# Tesla GPU架构分析 #

## GPU简介 ##

图形处理单元GPU英文全称Graphic Processing Unit，GPU是相对于CPU的一个概念，NVIDIA公司在1999年发布GeForce256图形处理芯片时首先提出GPU的概念。GPU使显卡减少了对CPU的依赖，并进行部分原本CPU的工作(主要是并行计算部分)。GPU具有强大的浮点数编程和计算能力，在计算吞吐量和内存带宽上，现代的GPU远远超过CPU。

# GPU组成 #

GPU从大的方面来讲，就是由显存和计算单元组成：

显存（Global Memory）：显存是在GPU板卡上的DRAM，类似于CPU的内存，就是那堆DDR啊，GDDR5啊之类的。特点是容量大（可达16GB），速度慢，CPU和GPU都可以访问。

计算单元（Streaming Multiprocessor）：执行计算的。每一个SM都有自己的控制单元（Control Unit），寄存器（Register），缓存（Cache），指令流水线（execution pipelines）。

## Streaming Multiprocessor (SM)



在GP100里，每一个SM有两个SM Processing Block（SMP），里边的绿色的就是CUDA Core，CUDA core也叫Streaming Processor（SP）。每一个SM有自己的指令缓存，L1缓存，共享内存。而每一个SMP有自己的Warp Scheduler、Register File等。要注意的是CUDA Core是Single Precision的，也就是计算float单精度的。双精度Double Precision是那个黄色的模块。所以一个SM里边由32个DP Unit，由64个CUDA Core，所以单精度双精度单元数量比是2:1。LD/ST 是load store unit，用来内存操作的。SFU是Special function unit，用来做cuda的intrinsic function的，类似于\_\_cos()这种。

##Fermi架构

Fermi是第一个完整的GPU计算架构。

+ 512个accelerator cores即所谓CUDA cores（包含ALU和FPU）

+ 16个SM，每个SM包含32个CUDA  core

+六个384位 GDDR5 DRAM，支持6GB global on-board memory

+ GigaThread engine（图左侧）将thread blocks分配给SM调度

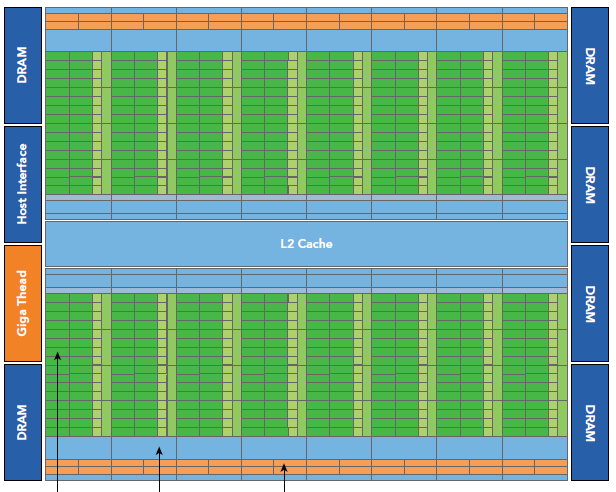
+ 768KB L2 cache

+每个SM有16个load/store单元，允许每个clock cycle为16个thread（即所谓half-warp，+不过现在不提这个东西了）计算源地址和目的地址

+ Special function units（SFU）用来执行sin cosine 等

+每个SM两个warp scheduler两个instruction dispatch unit，当一个block被分配到一个SM中后，所有该block中的thread会被分到不同的warp中。

+ Fermi（compute capability 2.x）每个SM同时可处理48个warp共计1536个thread。



每个SM由以下几部分组成：

+执行单元（CUDA cores）

+调度分配warp的单元

+ shared memory，register file，L1 cache