《计算机系统设计》课程设计报告

湖南大學

HUNAN UNIVERSITY

选题名称：Summit 架构分析

姓　　名：王倩

学　　号：201708010630

专业班级：物联 1702

Summit 超级计算机概要

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。

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。

工作原理

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。

性能数据

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。

Summit 超级计算机原定计算性能是 150petaflops，交付性能达到 200petaflops。中国的 TaihuLight 超级计算机性能指标是 93 petaflops，峰值性能是 124.5petaflops。IBM 这款超级计算机交易据说价值 3.25 亿美元。

产品应用

建成后，Summit 将可以解决一些世界上最紧迫的计算挑战。国内同时启动了三大百亿亿次超算研发，分别是国防科大/天津超算中心的天河三号、中科曙光的 E 级超算以及江南所/济南超算中心的神威 E 级。以上三套百亿亿次超算中，有一条要求是共同的，那就是核心处理器必须是国产的，神威· 太湖之光上已经用了国产申威 SW26010 处理器。

Summit 系统的层次架构

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。

前端到后端的相关技术

Infragistics 基础

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。

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:

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。

节点，机架和整体

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。

计算结点

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。

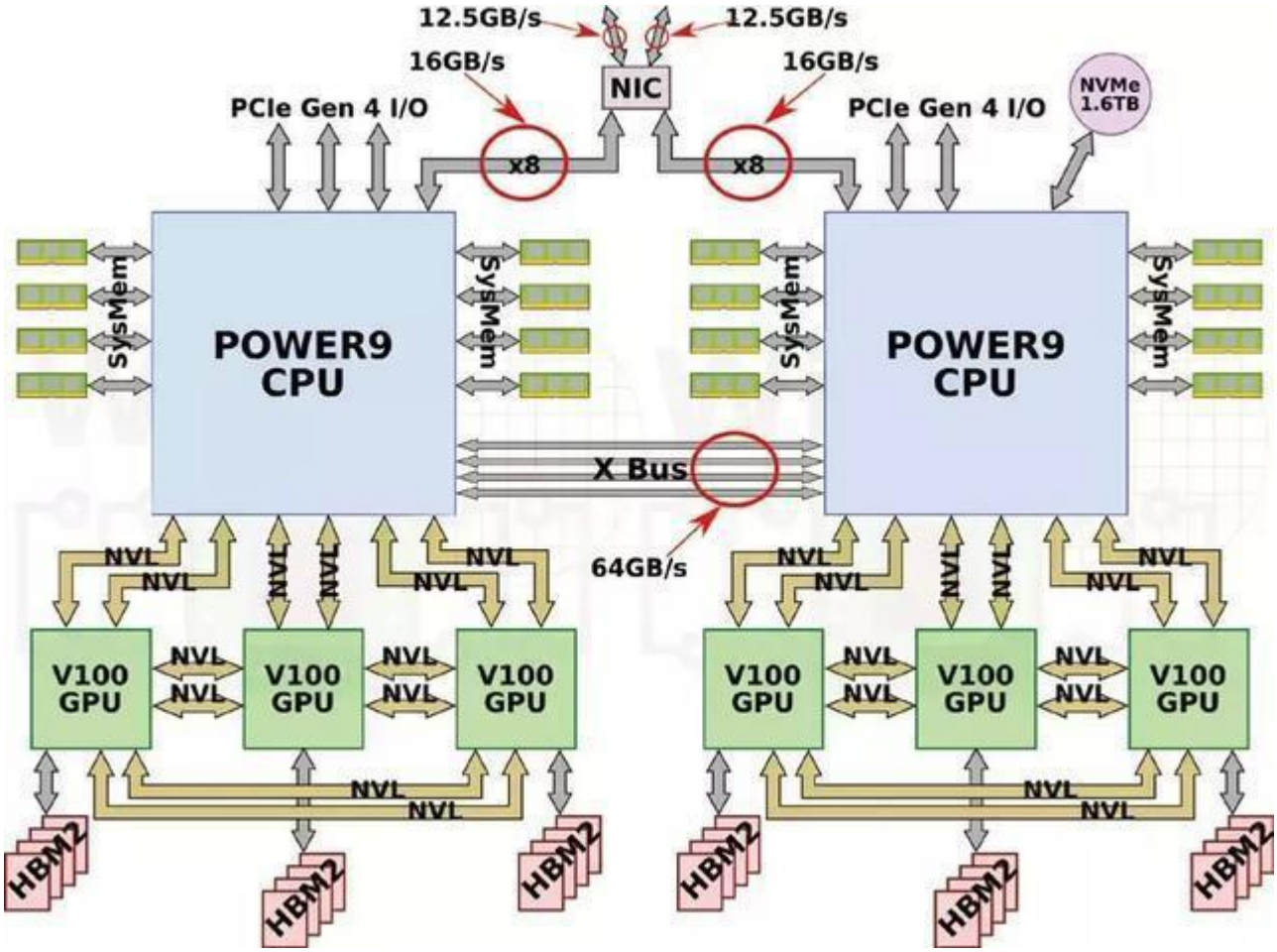
▲Summit 的一个计算节点，以及其内部设备

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。

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。

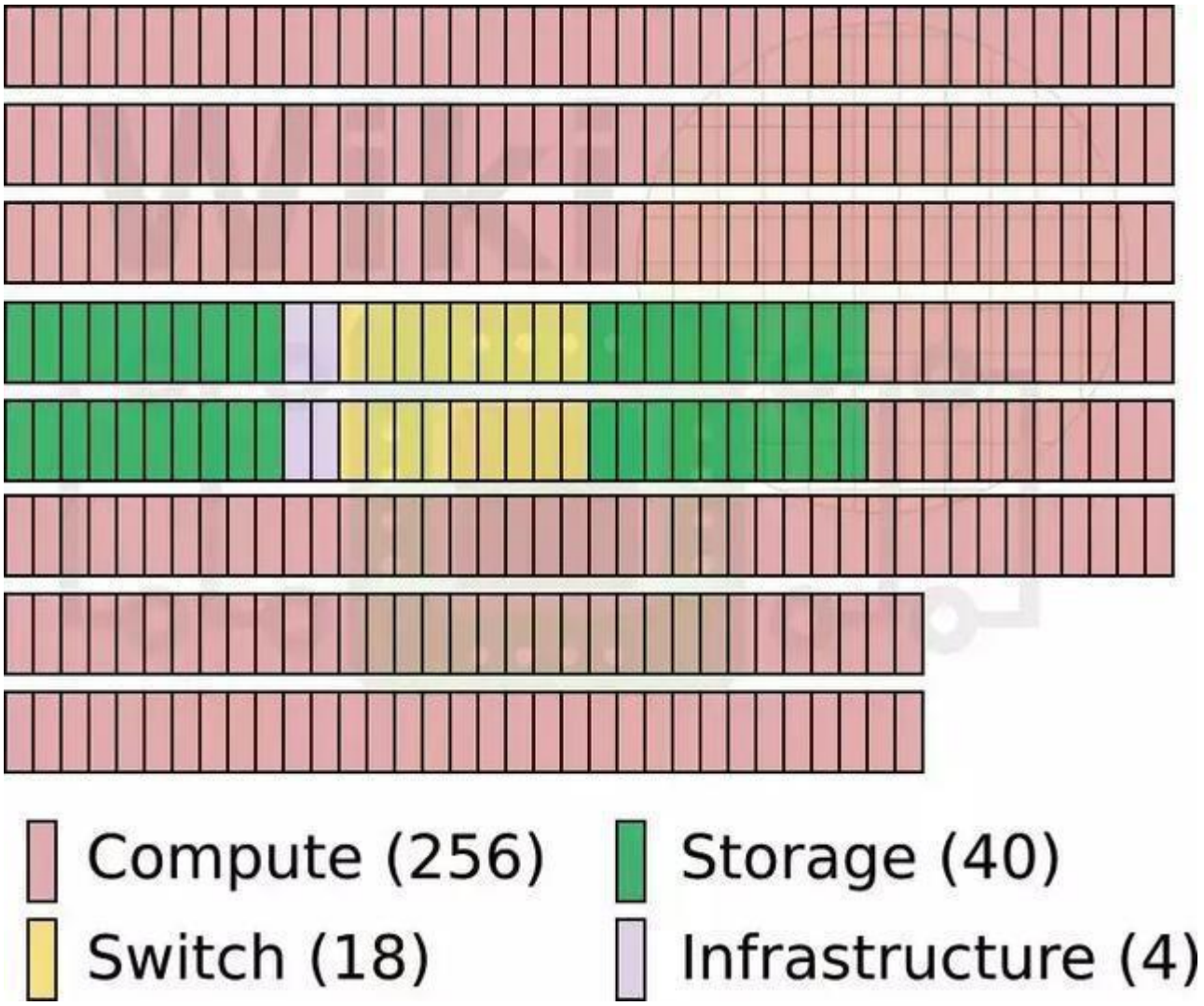
Summit的性能		
Summit峰值性能		
处理器	CPU	GPU
型号	POWER9	V100
数量	9,216 /2 × 18 × 256	27,648/ 6 × 18 ×256
峰值FLOPS	9.96 PF	215.7 PF
峰值AI FLOPS	N/A	3.456 EF

