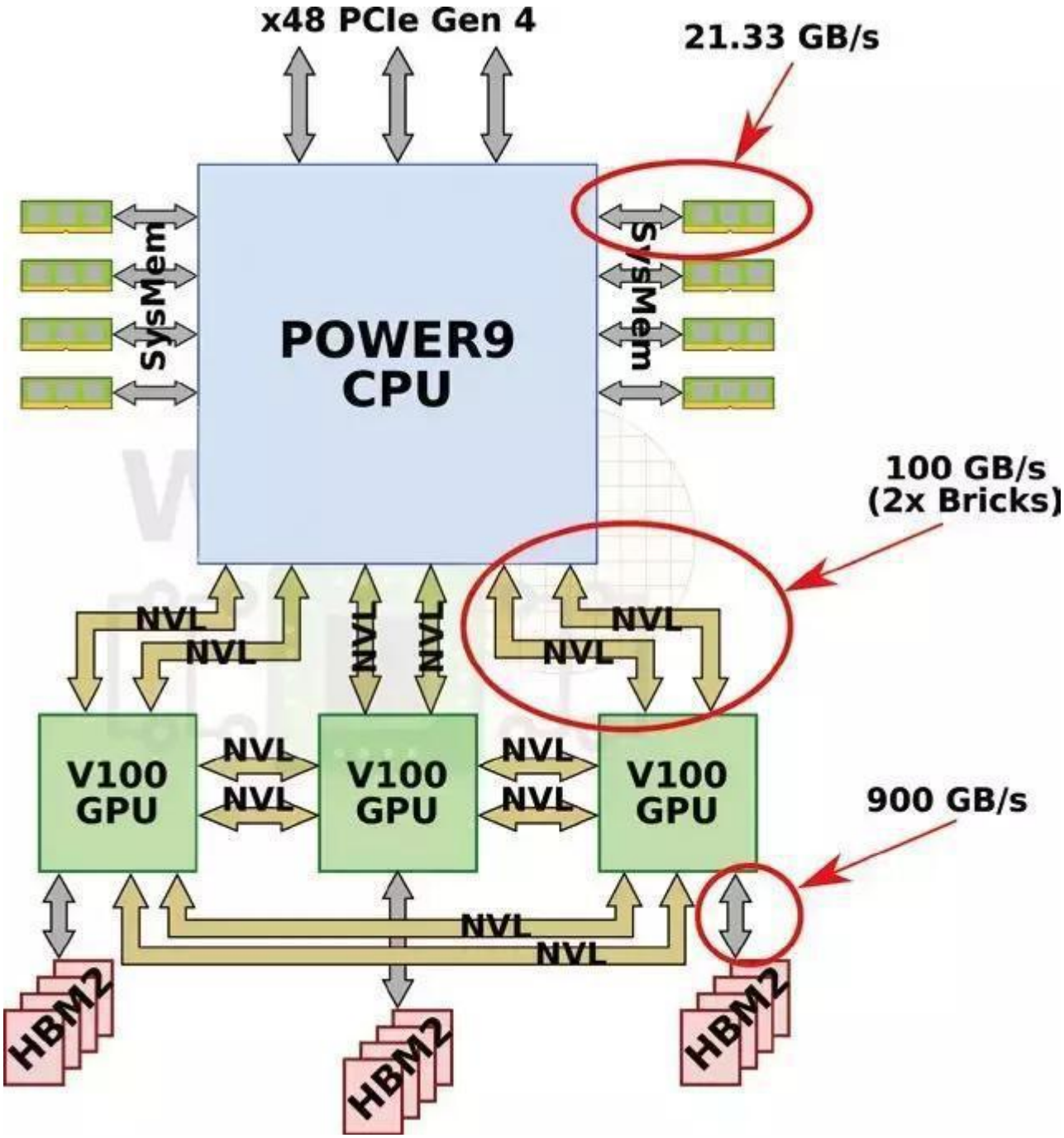
风生水起的 NVLink 2.0

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。



