Submit 架构分析

一、 Submit 简介

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。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。在 HPCG 排行榜中，Summit 仍然暂居第一名的位置，HPCG 性能为 2925.75TFlops/s。

二、 Summit 的硬件配置

从硬件架构方面来看，Summit 依旧采用的是异构方式。

1、概览

主 CPU 是 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 搭配了 27648 块 NVIDIA Tesla V100 计算卡，总内存为 2736TB。

Submit 有三种类型的节点，分类的标准就在于节点的用途。

节点类型	描述
Login	连接到 Summit 时，就会被置于登录节点上。在此类节点上可以编写/编辑/编译代码，管理数据，提交作业等。不可以在登录节点启动并行作业或者上运行线程化作业。登录节点是许多用户同时使用的共享资源。
Launch	当批处理脚本（或交互式批处理作业）开始运行时，它将在启动节点上执行。作业脚本中的所有命令（或在交互式作业中运行的命令）将在启动节点上运行。与登录节点一样，它们是共享资源，因此不能在启动节点上运行多处理器/线程程序。从启动节点启动 *jsrun* 命令是可以的。
Compute	Summit 上的大多数节点是计算节点。这些是您并行作业的执行位置。可通过 *jsrun* 命令访问它们。

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。

2、计算节点结构

▲Summit 的一个计算节点，以及其内部设备。

①Summit 采用的计算节点型号为 Power System AC922，后文我们将其简称为 AC922，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 NVIDIA Tesla V100 核心的计算卡。即每个计算节点都包含两个 IBM POWER9 处理器和六个 NVIDIA Volta V100 GPU 。

②内存：

每颗 CPU 处理器设计了 8 通道内存，每个内存插槽可以使用 32GB DDR4 2666MHz 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 采用了 SXM2 外形设计，每颗 GPU 配备 16GB 的 HBM2 内存，每个 HBM2 内存可提供 900 GB / s 的带宽。对每个节点而言，总计有 96GB 的 HBM2 显存和 5.4TBps/s 的带宽。

③非易失性存储器：

每个节点都有 1.6TB 的非易失性存储器，可用作突发缓冲区。

3、机架和系统

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，并使用非阻塞胖树拓扑（non-blocking fat-tree topology）交换结构，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。

