**Summit架构分析**

**智能1702班 尹淑青 201708010601**

1. Summit超级计算机简介

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。2018年11月12日，全球超级计算机500强榜单中，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。

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%。

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。

1. Summit系统架构概述

Summit系统是典型的MVC结构系统，其中View层称为SummitFT，基于微软C# .NET技术；Control层分为2部分，一部分为Java开发的通信中间层，另一部分为C/C++编写的Summit主体部分；最后，Model层作为Summit业务数据抽象、存取层，基于ENTITY实现，支持主流的Oracle/SQL Server以及Sybase数据库。下面，我们从前端到后端，对Summit使用的主要技术进行介绍。

SummitFT使用Infragistics的C#控件库作为基础，封装出了一套自己的控件。整个界面风格统一、控件布局合理，操作方便，对用户比较友好。作为对比，Calypso基于Java做的界面；Kondor基于C做的界面，操作体验上来说，跟SummitFT是没法比。

我们再来说Control层，Summit作为典型的CS程序，客户端与服务端通讯采用的不是TCP/IP直接通讯的方式，而是采用了HTTP协议和WebService的方式。其中，SummitFT通过HTTP协议与通讯中件层通讯；通讯中间层采用WebService与etoolkit进程通讯，达到使用Summit后端服务的目的。这种设计的好处就是Control层不仅可以对接SummitFT，还提供了一套灵活的供其他客户端调用的方式，比如Summit就支持VBA、Java等其他语言的直接调用。由此可以看出，Summit系统在设计时已经考虑到了系统的开放性。通讯中间层采用Java语言编写，负责接收SummitFT的HTTP连接，并负责HTTP协议报文与SOAP报文之间的转换。Summit Business Control层即上文提到的etoolkit，etoolkit使用C/C++开发，实际上就是一个WebService Server，负责处理中间层的请求，并将结果封闭成SOAP报文，返回给通讯中间层。

Model层依赖Summit数据抽象ENTITY以及关系型数据库，目前支持Oracle, Sybase以及SQL Server。Model层进行Summit数据的序列化与反序列化。ENTITY即Summit系统的元数据，在Summit系统中，所有的数据（交易数据、静态数据、系统基础数据）都以ENTITY进行抽象。ENTITY不仅包含属性(Properties)，还会包含接口(Interface)和具体的方法(Method)。因此，ENTITY完全可以用现在的面向对象来理解。

1. 分析组成Summit的硬件部分

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。

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。

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。

NVLink 2.0

AC922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。

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。

CPU之间的通讯 X总线

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。

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。

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。