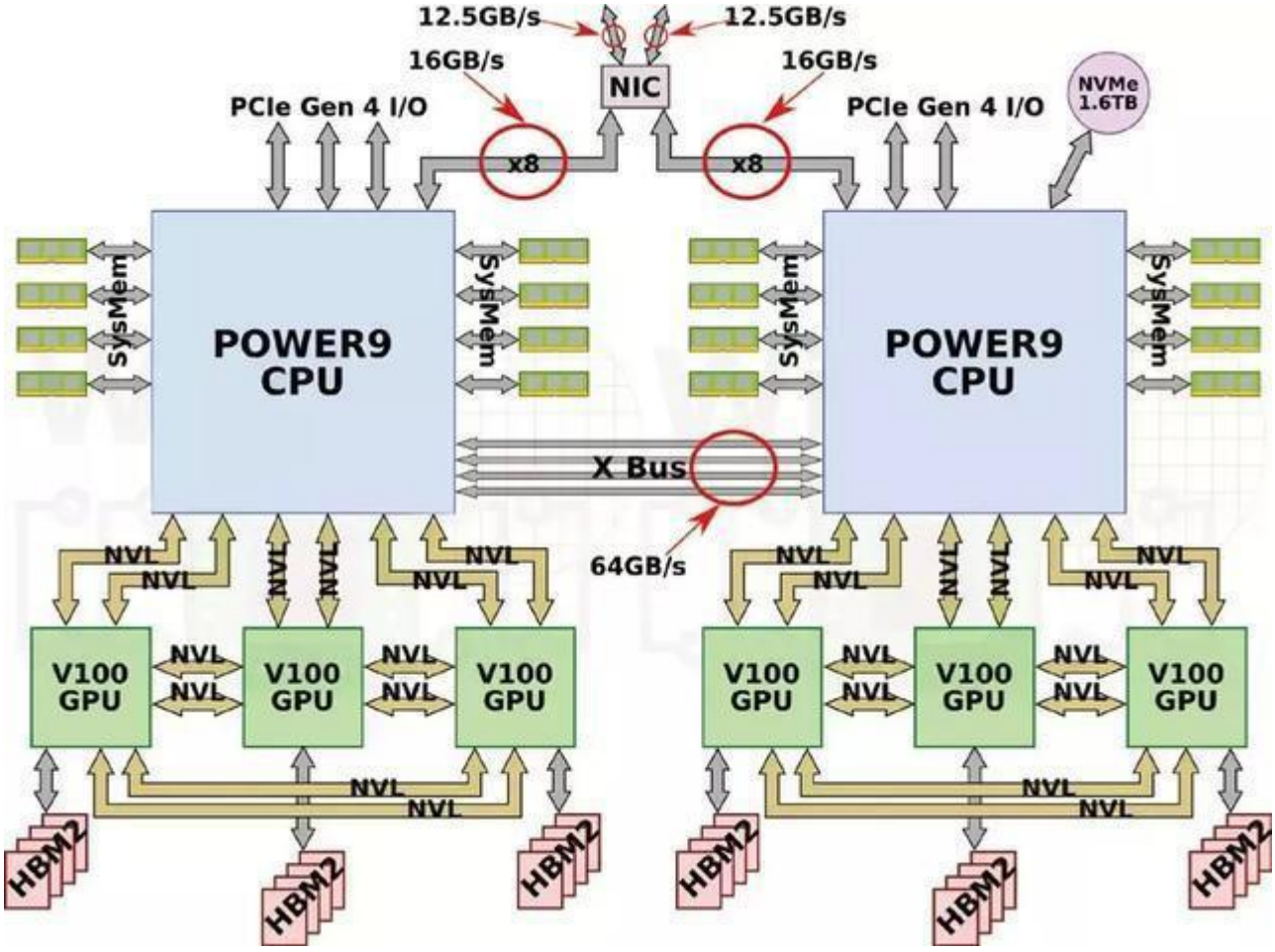
▲国外 WikiChip 机构制作的 Summit 内部 NVLink 2.0 连接示意图。

