Summit架构分析

智能1702 201708010528 莫诗幻

1. summit简介

1、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“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、发展历程：

2013年6月起，中国超算长期蝉联第一，美国的超级计算机再未问鼎全球超算top500榜单。而Summit的问世让这一宝座终易主。

1988年，ORNL的科学家们完成了首次G浮点（gigaflops）运算，1998年完成了首次T浮点（teraflops）运算，2008年完成了首次P浮点（petaflops）运算，2018年又完成了首次exaops计算。

超级计算机Summit的发布让美国向“2021年交付E级超算”的目标又迈进了一步。它将在能源研究、科学发现、经济竞争力和国家安全等方面带来深远影响，助力科学家们在未来应对更多新的挑战，促进科学发现和激发科技创新。

中国计划于2020年推出首台E级超算；美国能源部启动了“百亿亿次计算项目（Exascale Computing Project）”，希望于2021年至少交付一台E级超算，其中一台的名字为“极光（Aurora）”，初步规划峰值运算能力超过每秒130亿亿次，内存超过8PB，系统功耗约为40MW。

欧盟预计于2022年—2023年交付首台E级超算，使用的是美国、欧盟处理器，架构有可能类似ARM；日本发展E级超算的“旗舰2020计划”由日本理化所主导，完成时间也设定在2020年

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。

在Linpark和HPCG性能之外，由于Summit使用了目前深度学习领域最强的GV100芯片作为加速芯片，因此具有强大的深度学习计算能力，每秒可以进行30亿亿次混合精度计算（30 exaops），成为目前深度学习计算最强大的计算机。

1. summit构架：
2. 硬件构架：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。

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。

在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(NVLink采用点对点结构、串列传输，用于中央处理器（CPU）与图形处理器（GPU）之间的连接，也可用于多个图形处理器之间的相互连接。多为针对高性能运算应用领域，像是英伟达最新的Tesla P100运算卡。)来取代PCIe总线。

1. CPU间的通讯：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；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. 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。

4、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，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。