Summit 架构分析

1. 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。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“顶点”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“顶点”以每秒 14.86 亿亿次的浮点运算速度再次登顶。

2. 性能介绍

根据超算 Top500 排行的数据，Summit 超级计算机的峰值浮点性能为 187.7PFlops，Linpack 浮点性能为 122.3PFlops，功耗为 8805.5kW。相比之下，我国的神威太湖之光的峰值浮点性能为 125.4PFlops，Linpack 浮点性能为 93.0PFlops，功耗为 15371kW。第三到第六名分别是美国的 Sierra、中国天河 2A（升级了全新的 Matrix-2000 处理器，移除了之前的 Xeon Phi，性能提升至 61.4PFLops Linpark）、日本的 ABCI 以及瑞士的 Piz Daint。

除了 TOP500 排行榜外，在 HPCG 排行榜中，Summit 仍然暂居第一名的位置，HPCG 性能为 2925.75TFlops/s。第二名到第五名分别是美国的 Sierra、日本的 K、美国的 Trinity、瑞士的 Piz Daint。

Summit 性能参数：

每秒 20 亿亿次计算

250PB 的存储容量

9,216 个 IBM POWER9 CPU（中央处理器）

计算节点之间数据传输可达每秒 25GB

27,648 个 NVIDIA Tesla GPU（英伟达图形处理器）

3. Summit 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。

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。

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。

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。