《计算机系统设计》课程报告

湖南大學

HUNAN UNIVERSITY

选题名称：各测试系统架构分析

姓　　名：袁 纯 楸

学　　号：201708010225

专业班级：智能 1701

信息科学与工程学院

一、 Linpack 标准测试程序及其分析

Linpack 测试包括三类，Linpack100、Linpack1000 和 HPL。

Linpack100：求解规模为 100 阶的稠密线性代数方程组，它只允许采用编译优化选项进行优化，不得更改代码，甚至代码中的注释也不得修改；

Linpack1000：求解规模为 1000 阶的线性代数方程组，达到指定的精度要求，可以在不改变计算量的前提下做算法和代码上做优化。

HPL（高度并行计算基准测试）：对数组大小 N 没有限制，求解问题的规模可以改变，除基本算法（计算量）不可改变外，可以采用其它任何优化方法。

LINPACK 压力测试的目的主要为检测系统中 CPU 的工作的稳定性及内存访问的稳定性。

安装过程比较复杂，略：

相关配置文件：

在 HPL 测试中，使用的参数选择与测试的结果有很大的关系。HPL 中参数的设定是通过从一个配置文件 HPL.dat 中读取的，所以在测试前要改写 HPL.dat 文件，设置需要使用的各种参数，然后再开始运行测试程序。配置文件内容的结构如下：

```
HPLinpack benchmark input file //文件头
Innovative Computing Laboratory, University of Tennessee//说明性文字

HPL.out output file name (if any)//如果使用文件保留输出结果，设定文件名
6 device out (6=stdout,7=stderr,file)//输出方式选择（stdout,stderr 或文件）
“device out”=“6”时，测试结果输出至标准输出（stdout）
“device out”=“7”时，测试结果输出至标准错误输出（stderr）
“device out”为其它值时，测试结果输出至第三行所指定的文件中

2 # of problems sizes (N) //指出要计算的矩阵规格数量
```

1960 2048Ns //每种规格分别的数值

• 矩阵的规模 N 越大，有效计算所占的比例也越大，系统浮点处理性能也就越高；但矩阵规模 N 的增加会导致内存消耗量的增加——当系统实际内存空间不足，使用缓存、性能会大幅度降低。

• 计算规模的大小，应以设置的大页面内存总量做计算，计算方式为：N*N*8=大页内存总量*0.8，内存总量换算为字节，规模的大小最好为 384 的倍数。

2 # of NBs //指出使用不同的分块大小数量

60 80 NBs //分别指出每种大小的具体值

提高数据的局部性，从而提高整体性能，HPL 采用分块矩阵的算法。

分块的大小对性能有很大的影响，NB 的选择和软硬件许多因素密切相关 NB 值的选择主要是通过实际测试得到最优值；NB 大小的选择还跟通信方式、矩阵规模、网络、处理器速度等有关系。

1 PMAP process mapping (0=Row-,1=Column-major)

选择处理器阵列是按列的排列方式还是按行的排列方式。

按列的排列方式适用于节点数较多、每个节点内 CPU 数较少的系统；

按行的排列方式适用于节点数较少、每个节点内 CPU 数较多的大规模系统。

2 # of process grids (P x Q) //指出用进程组合方式数量

2 4 Ps //每对 PQ 具体的值

2 1 Qs

二维处理器网格（P×Q）的要求：

P×Q = 系统 CPU 数 = 进程数。（一般：一个进程对于一个 CPU 可以得到最佳性能；Intel Xeon：关闭超线程可以提高 HPL 性能）

• P≤Q：P 的值尽量取得小一点，列向通信量（通信次数和数据量）>>横向通信。

• P=2n：L 分解的列向通信采用二元交换法，此时性能最优。

• 在集群测试中，P×Q = 系统 CPU 总核数。

16.0 threshold //余数的阈值

说明测试的精度。做完线性方程组的求解以后，检测求解结果是否正确：

若误差在这个值以内就是正确，否则错误。

若是求解错误，其误差非常大；若正确，则很小。

1 # of panel fact //用分解方法数量

1 PFACTs (0=left, 1=Crout, 2=Right) //具体那种分解方法

0 left,1 crout,2 right

1 # of recursive stopping criterium //停止递归的判断标准

4 NBMINs (>= 1) //具体的标准数值（须不小于 1）

1 # of panels in recursion //递归中用分割法的数量

2 NDIVs //一种 NDIV 值为 2，即每次递归分成两块

1 # of recursive panel fact. //用递归分解方法的数量

2 RFACTs (0=left, 1=Crout, 2=Right) //这里每种都用到（左，右，crout 分解）

L 分解的方式：

在消元过程中，HPL 采用每次完成 NB 列的消元（L 分解），然后更新后面的矩阵。

对每一个小矩阵作消元时，都有 3 种算法：L、R、C，分别代表 Left、Right 和 Crout。

1 # of broadcast //用广播方法的数量

3 BCASTs (0=1rg,1=1rM,2=2rg,3=2rM,4=Lng,5=LnM) //指定具体哪种（有 1-ring,1-ring Modified,2-ring,2ring Modified,Long 以及 long-Modified）

L 的横向广播方式：

HPL 中提供了 6 种广播方式：前 4 种适合于快速网络；后 2 种采用将数据切割后传送的方式，主要适合于速度较慢的网络。目前，机群系统一般采用千兆以太网甚至光纤等高速网络。在小规模系统中，选择 0 或 1；对于大规模系统，选择 3

