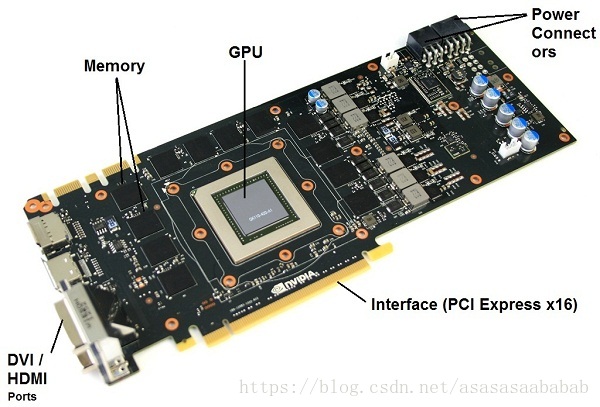
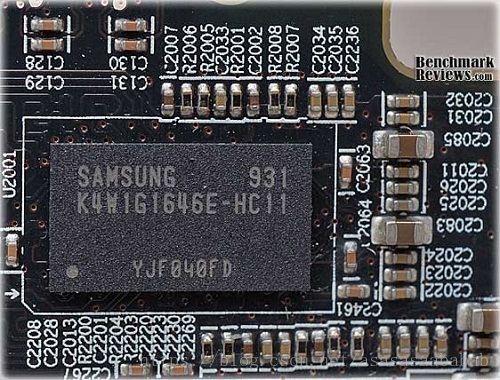
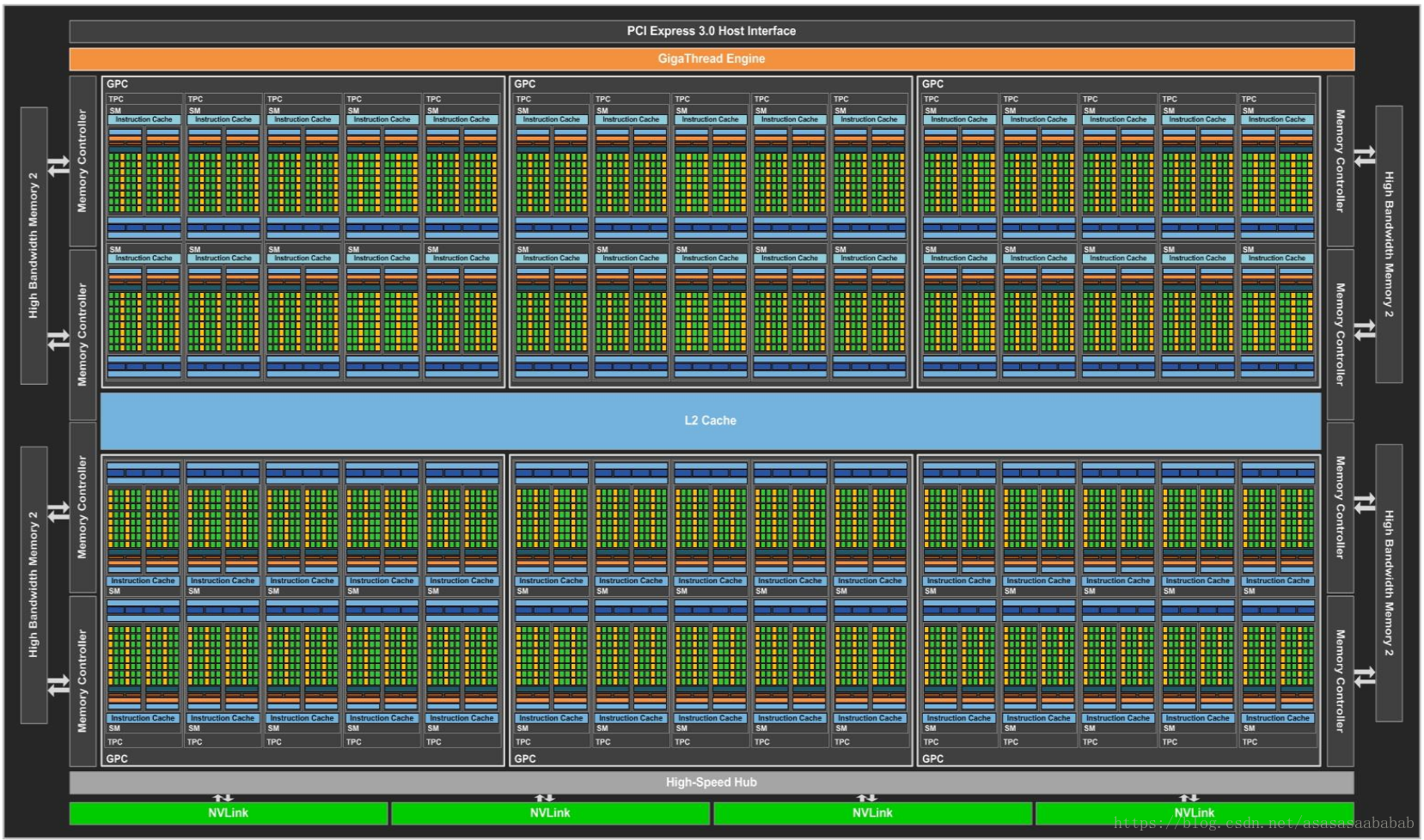
**背景**

