Tesla GPU 架构分析

1. GPU 架构

GPU 为 Graphics Processing Unit 的缩写，一般称为视觉处理单元。GPU 被广泛用于嵌入式系统、移动电话、个人电脑、工作站和电子游戏解决方案当中。现代的 GPU 对图像和图形处理是十分高效率的，这是因为 GPU 被设计为很高的并行架构这样使得比通用处理器 CPU 在大的数据块并行处理算法上更具有优势。

GPU 架构指的是硬件的设计方式，例如流处理器簇中有多少个 core、是否有 L1 or L2 缓存、是否有双精度计算单元等等。每一代的架构是一种思想，如何去更好完成并行的思想，而芯片就是对上述思想的实现。第一代的 GPU 架构的命名也是 Tesla，但现在基本已经没有这种设计的卡了。

NVIDIA 希望区分成三种选择，GeFore 用于家庭娱乐，Quadro 用于工作站，而 Tesla 系列用于服务器。

2. Pascal 架构

Pascal 是英伟达公司于 2016 年推出的新一代 GPU 架构，用于接替上一代的 Maxwell 架构。基于 Pascal 架构的 GPU 将会使用 16nm FinFET 工艺、HBM2、NVLink 2.0 等新技术 。Tesla P100 采用顶级大核心 GP100。

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

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。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。

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。

3. Volta 架构

新的 Volta GPU 架构的显著特征是它的 Tensor Core,新的 Tensor Core 是专门为深度学习设计的，有助于提高训练神经网络所需的性能。Tesla V100 的 Tensor Core 能够为训练、推理应用提供 120Tensor TFLOPS.相比于在 P100 FP32 上，在 Tesla V100 上进行深度学习训练有 12 倍的峰值 TFLOPS 提升。而在深度学习推理能力上，相比于 P100 FP16 运算，有了 6 倍的提升。

每个 Tensor Core 包含一个 4x4x4 的矩阵处理阵列来完成 D=A x B + C 的运算，其中 A、B、C、D 是 4×4 的矩阵，如下图所示。矩阵相乘的输入 A 和 B 是 FP16 矩阵，相加矩阵 C 和 D 可能是 FP16 矩阵或 FP32 矩阵。

每个 Tensor Core 每个时钟可执行 64 次浮点 FMA 混合精度运算（FP16 乘法与 FP32 累加），一个 SM 单元中的 8 个 Tensor Core 每个时钟可执行共计 1024 次浮点运算。相比于使用标准 FP32 计算的 Pascal GP100 而言，单个 SM 下的每个深度学习应用的吞吐量提升了 8 倍，所以这最终使得 Volta V100 GPU 相比于 Pascal P100 GPU 的吞吐量一共提升了 12 倍。Tensor Core 在与

FP32 累加结合后的 FP16 输入数据之上操作。FP16 的乘法得到了一个全精度结果，该结果在 FP32 和其他给定的 4x4x4 矩阵乘法点积的乘积运算之中进行累加。

在程序执行期间，多个 Tensor Core 通过一组 warp 线程的执行而同时使用。warp 内的线程提供了 Tensor Core 来处理大型 16×16×16 矩阵运算。CUDA 将这些操作作为 Warp-Level 矩阵运算在 CUDA C++ API 中公开。这些 C++接口提供了专门化的矩阵负载，如矩阵乘法和累加，矩阵存储操作可以有效地利用 CUDA C++程序中的 Tensor Core。

Tesla V100：人工智能计算和 HPC 的助推器

它的核心 GV100 GPU 包含 211 亿个晶体管，而芯片面积为前所未有的 815 平方毫米（Tesla GP100 为 610 平方毫米）。它采用了台积电（TSMC）的 12nm FFN 专属工艺打造。与其前身 GP100 GPU 及其他 Pascal 架构的显卡相比，GV100 提供了更强的计算性能，并增加了许多新功能。它进一步减小了 GPU 编程和应用程序移植难度，也通过制程的升级提高了 GPU 资源利用率。另外，GV100 也是一款能效极高的处理器，其在单位功耗的性能上表现卓越。

与前一代 Pascal GP100 GPU 类似，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 个 GPC、84 个 Volta SM、42 个 TPC（每个 TPC 包含了 2 个 SM）和 8 个 512 位的内存控制器（共 4096 位）。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。加上 84 个 SM，一个完整的 GV100 GPU 总共有 5376 个 FP32 核、5376 个 INT32 核、2688 个 FP64 核、672 个 Tensor Core 与 336 个纹理单元。每块内存控制器都连接了一个 768 KB 的 2 级缓存，每个 HBM2 DRAM 堆栈都由一对内存控制器控制。一个完整的 GV100 GPU 包括了总共 6144 KB 的二级缓存。图 4 展示了一个带有 84 个 SM 单元的完整 GV100 GPU（不同产品可以使用不同的 GV100 配置）。Tesla V100 加速器使用了 80 个 SM 单元。