1 # of lookahead depth //用几种向前看的步数

1 DEPTHs (>=0) //具体步数值（须大于等于 0）

横向通信的通信深度：依赖于机器的配置和问题规模的大小。

2 SWAP (0=bin-exch,1=long,2=mix)//交换算法（bin-exchange，long 或二者混合）

64 swapping threshold //采用混合的交换算法时使用的阈值

U 的广播算法：列向广播

HPL 提供了 3 种 U 的广播算法：二元交换法、Long 法和二者混合法。

SWAP="0"，采用二元交换法；SWAP="1"，采用 Long 法；SWAP="2"，采用混合法。

0 L1 in (0=transposed,1=no-transposed) form //L1 是否用转置形式

0 U in (0=transposed,1=no-transposed) form //U 是否用转置形式表示

L 和 U 的数据存放格式：若选择“transposed”->采用按列存放，否按行存放。

1 Equilibration (0=no,1=yes) //是否采用平衡状态

在回代中使用，一般使用其默认值

8 memory alignment in double (> 0) //指出程序运行时内存分配中的采用的对齐方式

为内存地址对齐设置：用于在内存分配中对齐地址。出于安全考虑-选择 8。

为调试出高的性能，必须考虑内存大小，网络类型以及拓扑结构，调试上面的参数，直到得出最高性能。

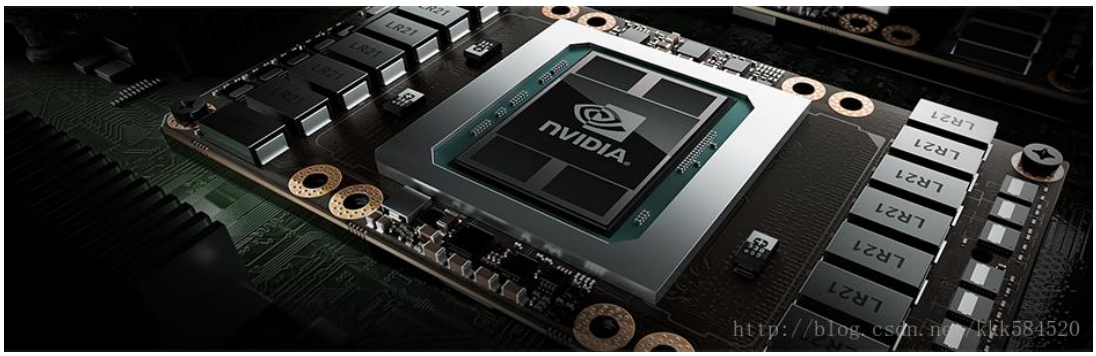
二、 Tesla GPU 架构分析

以 Tesla P00 为例：

其 NVIDIA 官网相关参数：

	适用于基于 PCIe 的服务器的 P100	适用于 NVLink 优化服务器的 P100
双精度浮点运算能力	4.7 teraFLOPS	5.3 teraFLOPS
单精度浮点运算能力	9.3 teraFLOPS	10.6 teraFLOPS
半精度浮点运算能力	18.7 teraFLOPS	21.2 teraFLOPS
NVIDIA NVLink 互联带宽	-	160 GB/s
PCIe x16 互联带宽	32 GB/s	32 GB/s
CoWoS HBM2 堆叠式显存容量	16 GB or 12 GB	16 GB
CoWoS HBM2 堆叠式显存带宽	732 GB/s or 549 GB/s	732 GB/s
提升使用页面迁移引擎编程的能力	✓	✓
ECC 保护助力实现可靠性	✓	✓
针对数据中心部署优化服务器	✓	✓

Tesla P100 采用顶级大核心 GP100，面积 610 mm^2，如图中间位置：



GP100 参数汇总如下：

芯片：GP100

架构：SM_60

工艺：16 nm FinFET

支持：双精度 FP64，单精度 FP32，半精度 FP16

功耗：250 W

CUDA 核心数：3584（56 SMs，64 SPs/SM）

GPU 时钟（Base/Boost）：1189 MHz/1328 MHz

PCIe：Gen 3 x16

显存容量：12/16 GB HBM2

显存位宽：3072/4096 bits

显存时钟：715 MHz

显存带宽：539/732 GB/s

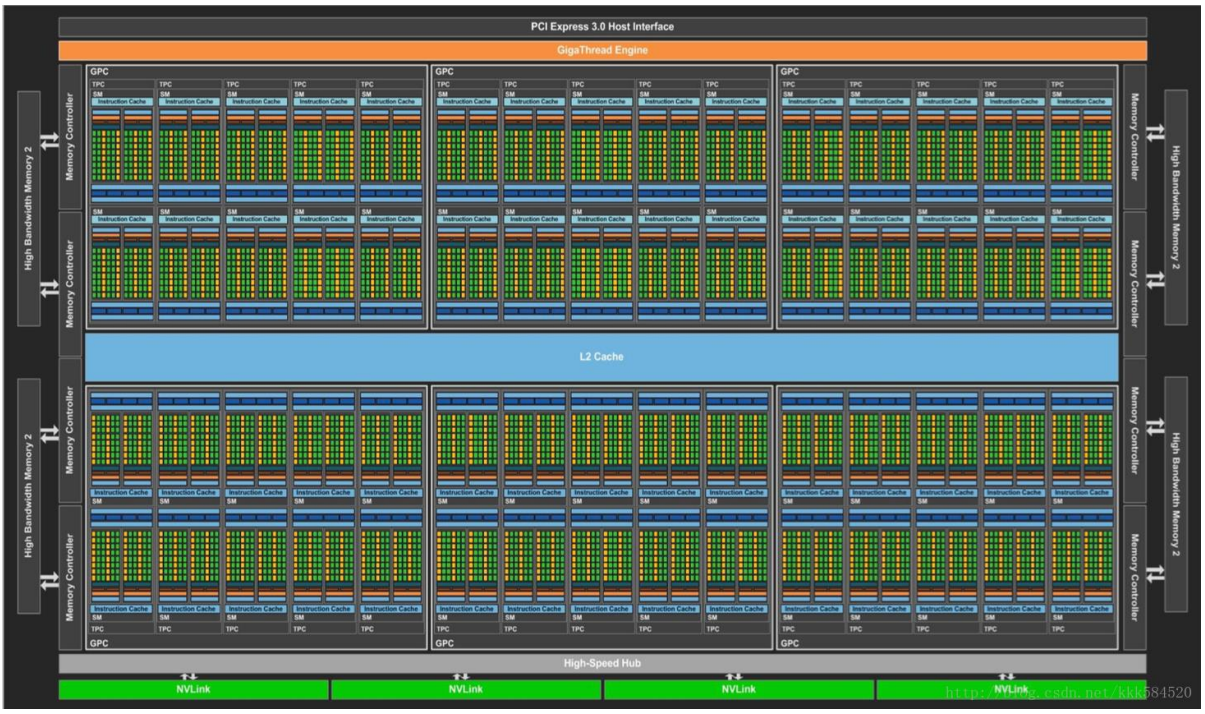
DGX-1 拥有 8 颗帕斯卡架构 GP100 核心的 Tesla P100 GPU，一共组成了 128 GB 的显存。

