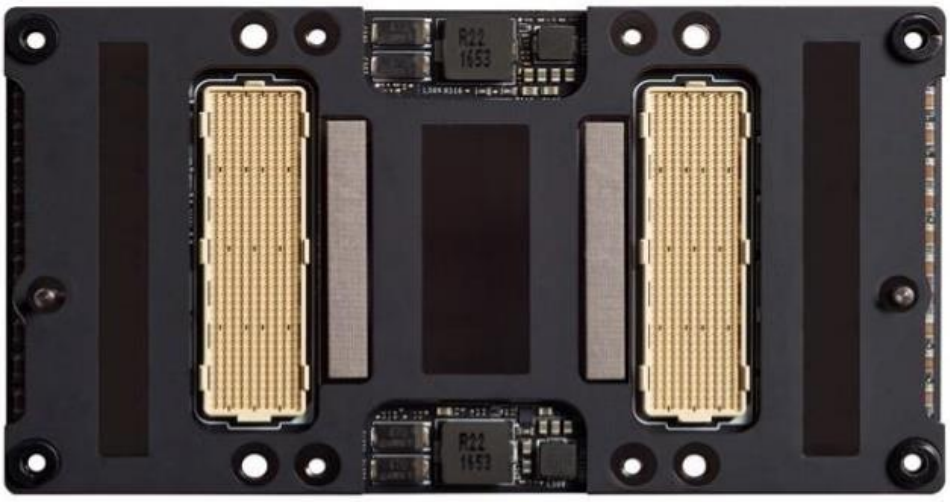
Tesla V100 GPU 架构分析

201708010814-李宗鸿

一、 简介

NVIDIA® Tesla® V100 Tensor Core 是有史以来极其先进的数据中心 GPU，能加快 AI、高性能计算 (HPC) 和图形技术的发展。其**采用 NVIDIA Volta 架构**，并带有 16 GB 和 32GB 两种配置，在单个 GPU 中即可提供高达 100 个 CPU 的性能。

Tesla V100 拥有 640 个 Tensor 内核，是世界上第一个突破 100 万亿次 (TFLOPS) 深度学习性能障碍的 GPU。新一代 NVIDIA NVLink™ 以高达 300 GB/s 的速度连接多个 V100 GPU，在全球打造出功能极其强大的计算服务器。



重要参数：

显存容量：16384MB
显存位宽：暂无数据
核心频率：1455MHz
显存频率：5012MHz
核心代号：HBM2
核心频率：1455MHz
显存类型：支持 DDR5
显存容量：16384MB
显存频率：5012MHz
显存带宽：900GB/s
DirectX 版本：DirectX 12
核心面积：815 平方毫米
晶体管数量：211 亿
流处理器数量：5120 个
单精度浮点性能：15 TFlops/S
双精度浮点性能：7.5 TFLOPS/s
发布日期：2017 年 05 月

Tesla V100 与过去五年历代 Tesla 加速器的参数对比：

Tesla Product	Tesla K40	Tesla M40	Tesla P100	Tesla V100
GPU	GK110 (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

二、 Tesla V100 的主要计算特征

- **为深度学习优化过的新型流式多处理器（SM）架构。**作为 GPU 处理器的核心组件，在

Volta 架构中 NVIDIA 重新设计了 SM，相比之前的 Pascal 架构而言，这一代 SM 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。专为深度学习设计的全新 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）。另外，由于全新的 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，因此在一般计算和寻址计算等混合场景下也能输出不错的效率。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。最后，一个新组合的 L1 高速数据缓存和共享内存子系统也显著提高了性能，同时大大简化了开发者的编程步骤。

- **第二代 NVLink。**第二代 NVIDIA NVLink 高速互连技术为多 GPU 和多 GPU/CPU 系统配置提供了更高的带宽，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 还支持基于 IBM Power 9 CPU 服务器的 CPU 控制和高速缓存一致性功能。

- **HBM2 显存：**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。相比上一代 Pascal GP100，来自三星的新一代 HBM2 内存与 Volta 的新一代内存控制器相结合，带宽提升 1.5 倍，并且在性能表现上也超过了 95% 的工作负载。

- **Volta 多处理服务。**Volta MPS 是 Volta GV100 架构的一项新特性，可以提供 CUDA MPS 服务器关键组件的硬件加速功能，从而在共享 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（QoS）。Volta MPS 还将 MPS 支持的客户端最大数量从 Pascal 时代的 16 个增加到 48 个。

- **增强统一存储和地址转换服务。**Volta GV100 中的 GV100 统一存储（GV100 Unified Memory)技术包括新型访问计数器，让访问网页最频繁的处理器能更准确的迁移存储页。

