TeslaGPU架构分析

智能1702 201708010528 莫诗幻

1. GPU简介

GPU缩写为Graphics Processing Unit的，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。

CPU和GPU架构差异很大，CPU功能模块很多，能适应复杂运算环境；而GPU构成则相对简单，目前流处理器和显存控制器占据了绝大部分晶体管。

CPU中大部分晶体管主要用于构建控制电路（比如分支预测等）和Cache，只有少部分的晶体管来完成实际的运算工作。而GPU的控制相对简单，且对Cache的需求小，所以大部分晶体管可以组成各类专用电路、多条流水线，使得GPU的计算速度有了突破性的飞跃，拥有了更强大的处理浮点运算的能力。

1. GPU的工作原理

GPU的图形（处理）流水线完成如下的工作：(并不一定是按照如下顺序)：

1）顶点处理：这阶段GPU读取描述3D图形外观的顶点数据并根据顶点数据确定3D图形的形状及位置关系，建立起3D图形的骨架。在支持DX8和DX9规格的GPU中，这些工作由硬件实现的VertexShader（定点着色器）完成。

2）光栅化计算：显示器实际显示的图像是由像素组成的，我们需要将上面生成的图形上的点和线通过一定的算法转换到相应的像素点。把一个矢量图形转换为一系列像素点的过程就称为光栅化。例如，一条数学表示的斜线段，最终被转化成阶梯状的连续像素点。

3）纹理帖图：顶点单元生成的多边形只构成了3D物体的轮廓，而纹理映射（texturemapping）工作完成对多变形表面的帖图，通俗的说，就是将多边形的表面贴上相应的图片，从而生成“真实”的图形。TMU（Texturemapping unit）即是用来完成此项工作。

4）像素处理：这阶段（在对每个像素进行光栅化处理期间）GPU完成对像素的计算和处理，从而确定每个像素的最终属性。在支持DX8和DX9规格的GPU中，这些工作由硬件实现的Pixel Shader（像素着色器）完成最终输出，由ROP（光栅化引擎）最终完成像素的输出，1帧渲染完毕后，被送到显存帧缓冲区。

1. Tesla GPU的特点

NVIDIA®Tesla®V100TensorCore是有目前最先进的数据中心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16GB和32GB两种配置，在单个GPU中即可提供高达100个CPU的性能。可以减少优化内存使用率的时间，提高效率。

Tesla GPU的20系列产品家族基于代号为“Fermi”的下一代CUDA架构，支持技术与企业计算所“必备”的诸多特性，其中包括C++支持、可实现极高精度与可扩展性的ECC存储器以及7倍于Tesla 10系列GPU的双精度性能。Tesla C2050与C2070 GPU旨在重新定义高性能计算并实现超级计算的平民化。与最新的四核CPU相比，Tesla C2050与C2070计算处理器以十分之一的成本和二十分之一的功耗即可实现同等超级计算性能。

1. 为深度学习优化过的新型流式多处理器（SM）架构。Volta 对 GPU 核心的 SM 处理器架构进行了重要的重新设计。新的 Volta SM 架构比前代 Pascal 设计能效高 50%，在同样的功率范围下 FP32 和 FP64 性能有重大提升。新的 Tensor Core 是专门为深度学习设计的，为浮点运算速度带来了 12 倍的提升。有了独立的、并行的整型和浮点型数据通路，Volta SM 在负载上也更高效，混合了计算与地址运算。Volta 新的独立线程调度能力使得并行线程之间的细粒度同步协同（finer-grain synchronization and cooperation）成为可能。最终，新型的 L1 Data Cache 与 Shared Memory 子系统的结合也能极大地提升性能，同时还简化了编程。
2. 第二代 NVLink。第二代英伟达 NVLink 高速互连技术能提供更高的带宽、更多连接，同时还改进了多 GPU 和多 GPU/CPU 系统配置的延展性。
3. HBM2 显存：更快、更高效。Volta 高度调整的 16GB HBM2 显存子系统提供了 900 GB/s 的峰值显存带宽。来自三星的新一代 HBM2 显存和 Volta 中的新一代显存控制器的组合实现的显存带宽是 Pascal GP100 的 1.5 倍，而且在许多负载上的显存带宽效率更高。Volta 多处理服务。
4. Volta 多服务处理（MPS：Multi-Process Service）是 Volta GV100 的一项新特性，能够为 CUDA MPS 服务器的关键组件提供硬件加速，从而能为共享该 GPU 的多个计算应用提供更高的性能、隔离和更好的服务质量（QoS）。Volta MPS 还将 MPS 客户端的最大数量从 Pascal 的 16 提升到了 Volta 的 48。
5. 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。
6. 协作组是 CUDA 9 中新的编程模型，用来组织通信线程组。Volta 增加了对新型同步模式的支持。最大性能和最大效率模式。在最大性能模式下，Tesla V100 加速器将不受限制的把 TDP（热设计功耗）水平提高到 300W，从而加速需要最快计算速度和最高数据吞吐的应用。最大效率模式下，数据中心管理员可以调整 Tesla V100 加速器的功率使用，从而用单位功耗下最优的性能进行运算。
7. 为 Volta 优化过的软件。Caffe2、MXNet、CNTK、TensorFlow 等这样的深度学习框架的新版本，能够利用 Volta 的性能来获得更快的训练速度、更高的多节点训练性能。GPU 加速库（比如 cuDNN、cuBLAS 等）的 Volta 优化版本利用 Volta GV100 架构的新特性能为深度学习和高性能计算应用提供更高的性能。
8. Tesla GPU（NVIDIA Volta）架构