浮点运算（FP16）达到了 170 TFlops（官方宣称相当于 250 个 x86 服务器），内置 7 TB 的 SSD，两颗 16 核心的 Xeon E5-2698v3 以及 512 GB 的 DDR4 内存，整机功耗 3200W，售价为 129000$。

其 GP100 GPU 硬件架构：

GP100 称：最高性能并行计算处理器。

GP100 包含一组 GPC（图形处理簇，Graphics Processing Clusters）、TPC（纹理处理簇，Texture Processing Clusters）、SM（流多处理器，Stream Multiprocessors）以及内存控制器。其架构如下：



一颗完整的 GP100 芯片包括 6 个图形处理簇，60 个 Pascal 流多处理器，30 个纹理处理簇和 8 个 512 位内存控制器（总共 4096 位）。

每个图形处理簇内部包括 10 个 流多处理器。

每个流多处理器内部包括 64 个 CUDA 核心和 4 个纹理单元。

Tesla P100 只用了 GP100 上 60 个 SM 中的 56 个。

GP100 的第六代 SM 架构提高了 CUDA 核心利用率和能效，核心频率更高，整体 GPU 性能有较大提升。

GP100 的 SM 包括 64 个单精度 CUDA 核心。而 Maxwell 和 Kepler 的 SM 分别有 128 和 192 个单精度 CUDA 核心。虽然 GP100 SM 只有 Maxwell SM 中 CUDA 核心数的一半，但总的 SM 数目增加了，每个 SM 保持与上一代相同的寄存器组，则总的寄存器数目增加了。这意味着 GP100 上的线程可以使用更多寄存器，也意味着 GP100 相比旧的架构支持更多线程、warp 和线程块数目。与此同时，GP100 总共享内存量也随 SM 数目增加而增加了，带宽显著提升不至两倍。

一个 SM 的架构：

其中绿色的“Core” 为单精度 CUDA 核心，共有 64 个，同时支持 32 位单精度浮点计算和 16 位半精度浮点计算，其中 16 位计算吞吐是 32 位计算吞吐的两倍。图中橘黄色的 “DP Unit” 为双精度计算单元，支持 64 位双精度浮点计算，数量为 32 个。每个 GP100 SM 双精度计算吞吐为单精度的一半。

下面处理性能表证实：

PERFORMANCE SPECIFICATION FOR NVIDIA TESLA P100 ACCELERATORS

	P100 for PCIe-Based Servers	P100 for NVLink-Optimized Servers
Double-Precision Performance	4.7 TeraFLOPS	5.3 TeraFLOPS
Single-Precision Performance	9.3 TeraFLOPS	10.6 TeraFLOPS
Half-Precision Performance	18.7 TeraFLOPS	21.2 TeraFLOPS

Pascal SM 架构图：一个 GP100 SM 分成两个处理块，每块有【 32768 个 32 位寄存器 + 32 个单精度 CUDA 核心 + 16 个双精度 CUDA 核心 + 8 个特殊功能单元(SFU) + 8 个存取单元 + 一个指令缓冲区 + 一个 warp 调度器 + 两个分发单元】

- Pascal SM 架构相比 Kepler 架构简化了数据通路，占用面积更小，功耗更低。
- Pascal SM 架构提供更高级的调度和重叠载入/存储指令来提高浮点利用率。
- GP100 中新的 SM 调度器架构相比 Maxwell 更智能，具备高性能、低功耗特性。
- warp 调度器（每个处理块共享 1 个）在一个时钟周期内分发两个 warp 指令。

双精度算法是很多 HPC 应用（如线性代数，数值模拟，量子化学等）的核心。GP100 的一个关键设计目标就是显著提升这些案例的性能。

GP100 中每个 SM 都有 32 个双精度（FP64）CUDA 核心，即单精度（FP32）CUDA 核心数目的一半。GP100 和以前架构相同，支持 IEEE 754-2008 标准，支持 FMA 运算，支持异常值处理。