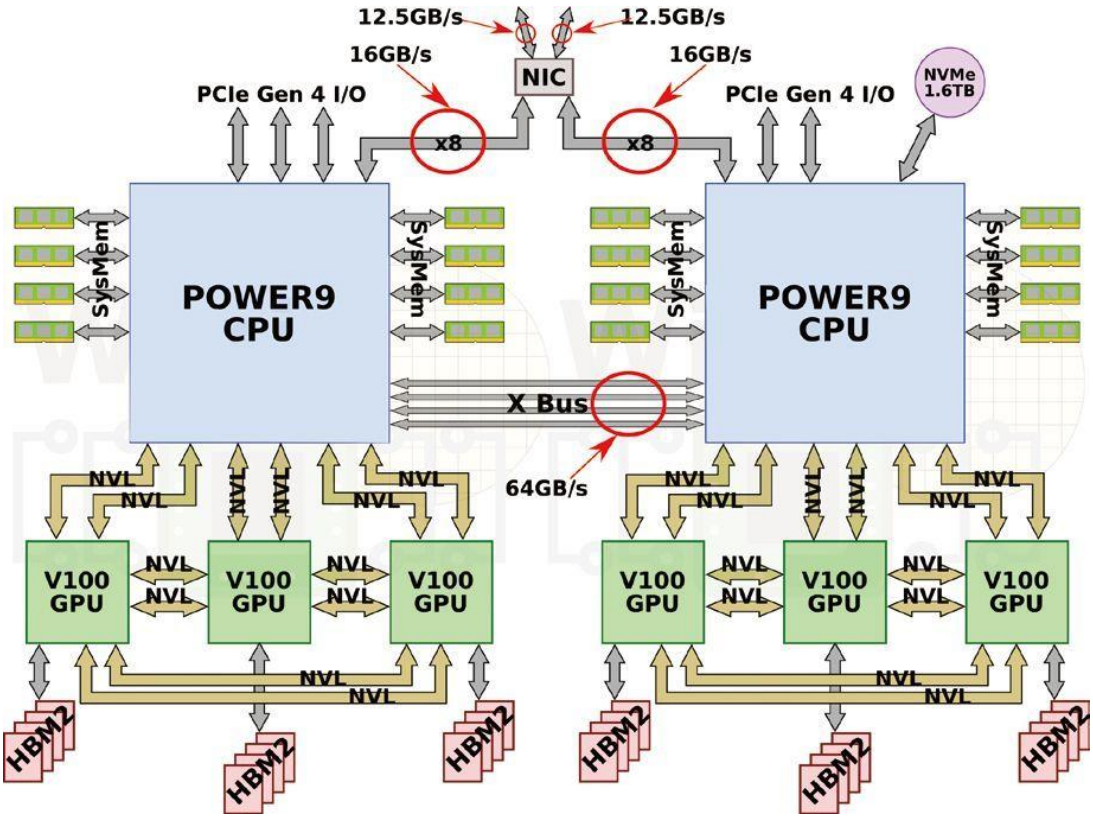
4、其他配置

POWER9 处理器和 V100 加速器采用冷板技术冷却。其余的组件通过更传统的方法进行冷却，尽管排气先经过柜后热交换器，然后再释放回室内。冷板和热交换器都使用中温水运行，与旧系统使用的冷水相比，中温水对中心的维护更具成本效益。

三、 Summit 节点内处理器通信方式

1、POWER9 处理器通信

两个 CPU 间通讯方面，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。

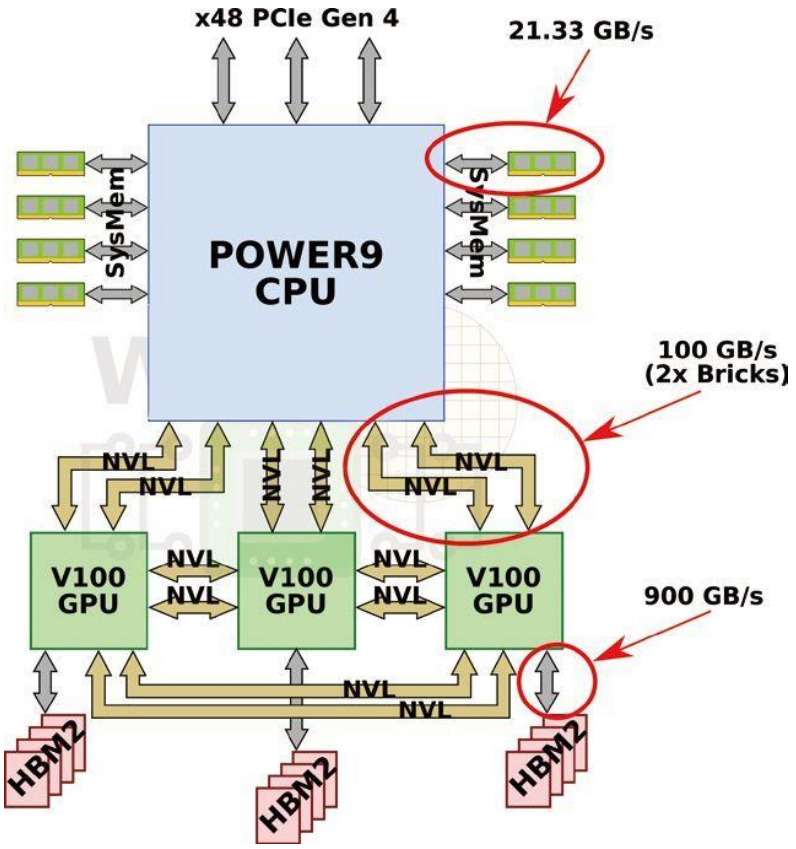


▲国外 WikiChip 机构制作的 Summit 内部 CPU 间通讯结构示意图

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。

2、POWER9 处理器与 NVIDIA Tesla V100 GPU 通信

③CPU 与 GPU 通信方面，在 Summit 上，CPU 和 GPU 之间的连接采用的是 NVLink 来取代 PCIe 总线。



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

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。

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。

四、 Summit 文件系统以及操作系统

Summit 已连接到 IBM Spectrum Scale™文件系统，该文件系统提供 250PB 的存储容量，峰值写入速度为 2.5 TB / s。Summit 还可以访问中心范围内基于 NFS 的文件系统（该文件系统提供用户和项目宿主区域），并可以访问中心的高性能存储系统（HPSS），用于用户和项目档案存储。

Summit 运行的操作系统是 Red Hat Enterprise Linux（RHEL）7.6 版本。

参考资料：

http://www.sohu.com/a/246227428_616364

https://docs.olcf.ornl.gov/systems/summit_user_guide.html