- **协作组（Cooperative Groups）和新的 Cooperative Launch API。**Cooperative Groups 是在 CUDA 9 中引入的一种新的编程模型，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（decompositions）。Kepler 系列以来，所有的 NVIDIA GPU 都支持基本 Cooperative Groups 特性。Pascal 和 Volta 系列还支持新的 Cooperative Launch API，通过该 API 可以实现 CUDA 线程块之间的同步。另外 Volta 还增加了对新的同步模式的支持。

- **最大性能和最大效率模式。**顾名思义，在最高性能模式下，Tesla V100 极速器将无限制地运行，达到 300W 的 TDP（热设计功率）级别，以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。而最高效率模式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。而且，Tesla V100 还支持在所有 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。

- **为 Volta 优化过的软件。**各种新版本的深度学习框架（包括 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。各种 Volta 优化版本的 GPU 加速库（包括 cuDNN，cuBLAS 和 TensorRT 等）也都可以在 Volta GV100 各项新特性的支持下，为深度学习和 HPC 应用提供更好的性能支持。此外，NVIDIA CUDA Toolkit 9.0 版也加入了新的 API 和对 Volta 新特性的支持，以帮助开发者更方便地针对这些新特性编程。

三、 GV100 GPU 硬件架构

GV100 GPU 由多个图形处理集群（Graphics Processing Cluster，GPC）、纹理处理集群（Texture Processing Cluster，TPC）、流式多处理器（Streaming Multiprocessor，SM）以及内存控制器组成。

一个完整的 GV100 GPU 由 6 组 GPC 单元，每组 GPC 单元由 14 组 SM 单元构成，满血版应该是 6 x 14 = 84 组 SM 单元，但 Tesla V100 只有 80 组，每组 SM 单元 64 个 CUDA 单元，因此共同构成 80 x 64 = 5120 个 CUDA 单元。每组 SM 单元中，FP32：FP64：Tensor 单元比例为 8:4:1，64 个 FP32 核、64 个 INT32 核、32 个 FP64 核与 8 个 Tensor Core。同时，每个 SM 也包含了 4 个纹理处理单元（texture units）。

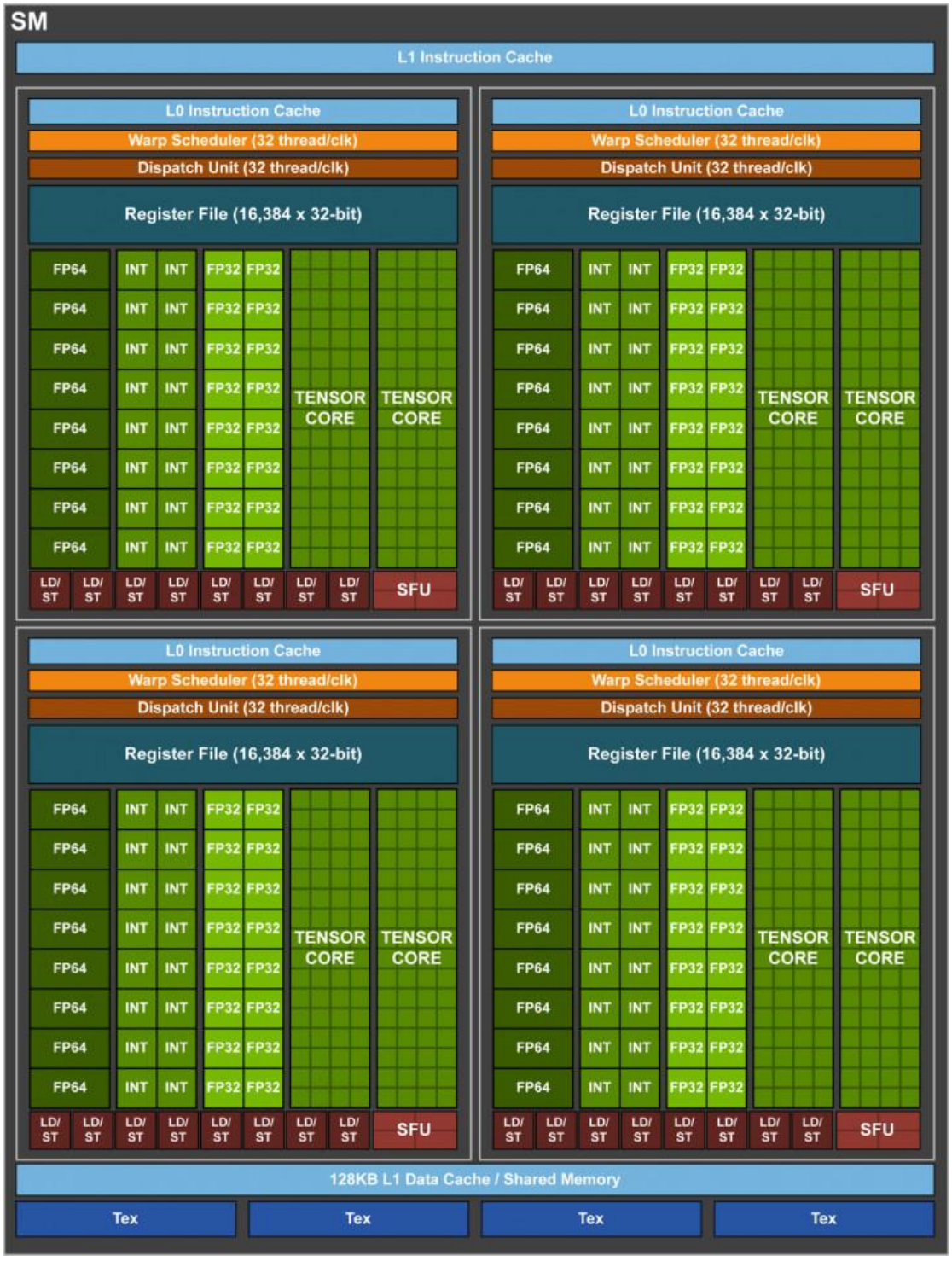
更具体地说，一个完整版 Volta GV100 中总共包含了 5376 个 FP32 核、5376 个 INT32 核、2688 个 FP64 核、672 个 Tensor Core 以及 336 个纹理单元。每个内存控制器都链接一个 768 KB 的 2 级缓存，每个 HBM2 DRAM 堆栈都由一对内存控制器控制。整体上，GV100 总共包含 6144KB 的二级缓存。下图展示了带有 84 个 SM 单元的完整版 Volta GV100