注意：在 Kepler GK110 中单精度核心数目：双精度核心数目为 3:1。

使用 FP16 计算相比 FP32 可以带来 2 倍性能提升，数据传输时间也可以大大降低。

注意：GP100 上，两个 FP16 运算可以使用一个双路指令完成。

相比上一代产品和上上代产品情况如下表所示：

Table 1. Tesla P100 Compared to Prior Generation Tesla products

Tesla Products	Tesla K40	Tesla M40	Tesla P100
GPU	GK110 (Kepler)	GM200 (Maxwell)	GP100 (Pascal)
SMs	15	24	56
TPCs	15	24	28
FP32 CUDA Cores / SM	192	128	64
FP32 CUDA Cores / GPU	2880	3072	3584
FP64 CUDA Cores / SM	64	4	32
FP64 CUDA Cores / GPU	960	96	1792
Base Clock	745 MHz	948 MHz	1328 MHz
GPU Boost Clock	810/875 MHz	1114 MHz	1480 MHz
Peak FP32 GFLOPs[1]	5040	6840	10600
Peak FP64 GFLOPs[1]	1680	210	5300
Texture Units	240	192	224
Memory Interface	384-bit GDDR5	384-bit GDDR5	4096-bit HBM2
Memory Size	Up to 12 GB	Up to 24 GB	16 GB
L2 Cache Size	1536 KB	3072 KB	4096 KB
Register File Size / SM	256 KB	256 KB	256 KB
Register File Size / GPU	3840 KB	6144 KB	14336 KB
TDP	235 Watts	250 Watts	300 Watts
Transistors	7.1 billion	8 billion	15.3 billion
GPU Die Size	551 mm^2	601 mm^2	610 mm^2
Manufacturing Process	28-nm	28-nm	16-nm FinFET

[1] The GFLOPS in this chart are based on GPU Boost Clocks.

表现出了优异的高性能和高能效。

Maxwell 架构相对 Kepler 做了改进，提升了效率。

Pascal 基于 Maxwell 架构，利用 TSMC 16 nm FinFET 工艺进一步降低了能耗。

计算能力：

GP100 GPU 支持最新 6.0 计算能力。

Table 2. Compute Capabilities: GK110 vs GM200 vs GP100

GPU	Kepler GK110	Maxwell GM200	Pascal GP100
Compute Capability	3.5	5.2	6.0
Threads / Warp	32	32	32
Max Warps / Multiprocessor	64	64	64
Max Threads / Multiprocessor	2048	2048	2048
Max Thread Blocks / Multiprocessor	16	32	32
Max 32-bit Registers / SM	65536	65536	65536
Max Registers / Block	65536	32768	65536
Max Registers / Thread	255	255	255
Max Thread Block Size	1024	1024	1024
Shared Memory Size / SM	16 KB/32 KB/48 KB	96 KB	64 KB

三、 Infiniband 网络结构分析

InfiniBand 技术不是用于一般网络连接的，它的主要设计目的是针对服务器端的连接问题的。因此，InfiniBand 技术将会被应用于服务器与服务器（比如复制，分布式工作等），服务器和存储设备（比如 SAN 和直接存储附件）以及服务器和网络之间（比如 LAN， WANs 和 the Internet）的通信。

1、 IB 基本概念：

IB 是以通道为基础的双向、串行式传输，在连接拓朴中是采用交换、切换式结构(Switched Fabric)，在线路不够长时可用 IBA 中继器(Repeater)进行延伸。每一个 IBA 网络称为子网(Subnet)，每个子网内最高可有 65,536 个节点(Node)，IBA Switch、IBARepeater 仅适用于 Subnet 范畴，若要通跨多个 IBASubnet 就需要用到 IBA 路由器(Router)或 IBA 网关器(Gateway)。

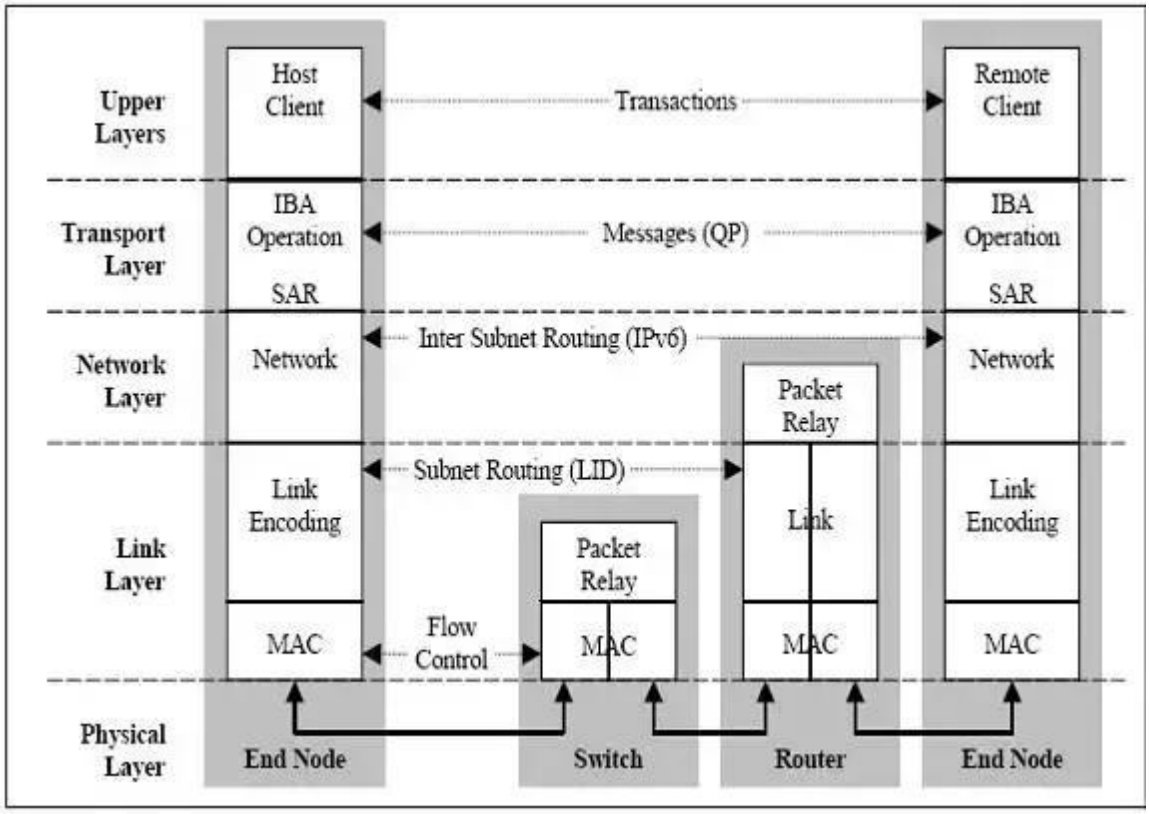
每个节点(Node) 必须透过配接器(Adapter)与 IBA Subnet 连接，节点 CPU、内存要透过 HCA(Host Channel Adapter)连接到子网；节点硬盘、I/O 则要透过 TCA(TargetChannel Adapter)连接到子网，这样的一个拓扑结构就构成了一个完整的 IBA。

