
湖南大學

HUNAN UNIVERSITY

# 计算机系统设计

选题名称：Tesla GPU 架构分析

姓　　名：肖若愚

学　　号：201708010619

专业班级：物联 1702

# 一、英伟达 GPU 发展历史

NVIDIA GPU 架构历经多次变革，从起初的 Tesla 发展到最新的 Turing 架构，发展史可分为以下时间节点：

- **2008 - Tesla**

  Tesla 最初是给计算处理单元使用的，应用于早期的 CUDA 系列显卡芯片中，并不是真正意义上的普通图形处理芯片。

- **2010 - Fermi**

  Fermi 是第一个完整的 GPU 计算架构。首款可支持与共享存储结合纯 cache 层次的 GPU 架构，支持 ECC 的 GPU 架构。

- **2012 - Kepler**

  Kepler 相较于 Fermi 更快，效率更高，性能更好。

- **2014 - Maxwell**

  其全新的立体像素全局光照 (VXGI) 技术首次让游戏 GPU 能够提供实时的动态全局光照效果。基于 Maxwell 架构的 GTX 980 和 970 GPU 采用了包括多帧采样抗锯齿 (MFAA)、动态超级分辨率 (DSR)、VR Direct 以及超节能设计在内的一系列新技术。

- **2016 - Pascal**

  Pascal 架构将处理器和数据集成在同一个程序包内，以实现更高的计算效率。1080 系列、1060 系列基于 Pascal 架构

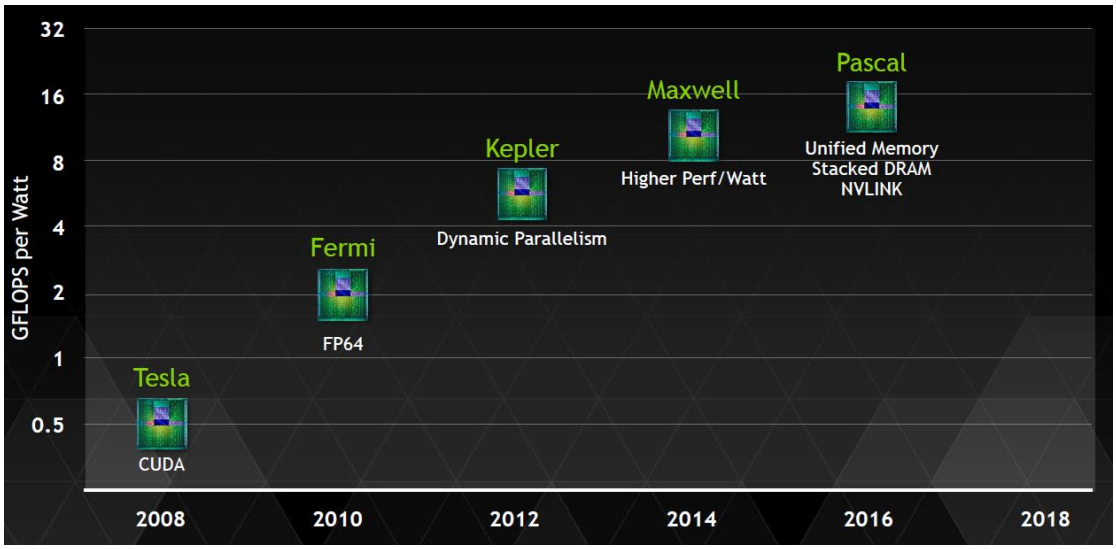
- **2017 - Volta**

  Volta 配备 640 个 Tensor 核心，每秒可提供超过 100 兆次浮点运算(TFLOPS) 的深度学习效能，比前一代的 Pascal 架构快 5 倍以上。

- **2018 - Turing**

  Turing 架构配备了名为 RT Core 的专用光线追踪处理器，能够以高达每秒 10 Giga Rays 的速度对光线和声音在 3D 环境中的传播进行加速计算。Turing 架构将实时光线追踪运算加速至上一代 NVIDIA Pascal™ 架构的 25 倍，并能以高出 CPU 30 多倍的速度进行电影效果的最终帧渲染。2060 系列、2080 系列显卡也是跳过了 Volta 直接选择了 Turing 架构。