➢ Volta SM（流式多处理器）

为提供更高的性能而设计的架构，Volta SM 比过去的 SM 设计有更低的指令与缓存延迟，也包括加速深度学习应用的新特性：

- 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。

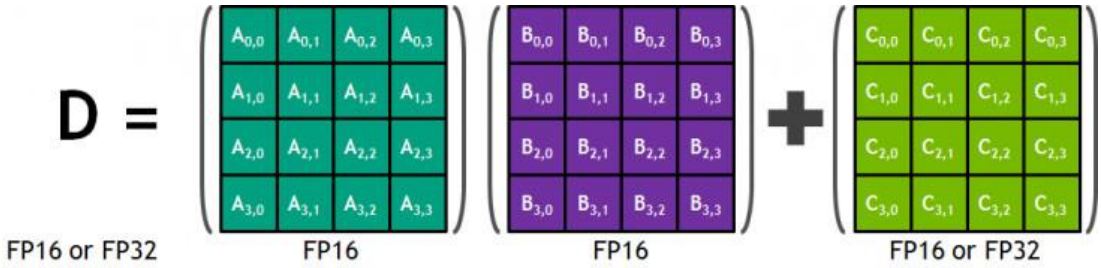
下图显示了 Volta GV100 SM 单元的基本结构。



➢ 新型混合精度 FP16/FP32 Tensor Core

全新的 Tensor Core 是 Volta GV100 架构中最重要的一项新特性，在训练超大型神经网络模型时，它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。具体来说，在深度学习的模型训练方面，相比于 P100 上的 FP32 操作，全新的 Tensor Core 可以在 Tesla V100 上实现最高 12 倍速的峰值 TFLOPS。而在深度学习的推断方面，相比于 P100 上的 FP16 操作，则可以实现最高 6 倍速的峰值 TFLOPS。Tesla V100 GPU 一共包含 640 个

Tensor Core，每个流式多处理器（SM）包含 8 个。
众所周知，矩阵乘法运算是神经网络训练的核心，在深度神经网络的每个连接层中，输入矩阵都要乘以权重以获得下一层的输入。在这个 Volta 架构中，每个 Tensor Core 都包含一个 4x4x4 的矩阵处理队列，来完成神经网络结构中最常见的 D=AxB+C 运算。其中 A、B、C、D 是 4 个 4×4 的矩阵，因此被称为 4x4x4。如下图所示，输入 A、B 是指 FP16 的矩阵，而矩阵 C 和 D 可以是 FP16，也可以是 FP32。



按照设计，Tensor Core 在每个时钟频率可以执行高达 64 次 FMA 混合精度浮点操作，也就是两个 FP16 输入的乘积，再加上一个 FP32。而因为每个 SM 单元都包含 8 个 Tensor Core，因此总体上每个时钟可以执行 1024 次浮点运算。这使得在 Volta 架构中，每个 SM 单元的深度学习应用吞吐量相比标准 FP32 操作的 Pascal GP100 大幅提升了 8 倍，与 Pascal P100 GPU 相比，Volta V100 GPU 的吞吐量总共提高了 12 倍。