IB 的传输方式和介质相当灵活，在设备机内可用印刷电路板的铜质线箔传递(Backplane 背板)，在机外可用铜质缆线或支持更远光纤介质。若用铜箔、铜缆最

远可至 17m，而光纤则可至 10km，同时 IBA 也支持热插拔，及具有自动侦测、自我调适的 Active Cable 活化智能性连接机制。

2、 IB 协议简介：

InfiniBand 也是一种分层协议(类似 TCP/IP 协议)，每层负责不同的功能，下层为上层服务，不同层次相互独立。 IB 采用 IPv6 的报头格式。其数据包报头包括本地路由标识符 LRH，全局路由标示符 GRH，基本传输标识符 BTH 等。



A、 物理层

物理层定义了电气特性和机械特性，包括光纤和铜媒介的电缆和插座、底板连接器、热交换特性等。定义了背板、电缆、光缆三种物理端口。

并定义了用于形成帧的符号(包的开始和结束)、数据符号(DataSymbols)、和数据包直接的填充(Idles)。详细说明了构建有效包的信令协议，如码元编码、成帧标志排列、开始和结束定界符间的无效或非数据符号、非奇偶性错误、同步方法

等。

B、 链路层

链路层描述了数据包的格式和数据包操作的协议，如流量控制和子网内数据包的路由。链路层有链路管理数据包和数据包两种类型的数据包。

C、 网络层

网络层是子网间转发数据包的协议，类似于 IP 网络中的网络层。实现子网间的数据路由，数据在子网内传输时不需网络层的参与。

数据包中包含全局路由头 GRH，用于子网间数据包路由转发。全局路由头部指明了使用 IPv6 地址格式的全局标识符(GID)的源端口和目的端口，路由器基于 GRH 进行数据包转发。GRH 采用 IPv6 报头格式。GID 由每个子网唯一的子网 标示符和端口 GUID 捆绑而成。

D、 传输层

传输层负责报文的分发、通道多路复用、基本传输服务和处理报文分段的发送、接收和重组。传输层的功能是将数据包传送到各个指定的队列(QP)中，并指示队列如何处理该数据包。当消息的数据路径负载大于路径的最大传输单元(MTU)时，传输层负责将消息分割成多个数据包。

接收端的队列负责将数据重组到指定的数据缓冲区中。除了原始数据报外，所有的数据包都包含 BTH，BTH 指定目的队列并指明操作类型、数据包序列号和分区信息。

E、 上层协议

InfiniBand 为不同类型的用户提供了不同的上层协议，并为某些管理功能定义了消息和协议。InfiniBand 主要支持 SDP、SRP、iSER、RDS、IPoIB 和 uDAPL 等上层协议。

• SDP(SocketsDirect Protocol)是 InfiniBand Trade Association (IBTA)制定的基于 infiniband 的一种协议，它允许用户已有的使用 TCP/IP 协议的程序运行在高速的 infiniband 之上。

• SRP(SCSIRDMA Protocol)是 InfiniBand 中的一种通信协议，在 InfiniBand 中将 SCSI 命令进行打包，允许 SCSI 命令通过 RDMA(远程直接内存访问)在不同的系统之间进行通信，实现存储设备共享和 RDMA 通信服务。

• iSER(iSCSIRDMA Protocol)类似于 SRP(SCSI RDMA protocol)协议，是 IB SAN 的一种协议 ，其主要作用是把 iSCSI 协议的命令和数据通过 RDMA 的方式跑到例如 Infiniband 这种网络上，作为 iSCSI RDMA 的存储协议 iSER 已被 IETF 所标准化。

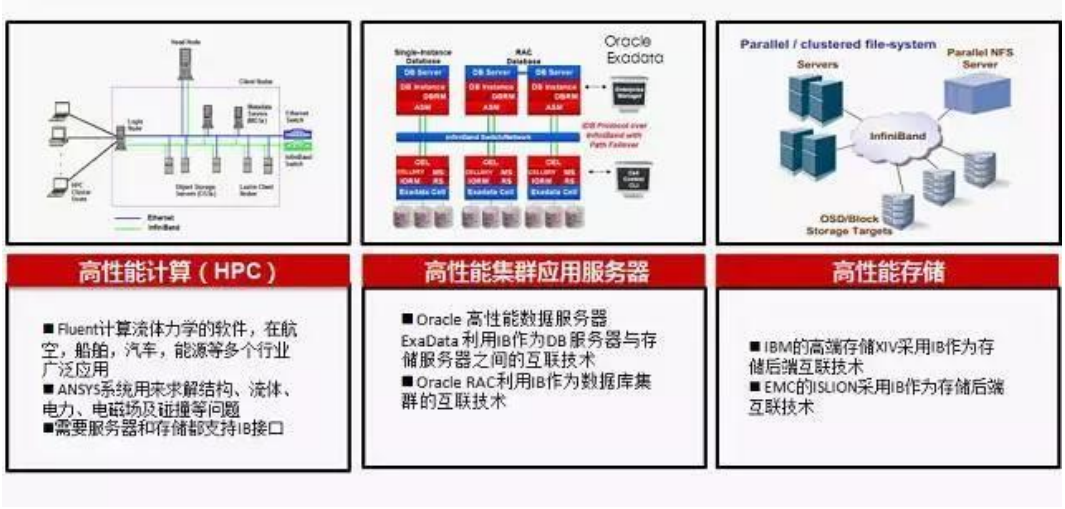
• RDS(ReliableDatagram Sockets)协议与 UDP 类似，设计用于在 Infiniband 上使用套接字来发送和接收数据。实际是由 Oracle 公司研发的运行在 infiniband 之上，直接基于 IPC 的协议。