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。

迈向百亿亿次计算时代

Summit 通过强大的 CPU 和 GPU 以及网络、系统等部分先进的技术综合和结构设计，成功登顶了全球第一超算的宝座，并且这可能不是 Summit 的终点，Summit 仅仅是美国能源部在探索百亿亿次超算道路上的一个中间站而已。

目前的消息显示，橡树岭国家实验室正在准备一款名为 Frontier 的百亿亿次超算，其性能应该可以达到 Summit 的 5~10 倍。目前尚不清楚新的超算是在 Summit 上升级而来还是全部重新建立，但是无论如何，百亿亿次级别超算正在朝我们一步步走来，时间节点在 2021 年左右。

▲美国橡树岭国家实验室的超算发展路线图

与我国超算之间的对比

目前看起来，神威太湖之光和天河系列超算短期内都没有更新和建设的新计划，包括新的神威系列超算和人们猜测中的天河 3 号等。目前国内也在尽全力冲刺百亿亿次级别超算，但是在工艺和设计上还有不少瓶颈和困难尚未解决，百亿亿次级别超算依旧在不断的研发和构建过程中。

在这种情况下，一些业内人士估计 Summit 可能在未来 3~5 个超算排行周期都暂居领先的态势，直到最新的百亿亿次超算正式登场。毕竟在超算争霸的战场上，没有谁是永远的赢家，只有不断问世、性能更强的超级计算机。所以在这个战场，没有最强，只有更强。

Summit 系统的优缺点

Summit 的缺点在于，系统非常复杂而难以掌握。Summit 系统设计理念即为高可扩展性，在 Summit 系统中，几乎什么都是可以配置的。也正是因为这个，每个功能都依赖于太多的配置，以至于需要大量时间去搞清楚该功能的依赖配置到底该如何配置。其初始学习曲线比较陡，但花 1~2 年熟悉 Summit 套路后，就可触类旁通，掌握基本用法。

Summit 系统前端使用 C#开发，因此，只能在 Windows 环境下运行。界面相当友好。后端使用 C/C++开发，由于历史遗留问题(毕竟 90 年代的系统)，几乎所有底层 API 都是使用 C 写的，后来 C++出现后，只是在外围包裹了一层 C++而已。Summit 系统还有一个中间层，用来进行前/后端通讯。

Summit 系统优秀的架构，提供了其技术层面上高可扩展性。一方面，Summit 系统的开发工作相对来说较容易，只需要掌握一些 API 的使用规则即可。大量的精力是花在搞懂业务规则上。另一方面，如果不使用 Summit API，自己也可以使用 C/C++和其开源库，写出很多好用、性能高的框架，然后套到 Summit 后端。笔者就写过很多平台类的组件，套到 Summit 整个框架内。

总结

Summit 系统的架构的真的非常优秀。90 年代，C 语言主流，C++/Java 刚刚兴起，Summit 系统的数据模型，可以做到面向对象。Summit 所有的数据结构，都可以像 Java 类一样，知道其属性、方法，可以直接进行属性扩展和方法扩展；Summit 系统使用的前/后端通信方式，不是当时流行的 TCP，而是 HTTP 协议；Summit 系统各个组件，License 管理、认证管理、数据库访问等层次分明，类似于 Linux 分层结构，非常优秀。