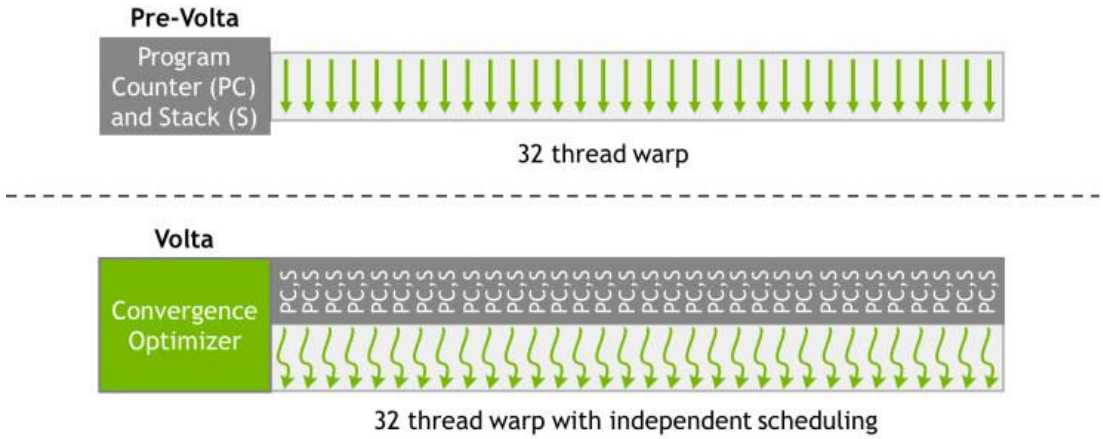
在程序执行期间，多个 Tensor Cores 通过 warp 单元协同工作。warp 中的线程同时还提供了可以由 Tensor Cores 处理的更大的 16x16x16 矩阵运算。CUDA 将这些操作作为 Warp-Level 级的矩阵运算在 CUDA C++ API 中公开。通过 CUDA C++ 编程，开发者可以灵活运用这些开放 API 实现基于 Tensor Cores 的乘法、加法和存储等矩阵操作。

➢ 增强的 L1 高速数据缓存和共享内存

Volta SM 的 L1 高速数据缓存和共享内存子系统相互结合，Volta 架构将数据高速缓存和共享内存功能组合到单个内存块中，在整体上为两种类型的内存访问均提供了最佳的性能。组合后的内存容量达到了 128 KB/SM，比老版的 GP100 高速缓存大 7 倍以上，并且所有这些都可以配置为不共享的独享 cache 块。另外，纹理处理单元也可以使用这些 cache。例如，如果共享内存被设置为 64KB，则纹理和加载/存储操作就可以使用 L1 中剩余的 64 KB 容量。

➢ Volta 架构的单指令多线程模式

Volta GV100 是第一个支持独立线程调度的 GPU，对每个线程，包括程序计数器和调用栈，Volta 都维护同一个执行状态。Volta 的独立线程调配机制允许 GPU 将执行权限让步于任何一个线程，这样做使线程的执行效率更高，同时也让线程间的数据共享更合理。为了最大化并行效率，Volta 有一个调度优化器，可以决定如何对同一个 warp 里的有效线程进行分组，并一起送到 SIMT 单元。这不仅保持了在 NVIDIA 之前的 GPU 里较高的 SIMT 吞吐量，而且灵活性更高：**现在，线程可以在 sub-warp 级别上分支和恢复，并且，Volta 仍将那些执行相同代码的线程分组在一起，让他们并行运行。**



四、 总结:

NVIDIA Tesla V100 无疑是目前世界上最先进的 GPU 之一，专门用于处理需要强大计算能力支持的密集型 HPC、AI、和图形处理任务。凭借最先进的 NVIDIA Volta 架构支持，Tesla V100 可以在单片 GPU 中提供 100 个 CPU 的运算性能，这使得数据科学家、研究人员和工程师们得以应对曾经被认为是不可能的挑战。搭载 640 个 Tensor cores，使得 Tesla V100 成为了目前世界上第一款突破 100 TFLOPS 算力大关的深度学习 GPU 产品。再加上新一代 NVIDIA NVLink 技术高达 300 GB/s 的连接能力，现实场景中用户完全可以将多个 V100 GPU 组合起来搭建一个强大的深度学习运算中心。这样，曾经需要数周时间的 AI 模型现在可以在几天之内训练完成。而随着训练时间的大幅度缩短，未来所有的现实问题或许都将被 AI 解决。

资料主要来源：https://devblogs.nvidia.com/inside-volta/