•IPoIB(IP-over-IB)是为了实现 INFINIBAND 网络与 TCP/IP 网络兼容而制定的协议，基于 TCP/IP 协议，对于用户应用程序是透明的，并且可以提供更大的带宽，也就是原先使用 TCP/IP 协议栈的应用不需要任何修改就能使用 IPoIB。

• uDAPL(UserDirect Access Programming Library)用户直接访问编程库是标准的 API，通过远程直接内存访问 RDMA 功能的互连（如 InfiniBand）来提高数据中心应用程序数据消息传送性能、伸缩性和可靠性。

3、 IB 应用场景

Infiniband 灵活支持直连及交换机多种组网方式，主要用于 HPC 高性能计算场景，大型数据中心高性能存储等场景，HPC 应用的共同诉求是低时延(<10 微秒)、低 CPU 占有率（<10%）和高带宽(主流 56 或 100Gbps)



• Infiniband 在主机侧采用 RDMA 技术释放 CPU 负载，可以把主机内数据处理的时延从几十微秒降低到 1 微秒；

• InfiniBand 网络的高带宽(40G、56G 和 100G)、低时延(几百纳秒)和无丢包特性吸取了 FC 网络的可靠性和以太网的灵活扩展能力。

四、 Summit 架构分析

根据 ORNL 官方网站的介绍，Summit 在运算能力方面的**特点**如下：

1、最高能够实现每秒 20 亿亿次的计算峰值，上一代超级计算机 Titan 的 8 倍；

2、在特定应用中，Summit 能实现每秒 330 亿亿次（3.3exaops）混合精度计算；

3、Summi 能够为能源、前沿材料和人工智能等领域提供前所未有的计算能力。

A. 从结构上：

Summit 由一排排硬件单元连接而成矩阵，总重 340 吨；通过长度为 185 英尺的光纤相连接，每分钟需要 4000 加仑的水来冷却，同时消耗的电量也很大。

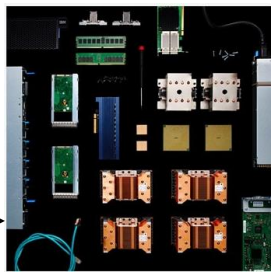
B. 从硬件上：

Summit 使用了 IBM AC922 系统（包含了 4608 个计算服务器，每个服务器包含了两个 22 核的 IBM Power9 处理器和 6 个 NVIDIA Tesla V100 GPU 加速器，它们之间通过双轨的 Mellanox EDR 100Gb/s InfiniBand 连接）；Summit 配备了 10 Petabytes 的存储空间，以及高速、高带宽的数据传输系统。

Summit 超算的浮点性能可达 200PFLOPS，也就是 20 亿亿次，它总共使用了 4608 个节点，每个节点性能 42TFLOPS，由 2 个 Power 9 22 核处理器及 6 个 Telsa V100 加速卡组成，搭配 512GB DDR4 内存及 96GB HBM 2 显存，每个节点搭配 1.6TB 非易失性内存，总的内存容量超过 10PB，存储系统容量高达 250PB，带宽 2.5TB/s。

Application Performance	200 PF
Number of Nodes	4,608
Node performance	42 TF
Memory per Node	512 GB DDR4 + 96 GB HBM2
NV memory per Node	1600 GB
Total System Memory	>10 PB DDR4 + HBM2 + Non-volatile
Processors	2 IBM POWER9™ 9,216 CPUs 6 NVIDIA Volta™ 27,648 GPUs
File System	250 PB, 2.5 TB/s, GPFS™
Power Consumption	13 MW
Interconnect	Mellanox EDR 100G InfiniBand
Operating System	Red Hat Enterprise Linux (RHEL) version 7.4

*IBM AC922 系统：



I/O 速度更快 - I/O 带宽可高达 x86 服务器的 5.6 倍

Power AC922 包含各种下一代 I/O 架构，包括 PCIe Gen4、CAPI 2.0、OpenCAPI 和 NVIDIA(R) NVLINK(TM)，相比 x86 服务器，可为数据密集型工作负载提供高达 5.6 倍的带宽。

立即观看视频

PCIe Gen4 - 是 PCIe Gen3 带宽的 2 倍

第三代 Power AC922 是业界首款采用下一代业界标准 PCIe 互连技术的服务器。第四代 PCIe 提供的数据带宽大约是 x86 服务器中采用的第三代 PCIe 互连线路的 2 倍。

POWER9 处理器 - 最新 POWER 处理器，专为 AI 设计

POWER9 专为 AI 时代而构建，与 x86 竞争产品相比，可提供超过 5.6 倍的 I/O 性能，支持超过 2 倍的线程。POWER9 可用于在 Power AC922 服务器内包含 16 核到最多 44 核的任意配置。

高级 GPU - 提供最多 6 个含 NVLink 的 NVIDIA® Tesla® V100 GPU

Power AC922 将 IBM POWER9 CPU 与带有 NVLink GPU 的 NVIDIA Tesla V100 配对，打造出一台能够将性能提升多达 5.6 倍的服务器。这将能够提供大规模的吞吐能力来支持 HPC、深度学习和 AI 工作负载。

连贯性 - 跨 CPU 和 GPU 共享 RAM

Power AC922 中 CPU 与 GPU 的一致性解决了增长问题，允许加速后的应用将系统内存用作 GPU 内存。此连贯性有助于通过消除数据移动需求和定位需求来简化编程工作。

专为可从一台服务器扩展到超级计算机而构建

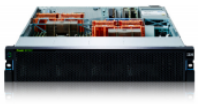
Power AC922 是美国能源部的 Summit 和 Sierra 超级计算机的主干。利用 AI 部署的可扩展性、高效率和易用性，它也是帮助您的企业达成 AI 夙愿的理想选择。

*IBM Power9 处理器：

Enterprise AI, Deep Learning & Machine Learning

and AI frameworks

These servers provide the fastest, simplest way to deploy deep learning frameworks—with enterprise-class support—to fuel new thinking and capabilities across your organization.