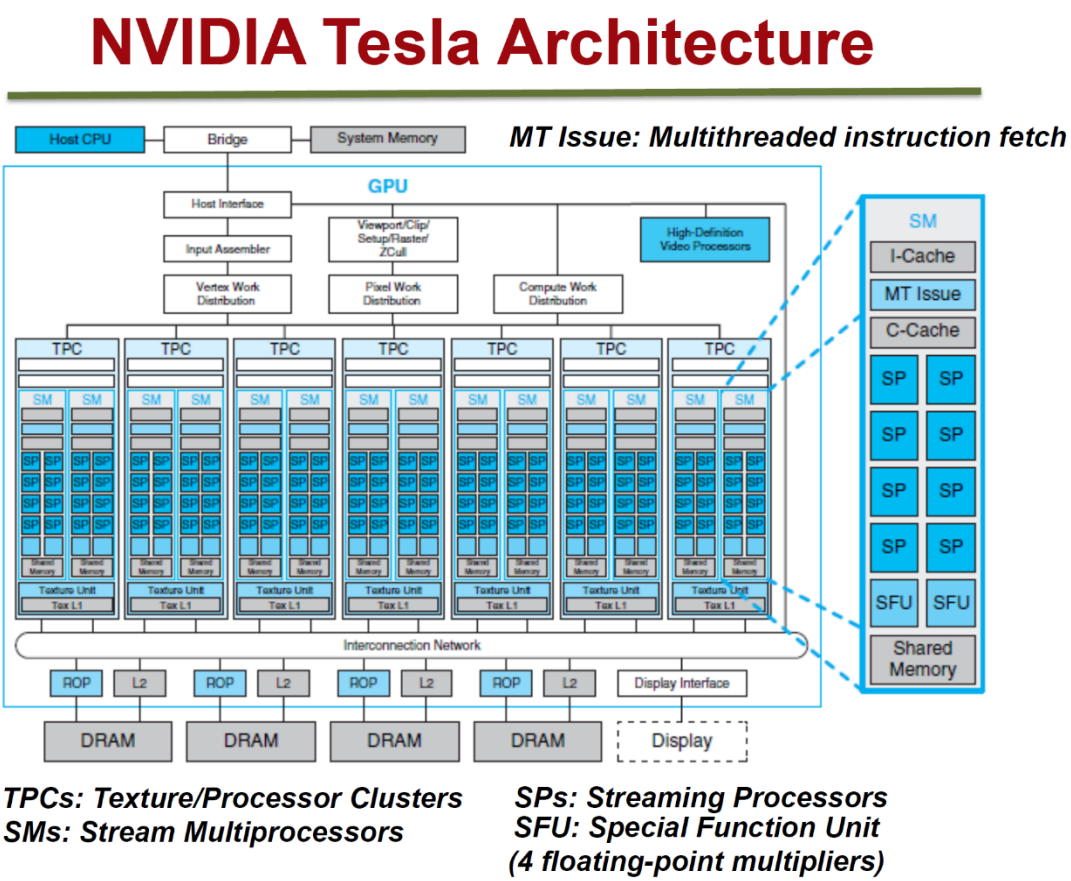
下图是部分 GPU 架构的发展历程：




英伟达 GPU 的发展

| 计算能力 | 微架构 | GPU核代 | 代表 |
|---|---|---|---|
| 1.0 | Tesla | G80 | GeForce 8800 Ultra |
| 1.1 | Tesla | G92, G94, G96, G98, G84, G86 | GeForce GTS 250, Quadro FX 4700 X2 |
| 1.2 | Tesla | GT218, GT216, GT215 | GeForce GT 340, *GeForce GT 330*,Quadro FX 380 Low Profile |
| 1.3 | Tesla | GT200, GT200b | GeForce GTX 295, Quadro FX 5800, Tesla C1060 |
| 2.0 | Fermi | GF100, GF110 | GeForce GTX 590, GeForce GTX 580,Quadro 6000,Tesla C2075 |
| 2.1 | Fermi | GF104, GF106 GF108, GF114, GF116, GF117, GF119 | GeForce GTX 560 Ti, GeForce GTX 550 Ti, Quadro 2000, Quadro 2000D |
| 3.0 | Kepler | GK104, GK106, GK107 | GeForce GTX 770, GeForce GTX 760,Quadro K5000,Tesla K10 |
| 3.2 | Kepler | GK20A | Tegra K1, Jetson TK1 |
| 3.5 | Kepler | GK110, GK208 | GeForce GTX Titan Z, GeForce GTX Titan Black, GeForce GTX Titan, GeForce GTX 780 Ti,Quadro K6000, Tesla K40 |
| 3.7 | Kepler | GK210 | Tesla K80 |
| 5.0 | Maxwell | GM107, GM108 | GeForce GTX 750 Ti, Quadro K1200, Quadro K620, Quadro M2000M, Tesla M10 |
| 5.2 | Maxwell | GM200, GM204, GM206 | GeForce GTX Titan X, GeForce GTX 980 Ti, Quadro M3000M, Tesla M4, Tesla M40 |
| 5.3 | Maxwell | GM20B | Tegra X1, Jetson TX1, |
| 6.0 | Pascal | GP100 | Quadro GP100, Tesla P100 |
| 6.1 | Pascal | GP102, GP104, GP106, GP107, GP108 | Titan X, GeForce GTX 1080 Ti,Tesla P40, Tesla P6, Tesla P4,Quadro P6000 |
| 6.2 | Pascal | GP10B | Drive PX2 with Tegra X2 |
| 7.0 | Volta | GV100 | NVIDIA TITAN V, Tesla V100 |