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。

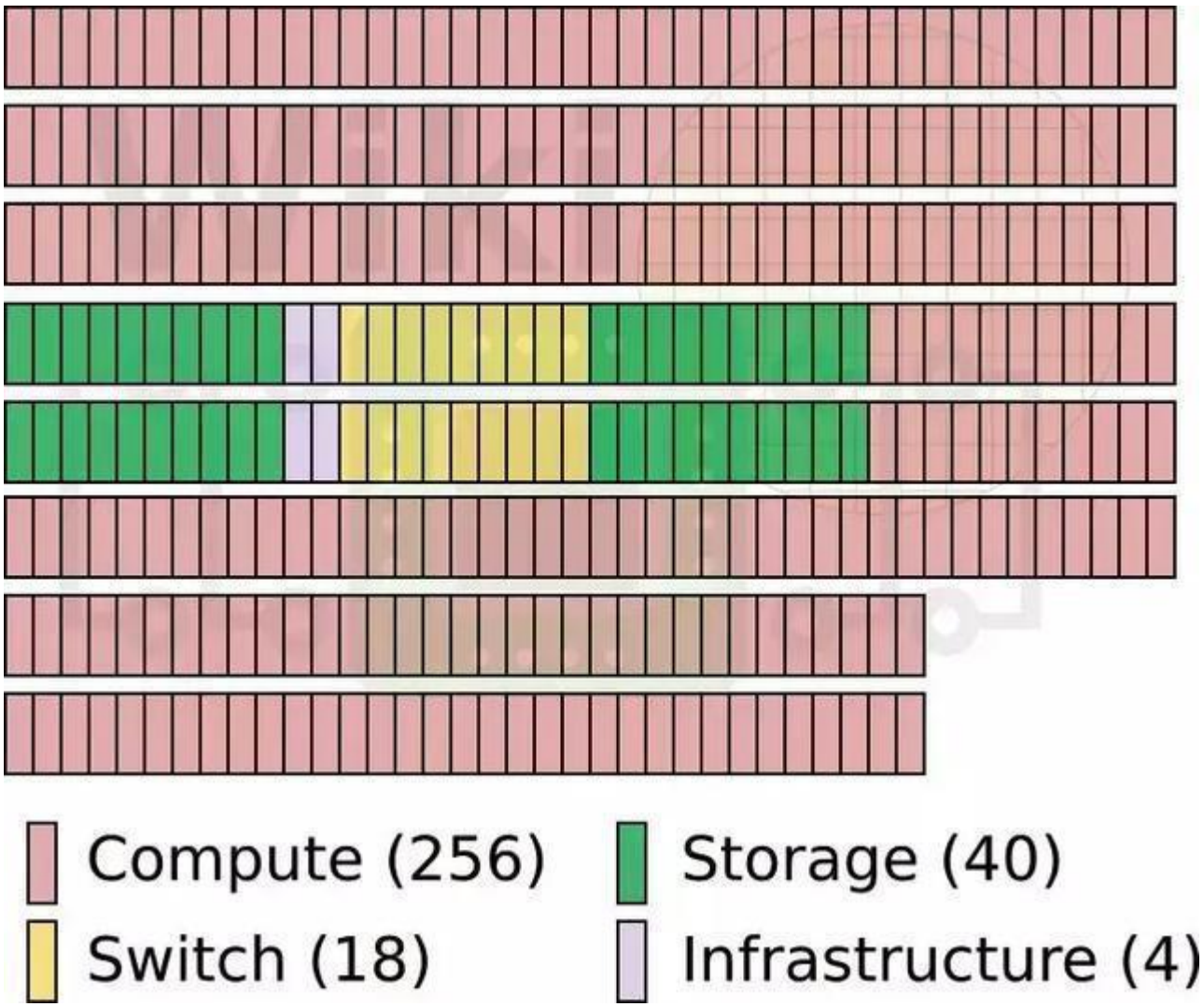
Summit的性能		
Summit峰值性能		
处理器	CPU	GPU
型号	POWER9	V100
数量	9,216 /2 × 18 × 256	27,648/ 6 × 18 ×256
峰值FLOPS	9.96 PF	215.7 PF
峰值AI FLOPS	N/A	3.456 EF

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。

三、参考资料

• 百万兆级计算机 Summit 是什么？Summit 介绍：
http://www.ccy.com.cn/c/2018071673642.html
• 性能突破 200PFLOPS！世界第一超级计算机 Summit 解析：
https://baijiahao.baidu.com/s?id=1608375900787917940&wfr=spider&for=pc