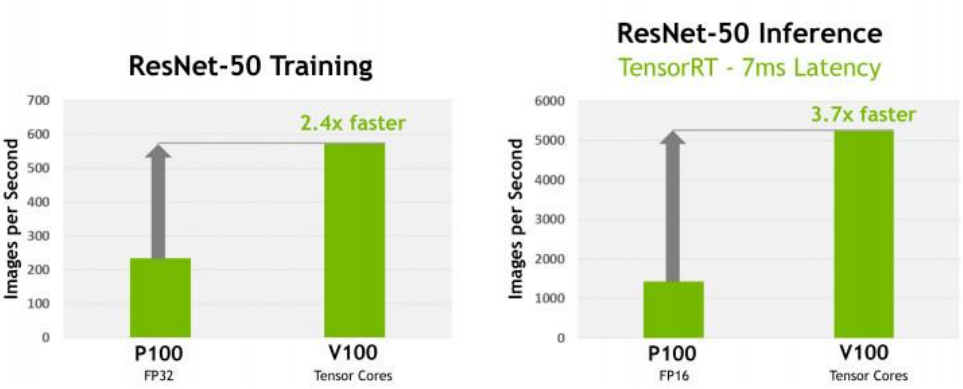
Feature	AC922	LC922
MTM	8335-GTH | 8335-GTX	9006-22P
System Packaging	2U	2U
Processor Socket	2S	2S
# of cores	Up to 44 cores	Up to 44 cores
Number of GPUs	4 or 6 Nvidia Tesla GPU processors (NVLink 2.0 attached)	Not Available
Memory DIMM Slots	16	16
Memory–Max	1TB	1TB
HDD/SSD	Two SFF (2.5") SATA drives for Max 4 TB (HDD) Max 7.68 TB (SSD)	12 SFF/LFF (HDD/SSD) (4x NVMe enabled) Max 120 TB (HDD) Max 45.6 TB (SSD)
PCIe G4 Slot	4 Slots	6 Slots

*NVIDIA Tesla V100 GPU 加速器：

AI 计算和 HPC 的源动力，。

NVIDIA Tesla V100 加速器的核心（基于）Volta GV100 GPU 处理器。NVIDIA 设计

的最新 12nm FFN 高精度制程封装技术，GV100 在 815 平方毫米的芯片尺寸中，内部集成了高达 211 亿个晶体管结构。



• Tesla V100 的关键特性：

1) 针对深度学习优化的流式多处理器（SM）架构：提高了约 50% 的能效，在同样的功率范围内可以大幅提升 FP32（单精度浮点）和 FP64（双精度浮点）的运算性能；全新 Tensor Core 在模型训练场景中，最高可以达到 12 倍速的 TFLOP（每秒万亿次浮点运算）；全新的 SM 架构对整型和浮点型数据采取了相互独立且并行的数据通路；Volta 架构新的独立线程调度功能还可以实现并行线程之间的细粒度同步和协作；L1 高速数据缓存和共享内存子系统显著提高了性能。
2) 第二代 NVLink：提供了更高的带宽，更多的连接和更强的可扩展性。GV100 GPU 最多支持 6 个 NVLink 链路，每个 25 GB/s，总共 300 GB/s；NVLink 支持基于 IBM Power 9 CPU 服务器的 CPU 控制和高速缓存一致性功能；NVIDIA DGX-1V 超级 AI 计算机使用 NVLink 技术为超快速的深度学习模型训练提供了更强的扩展性。
3) HBM2 内存：更快，更高效。Volta 高度优化的 16GB HBM2 内存子系统可提供高达 900 GB/s 的峰值内存带宽。
4) Volta 多处理器服务：Volta MPS 提供 CUDA MPS 服务器关键组件的硬件加速功能（共享 GPU 的多计算任务场景中显著提升计算性能、隔离性和服务质量）；Volta MPS 将 MPS 支持的客户端最大数量提升至 48 个。
5) 增强的统一内存和地址转换服务：Volta GV100 中的 GV100 统一内存技术实现了一个新的访问计数器，可以根据每个处理器的访问频率精确调整内存页的寻

址；在 IBM Power 平台上，新的地址转换服务（ATS）允许 GPU 直接访问 CPU 的存储页表。

6) Cooperative Groups（协作组）和新的 Cooperative Launch API（协作启动 API）：

• Cooperative Groups 是在 CUDA 9 中用于组织通信线程组。

• Cooperative Groups 允许开发人员表达线程之间的沟通粒度，帮助他们更丰富、更有效地进行并行分解 Kepler 系列以来。

Volta 系列支持新的 Cooperative Launch API 和 Cooperative Groups 特性，通过该 API 可以实现 CUDA 线程块之间的同步；增加了对新的同步模式的支持。