## 二、Tesla 架构




Tesla 微观架构总览图如上。下面将阐述它的特性和概念：

- 拥有 7 组 TPC（Texture/Processor Cluster，纹理处理簇）
- 每个 TPC 有两组 SM（Stream Multiprocessor，流多处理器）
- 每个 SM 包含：
  - 6 个 SP（Streaming Processor，流处理器）
  - 2 个 SFU（Special Function Unit，特殊函数单元）
  - L1 缓存、MT Issue（多线程指令获取）、C-Cache（常量缓存）、共享内存
- 除了 TPC 核心单元，还有与显存、CPU、系统内存交互的各种部件。

Tesla 体系架构是一块具有可扩展处器数量的处理器阵列。每个 GT200 GPU 包含 240 个流处理器（streaming processor,SP），每 8 个流处理器又组成了一个流多处理器(streaming multiprocessor,SM)，因此共有 30 个流多处理器。GPU 在工作时，工作负载由 PCI-E 总线从 CPU 传入 GPU 显存，按照体系架构的层次自顶向下分发。PCI-E 2.0 规范中，每个通道上下行的数据传输速度达到了 5.0Gbit/s，这样 PCI-E2.0×16 插槽能够为上下行数据各提供了 5.0*16Gbit/s=10GB/s 的带宽，故有效带宽为 8GB/s,而 PCI-E 3.0 规范的上下行数据带宽各为 20GB/s。但是由于 PCI-E 数据封包的影响，实际可用的带宽大约在 5-6GB/s（PCI-E 2.0 × 16）。 Normal 0 7.8 磅 0 2 false false false EN-US ZH-CN X-NONE

在 GT200 架构中，每 3 个 SM 组成一个 TPC（Thread Processing Cluster，线程处理器集群），而在 G80 架构中，是两个 SM 组成一个 TPC，G80 里面有 8 个 TPC，因为 G80 有 128(2*8*8)个流处理器，而 GT200 中 TPC 增加到了 10(3*10*8)个，其中，每个 TPC 内部还有一个纹理流水线。

大多数时候，称呼 streaming processor 为流处理器，其实并不太正确，因为如果称 streaming processor 为流处理器的话，自然是隐式的与 CPU 相对，但是 CPU 有独立的一套输入输出机构，而 streaming processor 并没有，不能在 GPU 编程中使用 printf 就是一个例证。将 SM 与 CPU 的核相比更加合适。和现在的 CPU 的核一样，SM 也拥有完整前端。
GT200 和 G80 的每个 SM 包含 8 个流处理器。流处理器也有其他的名称，如线程处理器，“核”等，而最新的 Fermi 架构中，给了它一个新的名称:CUDA Core。 SP 并不是独立的处理器核，它有独立的寄存器和程序计数器(PC)，但没有取指和调度单元来构成完整的前端（由 SM 提供）。因此，SP 更加类似于当代的多线程 CPU 中的一条流水线。SM 每发射一条指令，8 个 SP 将各执行 4 遍。因此由 32 个线程组成的线程束（warp）是 Tesla 架构的最小执行单位。由于 GPU 中 SP 的频率略高于 SM 中其他单元的两倍，因此每两个 SP 周期 SP 才能对片内存储器进行一次访问，所以一个 warp 中的 32 个线程又可以分为两个 half-warp，这也是为什么取数会成为运算的瓶颈原因。Warp 的大小对操作延迟和访存延迟会产生影响，取 Warp 大小为 32 是 NVIDIA 综合权衡的结果。

