Tesla GPU 架构分析

·介绍

显卡巨头 Nvidia 公司的产品可以分为三大类：科学计算卡（Tesla）、专业图形卡（Quadro）和家用显卡（Geforce）类。NVIDIA GeForce 和 NVIDIA Quadro 分别是为消费级图形处理和专业可视化而设计的，只有 Tesla 产品系列是完全针对并行计算而设计的，可提供独有的计算特性。

·GPU 的功能：

现代 GPU 除了绘制图形外，还担当了很多额外的功能，综合起来如下几方面：

图形绘制。

这是 GPU 最传统的拿手好戏，也是最基础、最核心的功能。为大多数 PC 桌面、移动设备、图形工作站提供图形处理和绘制功能。

物理模拟。

GPU 硬件集成的物理引擎（PhysX、Havok），为游戏、电影、教育、科学模拟等领域提供了成百上千倍性能的物理模拟，使得以前需要长时间计算的物理模拟得以实时呈现。

海量计算。

计算着色器及流输出的出现，为各种可以并行计算的海量需求得以实现，CUDA 就是最好的例证。

AI 运算。

近年来，人工智能的崛起推动了 GPU 集成了 AI Core 运算单元，反哺 AI 运算能力的提升，给各行各业带来了计算能力的提升。

其它计算。

音视频编解码、加解密、科学计算、离线渲染等等都离不开现代 GPU 的并行计算能力和海量吞吐能力。

·Tesla 架构

Tesla 微观架构总览图如下。下面阐述它的特性和概念：

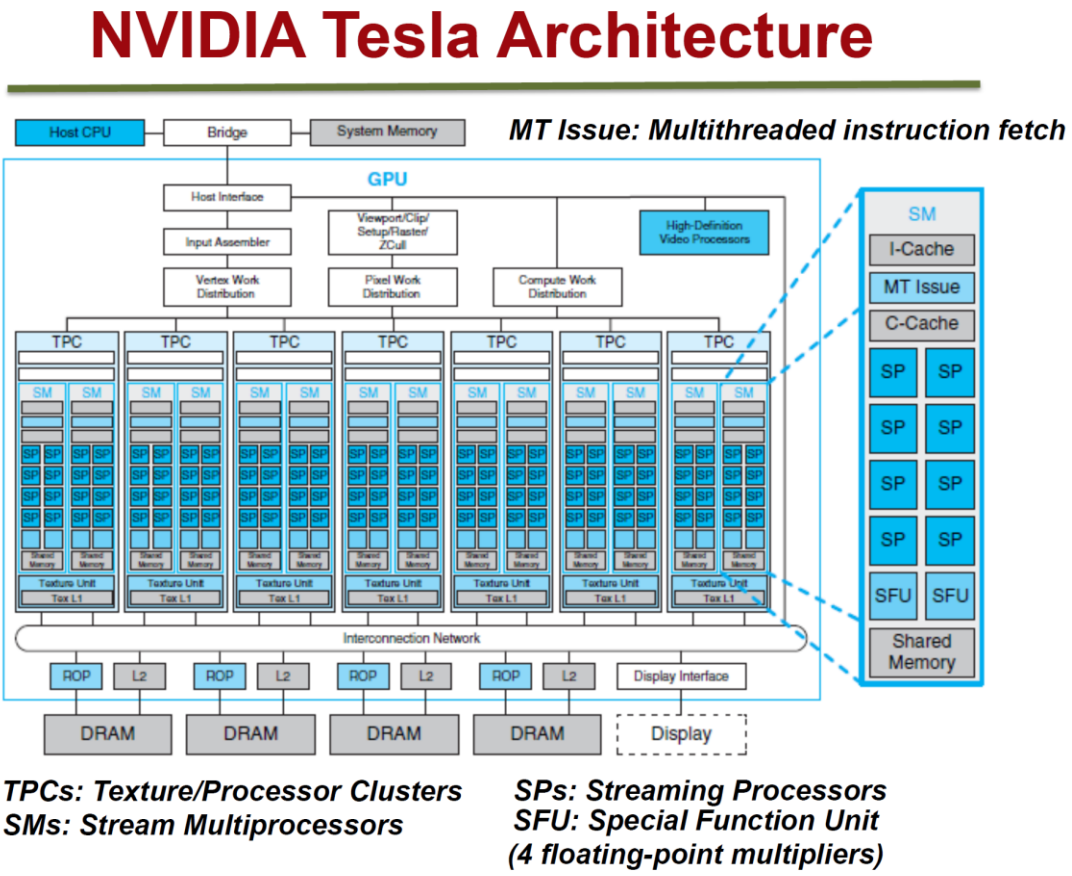
拥有 7 组 TPC（Texture/Processor Cluster，纹理处理簇）

