**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）

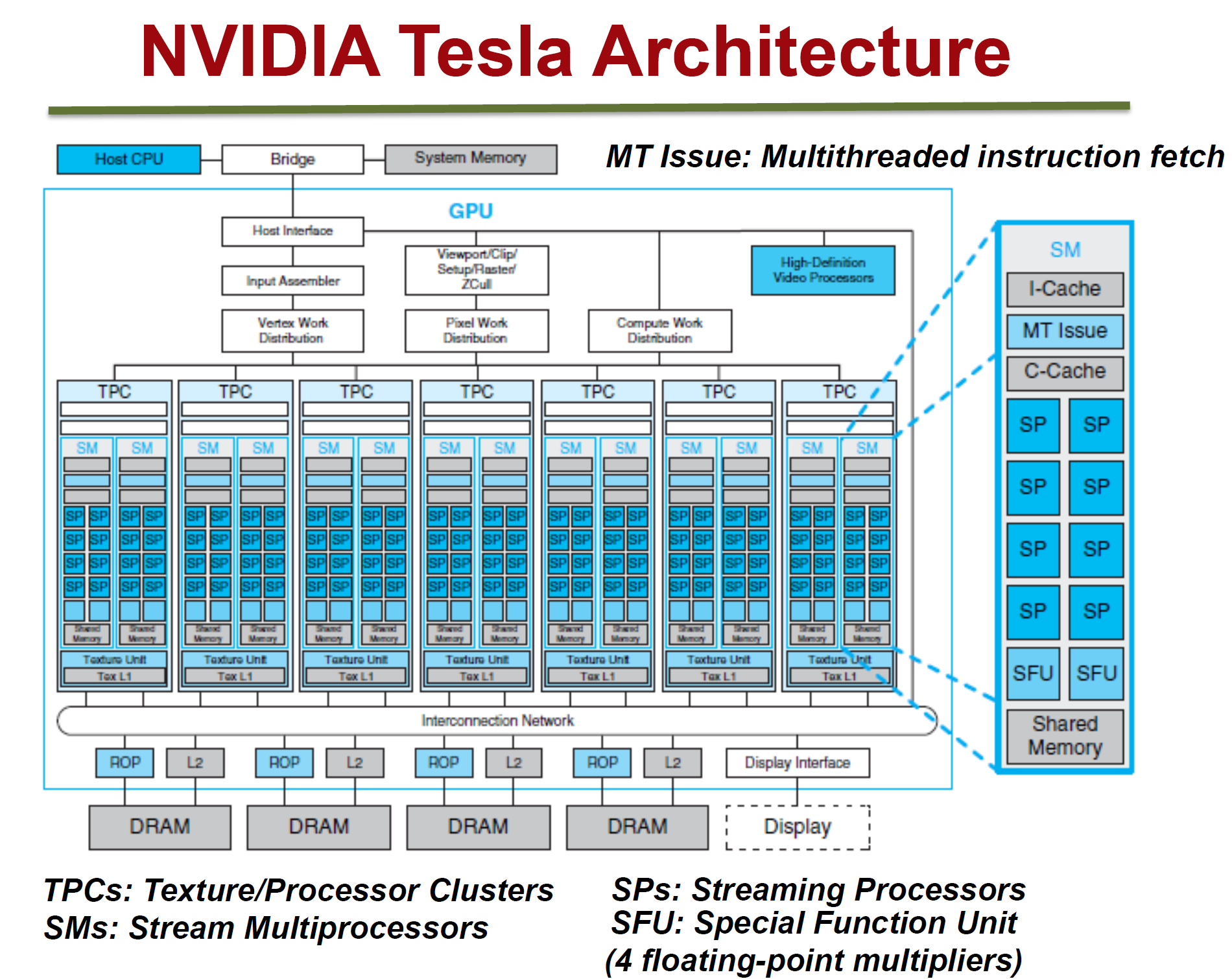
每个SM包含：

6个SP（Streaming Processor，流处理器）

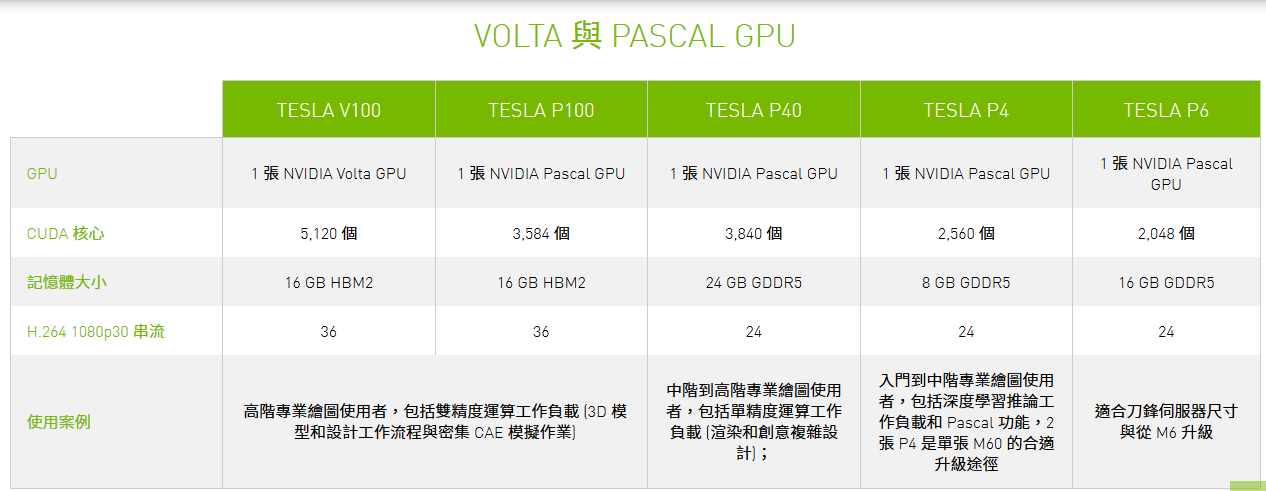
2个SFU（Special Function Unit，特殊函数单元）

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。



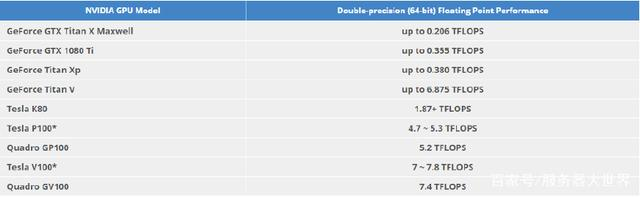
不同型号Tesla架构GPU比较：



**·Tesla GPU的优势**

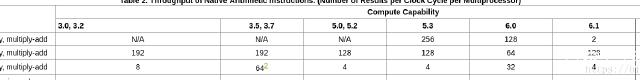
**①FP64双精度浮点计算能力强**

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：



**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。



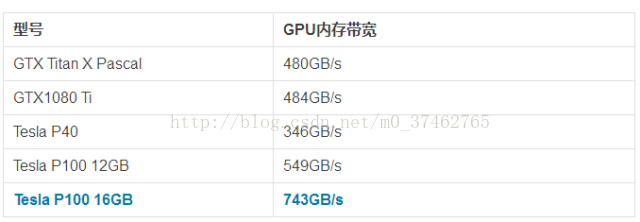
**③ECC内存的错误检测和纠正**

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：



最后附上 Tesla 20系列 CUDA架构主要参数：