1、特斯拉V100 拥有640个Tensor core：

Tensor core是一种新型处理核心，它执行一种专门的矩阵数学运算，适用于深度学习和某些类型的HPC。Tensor Core执行融合乘法加法，其中两个4\*4 FP16矩阵相乘，然后将结果添加到4\*4 FP16或FP32矩阵中，最终输出新的4\*4 FP16或FP32矩阵。Tensor core与标准的ALU流水线并没有太大差别，只不过Tensor Core处理的是大型矩阵运算，而不是简单地单指令流多数据流标量运算。Tensor Core是灵活性和吞吐量权衡的选择，它在执行标量运算时的表现很糟糕，但它可以将更多的操作打包到同一个芯片区域。特斯拉是世界上第一个突破深度学习性能的100teraFLOPS（TFLOPS）障碍的GPU。下一代NVIDIA NVLink™可以以高达300 GB / s的速度连接多个V100 GPU。

在架构设计方面，Volta架构的每个SM被划分为4四个区块，每个区块包含16个FP32计算单元、8个FP64计算单元、16个INT32计算单元、2个新的混合精度张量单元和一个新的L0指令缓存、1个warp调度器、1个调度单元和1个64KB文件寄存器。相比之下，上代Pascal架构中，GP100核心的每个SM只有2个区块，每个区块包含了32个FP32计算单元、16个FP64计算单元、1个指令缓冲区、1个warp调度器、2个调度单元和一个128KB的文件寄存器。

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。

在Tensor Core执行实际指令时，即使在使用NVVM IR（LLVM）的编译器级别上，也仅存在用于warp级矩阵操作的本征，对于CUDA++和PTX ISA，warp级别仍然是唯一级别。加载输入矩阵的形式是每个扭曲线程持有一个片段，其分布和身份均未指定。从广义上讲，它遵循标准CUDA核心的基于线程级别拼接的GEMM计算的相同模式。

一般而言，给定AB+C Tensor Core操作，片段由A的8个FP162元素（即16个FP16元素）和B的另外8个FP162元素，以及FP16累加器的4个FP162元素或 FP32累加器的8个FP32元素组成。

在矩阵乘法累加运算之后，计算结果会分散在每个线程的目标寄存器片段中，需要在整个范围内统一，如果其中一个warp线程退出，这些低级操作基本上就会失败。

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。除 CUDA C++接口可直接编程 Tensor Core 外，CUDA 9 cuBLAS 和 cuDNN 库还包含了使用 Tensor Core 开发深度学习应用和框架的新库接口。

1. Volta SM 的 L1 数据缓存和共享显存子系统的组合能显著提高性能，同时也简化了编程并减少了达到或接近峰值应用性能所需的时间成本。
2. 独立的线程调配

Volta 架构旨在设计为比以前的 GPU 更容易编程，令用户能在更复杂和多样的应用程序上高效地工作。Volta GV100 是第一款支持独立线程调配的 GPU，其在并行线程指令中可以实现细粒度（finer-grain）的同步和协作。Volta 主要的设计目标是减少在 GPU 中运行指令所需的工作量，并在线程合作中实现更大的灵活度，这样从而为细粒度并行算法提供更高的效率。英伟达同时也展示了他们如何对 SIMT（单指令多线程）做出重大改进以推进 Volta 架构。32 线程内单个独立的 CUDA 核现在只有有限的自主性；线程现在可以在一个细粒度层面上进行同步，并且仍然处于 SIMT 范式下，所以这就意味着更高的整体效率。更重要的是，独立的线程现在可以放弃再重新安排在一起。这就意味着英伟达的 GPU 有一定数量的调度硬件（scheduling hardware）。