  在深度学习大热的年代，并行计算也跟着火热了起来。深度学习变为可能的一个重要原因就是[算力的提升](https://blog.csdn.net/asasasaababab/article/details/78168187)。作为并行计算平台的一种，GPU及其架构本身概念是非常多的。下面就进行一个[概念阐述](http://www.icl.utk.edu/~luszczek/teaching/courses/fall2016/cosc462/pdf/GPU_Fundamentals.pdf)，以供参考。

**GPU：显存+计算单元**

  GPU从大的方面来讲，就是由显存和计算单元组成：

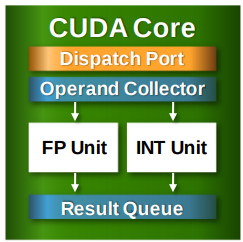
1. 显存（Global Memory）：显存是在GPU板卡上的DRAM，类似于CPU的内存，就是那堆DDR啊，GDDR5啊之类的。特点是容量大（可达16GB），速度慢，CPU和GPU都可以访问。
2. 计算单元（Streaming Multiprocessor）：执行计算的。每一个SM都有自己的控制单元（Control Unit），寄存器（Register），缓存（Cache），指令流水线（execution pipelines）。

  我们可以看一下[图](https://www.akshatblog.com/graphics-card-components-explained-in-detail/)：  
  
  
  所以其实Global Memory的真身就是那个内存颗粒。  
  来看下GPU里边的东西，是时候对密集恐惧症患者放出大招了：  
  
这个是英伟达PASCAL架构的GP100 GPU的架构。由非常多的Streaming Multiprocesser组成的（里边的各种外设先不说了）。下面我们看一下Streaming Multiprocessor的内容。

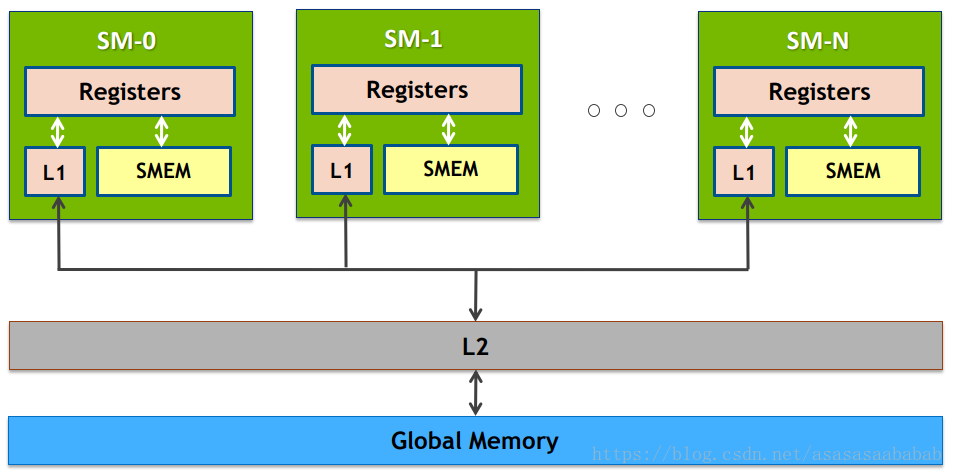
**Streaming Multiprocessor (SM)**

  下面这个图是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），这俩是一个意思。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。

**CUDA Core**

  下面这个图是CUDA Core的结构：  
  
  包括控制单元Dispatch Port、Operand Collector，以及浮点计算单元FP Unit、整数计算单元Int Unit，另外还包括计算结果队列。当然还有Compare、Logic、Branch等。相当于微型CPU。

**GPU内存架构**

  贴一张图：  
  
  越靠近SM的内存就越快。

1. L1 Cache：Pascal架构上，L1 Cache和Texture已经合为一体（Unified L1/Texture Cache），作为一个连续缓存供给warp使用。
2. L2 Cache：用来做Global Memory的缓存，容量大，给整个GPU使用。

**关于CUDA方面的一些参考文献**

  我发现Nvidia的文献非常分散，下面列举一些常用的。btw，PASCAL啊，VOLTA都是英伟达GPU架构代号。

1. [CUDA C Programming Guide](https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html)
2. [CUDA C Best Practices Guide](https://docs.nvidia.com/cuda/cuda-c-best-practices-guide/index.html)
3. [Pascal White Paper](https://images.nvidia.com/content/pdf/tesla/whitepaper/pascal-architecture-whitepaper.pdf), [Volta White Paper](https://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf)
4. [cuBLAS](https://docs.nvidia.com/cuda/cublas/index.html): 基础线性代数库，汇编级优化。
5. [cuDNN](https://docs.nvidia.com/deeplearning/sdk/cudnn-developer-guide/index.html): 深度学习库