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。



不同型号 Tesla 架构 GPU 比较：

VOLTA 與 PASCAL GPU

	TESLA V100	TESLA P100	TESLA P40	TESLA P4	TESLA P6
GPU	1 張 NVIDIA Volta GPU	1 張 NVIDIA Pascal GPU	1 張 NVIDIA Pascal GPU	1 張 NVIDIA Pascal GPU	1 張 NVIDIA Pascal GPU
CUDA 核心	5,120 個	3,584 個	3,840 個	2,560 個	2,048 個
記憶體大小	16 GB HBM2	16 GB HBM2	24 GB GDDR5	8 GB GDDR5	16 GB GDDR5
H.264 1080p30 串流	36	36	24	24	24
使用案例	高階專業繪圖使用者，包括雙精度運算工作負載 (3D 模型和設計工作流程與密集 CAE 模擬作業)		中階到高階專業繪圖使用者，包括單精度運算工作負載 (渲染和創意複雜設計)；	入門到中階專業繪圖使用者，包括深度學習推論工作負載和 Pascal 功能，2 張 P4 是單張 M60 的合適升級途徑	適合刀鋒伺服器尺寸與從 M6 升級

·Tesla GPU 的优势

①FP64 双精度浮点计算能力强

对于专业卡而言，仅强调 FP32 单精度运算速度是不够的，毕竟进行生化模拟，比如化学分析和生物遗传学对数学精度的要求远远高于图形成像要求。要展示一个清晰的图像，我们使用能计算到小数点后 23 位的 Geforce 卡能满足。但是对于科学家而言，小数点后 23 位可能会产生误差，这种误差可能导致药物研发/航空探索等科学研究出现重大失误。这时

就需要双精度（FP64=52 位小数）进行更加精准的计算。下图为不同显卡的双精度浮点运算能力：

NVIDIA GPU Model	Double-precision (64-bit) Floating Point Performance
GeForce GTX Titan X Maxwell	up to 0.206 TFLOPS
GeForce GTX 1080 Ti	up to 0.355 TFLOPS
GeForce Titan Xp	up to 0.380 TFLOPS
GeForce Titan V	up to 6.875 TFLOPS
Tesla K80	1.87+ TFLOPS
Tesla P100*	4.7 ~ 5.3 TFLOPS
Quadro GP100	5.2 TFLOPS
Tesla V100*	7 ~ 7.8 TFLOPS
Quadro GV100	7.4 TFLOPS

②FP16 半精度计算能力强

半精度浮点计算通常应用于深度学习/人工智能应用中，同样是帕斯卡架构，只有 P100 完整核心的拥有完整的计算速度。下图是不同架构中半精度、单精度和双精度吞吐量对比，6.0 代表帕斯卡架构完整核心，7.0 代表最新的 volta 架构完整核心。从图中可以看出不同架构中完整核心都是支持所有精度计算模式的。

	Compute Capability					
	3.0, 3.2	3.5, 3.7	5.0, 5.2	5.3	6.0	6.1
, multiply-add	N/A	N/A	N/A	256	128	2
, multiply-add	192	192	128	128	64	128
, multiply-add	8	64	4	4	32	4

③ECC 内存的错误检测和纠正

在运行 3D 游戏的 GeForce 显卡上，即使出现一些内存错误通常也不会造成什么严重的问题，对于个人用户来说，显示的画面偶尔出现些许的错误完全可以容忍甚至会被忽视。但对于计算领域来说，就非常依赖于 GPU 返回数据的准确性，即使内存出现单比特错误也可能导致最终计算结果的极大误差。

GeForce 系列显卡不具备错误检测和纠正的功能， 但 Tesla 系列 GPU 因为 GPU 核心内部的寄存器、L1/L2 缓存和显存都支持 ECC 校验功能，所以 Tesla 不仅能检测并纠正单比特错误也可以发现并警告双比特错误，这对保证计算结果的准确性来说非常重要。

④GPU 内存性能

计算密集型应用程序不仅需要 GPU 提供高性能计算单元，也需要 GPU 提供快速访问数据的能力，否则再好的 GPU 核心也将成为巧妇难为无米之炊。 对于许多 HPC 应用程序，GPU 内存性能的差异对最终结果的影响甚至比计算能力更明显，Tesla GPU 可以提供比 GeForce GPU 更好的内存带宽：

型号	GPU内存带宽
GTX Titan X Pascal	480GB/s
GTX1080 Ti	484GB/s
Tesla P40	346GB/s
Tesla P100 12GB	549GB/s
Tesla P100 16GB	**743GB/s**

最后附上　　Tesla 20 系列 CUDA 架构主要参数：

尺寸规格	9.75英寸PCIe x16规格
Tesla GPU的数量	1
CUDA核心数量	448
CUDA核心频率	1.15 GHz
双精度浮点性能（峰值）	515 Gflops
单精度浮点性能（峰值）	1.03 Tflops
专用存储器总容量* **Tesla C2050** **Tesla C2070**	3GB GDDR5 6GB GDDR5
存储器频率	1.5 GHz
存储器接口	384位
存储器带宽	144 GB/秒
功耗	247W热设计功耗
系统接口	PCIe x16 Gen2
散热解决方案	主动式风扇散热器
软件开发工具	CUDA C/C++/Fortran、OpenCL以及DirectCompute工具包。针对Visual Studio的NVIDIA®（英伟达？）Parallel Nsight?