SM 最主要的执行资源是 8 个 32bit ALU 和 MAD（multiply-add units，乘加器）。它们能够对符合 IEEE 标准的单精度浮点数（对应 float 型）和 32-bit 整数（对应 int 型，或者 unsigned int 型）进行运算。每次运算需要 4 个时钟周期（SP 周期，并非核心周期）。因为使用了四级流水线，因此在每个时钟周期，ALU 或 MAD 都能取出一个 warp 的 32 个线程中的 8 个操作数，在随后的 3 个时钟周期内进行运算并写回结果。

每个 SM 中，还有一个共享存储器(Shared memory),共享存储器用于通用并行计算时的共享数据和块内线程通信，但是由于它采用的是片上存储器，其速度极快，因此也被用于优化程序性能。

每个 SM 通过使用两个特殊函数(Special Function Unit,SFU)单元进行超越函数和属性插值函数（根据顶点属性来对像素进行插值）计算。SFU 用来执行超越函数、插值以及其他特殊运算。SFU 执行的指令大多数有 16 个时钟周期的延迟，而一些由多个指令构成的复杂运算，如平方根或者指数运算则需要 32 甚至更多的时钟周期。SFU 中用于插值的部分拥有若干个 32-bit 浮点乘法单元，可以用来进行独立于浮点处理单元(Float Processing Unit,FPU)的乘法运算。SFU 实际上有两个执行单元，每个执行单元为 SM 中 8 条流水线中的 4 条服务。向 SFU 发射的乘法指令也只需要 4 个时钟周期。
在 GT200 中，每个 SM 还有一个双精度单元，用于双精度计算，但是其计算能力不到单精度的 1/8。

控制流指令（CMP,比较指令）是由分支单元执行的。GPU 没有分支预测机制，因此在分支得到机会执行之前，它将被挂起，直到所有的分支路径都执行完成，这会极大的降低性能。

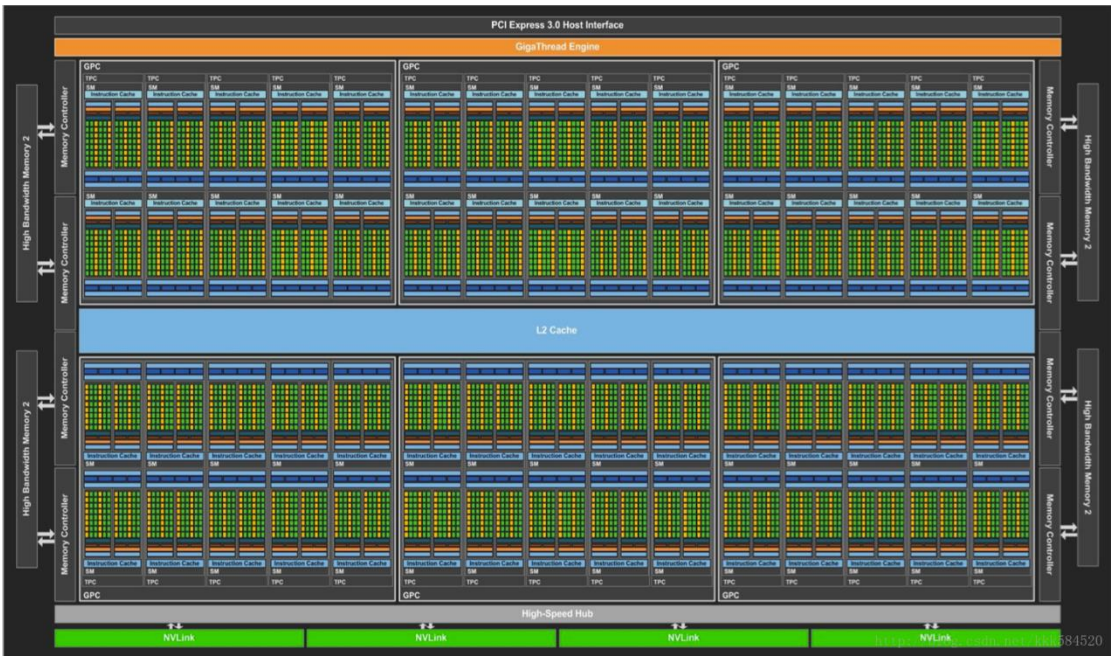
# 三、Tesla P100 架构分析

Tesla P100 采用的是 Pascal 架构

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。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。

进一步细看 GP100 SM 的架构。

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。


图中为一个 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。

Pascal SM 架构图中可以看到，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。

一个 warp 调度器（每个处理块共享一个）在一个时钟周期内可以分发两个 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。