7) 最大性能和最高效率两种模式：在最高性能模式下，Tesla V100 极速器将达到 300W 的 TDP（热设计功率）级别，满足那些需要最快计算速度和最高数据吞吐量的应用需求；在最高效率模式下，则允许数据中心管理员调整 Tesla V100 的功耗水平，以每瓦特最佳的能耗表现输出算力，Tesla V100 支持在所有 GPU 中设置上限功率，在大大降低功耗的同时，最大限度地满足机架的性能要求。

8) 针对 Volta 优化的软件：各新版本的深度学习框架（Caffe2，MXNet，CNTK，TensorFlow 等）都可以利用 Volta 大大缩短模型训练时间，同时提升多节点训练的性能。

Tesla V100 峰值运算性能如下（基于 GPU Boost 时钟频率）：

• 双精度浮点（FP64）运算性能：7.5 TFLOP/s；

• 单精度（FP32）运算性能：15 TFLOP/s;

• 混合精度矩阵乘法和累加：120 Tensor TFLOP/s。

• GV100 GPU 硬件架构：

1) 由许多图形处理集群（GPC）、纹理处理集群（TPC）、流式多处理器（SM）以及内存控制器组成。

2) 由 6 个 GPC、84 个 Volta SM、42 个 TPC（每个 TPC 包含 2 个 SM）和 8 个 512 位的内存控制器（共 4096 位）。其中，每个 SM 有 64 个 FP32 核、64 个 INT32 核、32 个 FP64 核与 8 个全新的 Tensor Core。同时，每个 SM 也包

含了 4 个纹理处理单元。

3) 总共包含了 5376 个 FP32 核、5376 个 INT32 核、2688 个 FP64 核、672 个 Tensor Core 以及 336 个纹理单元

4) GV100 总共包含 6144KB 的二级缓存（每个内存控制器都链接一个 768 KB 的 2 级缓存，每个 HBM2 DRAM 堆栈都由一对内存控制器控制）

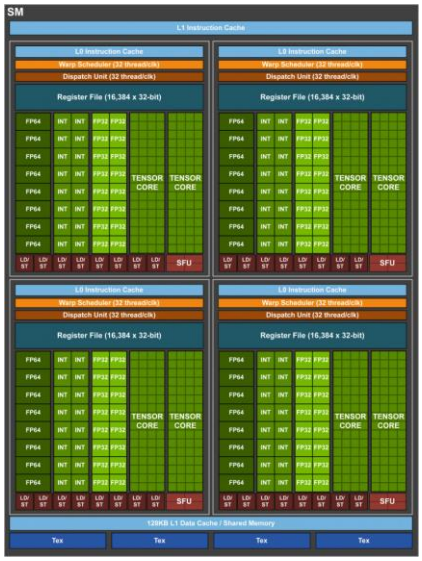
Tesla V100 与历代 Tesla 系列加速器的参数对比：

Tesla Product	Tesla K40	Tesla M40	Tesla P100	Tesla V100
GPU	GK180 (Kepler)	GM200 (Maxwell)	GP100 (Pascal)	GV100 (Volta)
SMs	15	24	56	80
TPCs	15	24	28	40
FP32 Cores / SM	192	128	64	64
FP32 Cores / GPU	2880	3072	3584	5120
FP64 Cores / SM	64	4	32	32
FP64 Cores / GPU	960	96	1792	2560
Tensor Cores / SM	NA	NA	NA	8
Tensor Cores / GPU	NA	NA	NA	640
GPU Boost Clock	810/875 MHz	1114 MHz	1480 MHz	1455 MHz
Peak FP32 TFLOP/s*	5.04	6.8	10.6	15
Peak FP64 TFLOP/s*	1.68	2.1	5.3	7.5
Peak Tensor Core TFLOP/s*	NA	NA	NA	120
Texture Units	240	192	224	320
Memory Interface	384-bit GDDR5	384-bit GDDR5	4096-bit HBM2	4096-bit HBM2
Memory Size	Up to 12 GB	Up to 24 GB	16 GB	16 GB
L2 Cache Size	1536 KB	3072 KB	4096 KB	6144 KB
Shared Memory Size / SM	16 KB/32 KB/48 KB	96 KB	64 KB	Configurable up to 96 KB
Register File Size / SM	256 KB	256 KB	256 KB	256KB
Register File Size / GPU	3840 KB	6144 KB	14336 KB	20480 KB
TDP	235 Watts	250 Watts	300 Watts	300 Watts
Transistors	7.1 billion	8 billion	15.3 billion	21.1 billion
GPU Die Size	551 mm²	601 mm²	610 mm²	815 mm²
Manufacturing Process	28 nm	28 nm	16 nm FinFET+	12 nm FFN

！关于组成细节分析：

Volta SM（流式多处理器）：

- 为深度学习矩阵计算建立的新型混合精度 FP16/FP32 Tensor Core；
- 为更高性能、更低延迟而强化的 L1 高速数据缓存；
- 为简化解码和缩短指令延迟而改进的指令集；
- 更高的时钟频率和能效。



Tensor Core（运算指令与数据格式）：

全新的 Tensor Core 是 Volta GV100 架构中最重要的一项新特性，可以为系统提供强劲的运算性能。Tesla V100 的 Tensor Core 可以为深度学习相关的模型训练和推断应用提供高达 120 TFLOPS 的浮点张量计算。Tesla V100 GPU 一共包含 640 个 Tensor Core，每个流式多处理器（SM）包含 8 个。

C. Summit 一个重要特征：AI 方面。

1、科学建模和仿真；

2、为 AI 与科学发现的整合提供可能性，允许研究人员将机器学习和深度学习应用于人类健康、高能物理、材料发现以及其他技术的相关问题中去。

3、随时准备代表美国能源部和 ORNL 对白宫所提出的人工智能相关任务回应。

D. 未来：

Summit 将会对特定的项目进行开放，可能涉及的领域包括天体物理、材料、癌症监测、系统生物学等等。