

AI 芯片 - GPU 详解

GPU基础概念



ZOMI



BUILDING A BETTER CONNECTED WORLD

Ascend

www.hiascend.com

Talk Overview

I. AI 计算体系

- 深度学习计算模式
- 计算体系与矩阵运算

2. AI 芯片基础

- 通用处理器 CPU
- 从数据看 CPU 计算
- 通用图形处理器 GPU
- AI 专用处理器 NPU/TPU
- 计算体系架构的黄金10年

I. 硬件基础

- GPU 工作原理
- GPU AI 编程本质

2. 英伟达 GPU 架构

- 从 Fermi 到 Hopper 架构
- Tensor Core 和 NVLink 详解

3. GPU 图形处理流水线

- 图形流水线基础
- GPU 逻辑模块划分
- 图形处理算法到硬件

Talk Overview

I. 硬件基础

- GPU 工作原理
- GPU AI 编程本质

2. 英伟达 GPU 架构

- GPU 基础概念
- 从 Fermi 到 Volta 架构
- Turing 到 Hopper 架构
- Tensor Core 和 NVLink 详解

3. GPU 图形处理

- GPU 逻辑模块划分
- 算法到 GPU 硬件
- GPU 的软件栈
- 图形流水线基础
- 流水线不可编译单元
- 光线跟踪流水线

Talk Overview

I. GPU 基础概念

- SP , Stream Processor
- SM , Stream Multiprocessors
- Warp , Thread Warp
- CUDA Core
- CUDA
- 峰值算力

SP，流处理器 (Stream Processor)

- SP 被称作流处理器，在G80时英文也曾称（ Scalar Processors ）、（ Scalar Streaming Processors ）。这种称呼主要因为之前 SP 单元是被当作标量处理单元。 Fermi 架构后，SP被改称为CUDA Core。所以对于现在的N卡架构来说，流处理器数量即CUDA Core数量。
- 从 DX9 时代过来的人应该都知道那场大变革，顶点渲染和像素渲染在 DX10 被合并成 “统一渲染” 。虽然对于开发者来说， vertex shader 和 fragment(pixel) shader 概念依旧存在，但底下执行单元却被统一了，即由流处理器（ SP ）来进行处理。

SM (Streaming Multiprocessor)

- SM 从 G80 提出的概念，中文称流多处理器，包含算术单元以及块和线程专用的其他资源（例如每个块共享内存和寄存器文件），组成了一套完整的线程运行系统。主要包括：
 1. CUDA Core : 向量运行单元 (FP32-FPU、FP64-DPU、INT32-ALU) ;
 2. Tensor Core : 张量运算单元 (FP16、BF16、INT8、INT4) ;
 3. Special Function Units : 特殊函数单元 SFU (超越函数和数学函数，e.g. 反平方根、正余弦等) ;
 4. Multi level Cache : 多级缓存 (L0/L1 Instruction Cache、L1 Data Cache & Shared Memory) ;
 5. Load/Store : 访问存储单元 LD/ST (负责数据处理) ;
 6. Warp Scheduler : 线程束调度器 ;
 7. Dispatch Unit : 分配单元 ;
 8. Register File : 寄存器堆 ;

Warp

- Warp 也称线程束。逻辑上，所有 Thread 是并行，但是，从硬件的角度来说，并不是所有的 Thread 能够在同一时刻执行，这里就需要 Warp 的引入。
- Warp 是 SM 基本执行单元，一个 Warp 包含 32 个并行 Thread (`warp_size = 32`)，这 32 个 Thread 执行于 SIMT (Single-Instruction Multiple-Thread) 模式。也就是说所有 Thread 以锁步的方式执行同一条指令，但每个 Thread 会使用各自的 Data 执行指令分支。如果在 Warp 中没有 32 个 Thread 需要工作，那么 Warp 虽然还是作为一个整体运行，但这部分 Thread 是处于非激活状态的。
- Thread 是最小的逻辑单位，Warp 是最小的硬件执行单位。

CUDA Core

- CUDA Core 在 Fermi 架构里提出，是最小的运算执行单元。
- 在 Fermi 架构中，一个 SM 中包含了有 2 组各 16 个 CUDA Core，每个 CUDA Core 包含了一个整数运算单元 ALU (Integer Arithmetic Logic Unit) 和一个浮点运算单元 FPU (Floating Point Unit)。



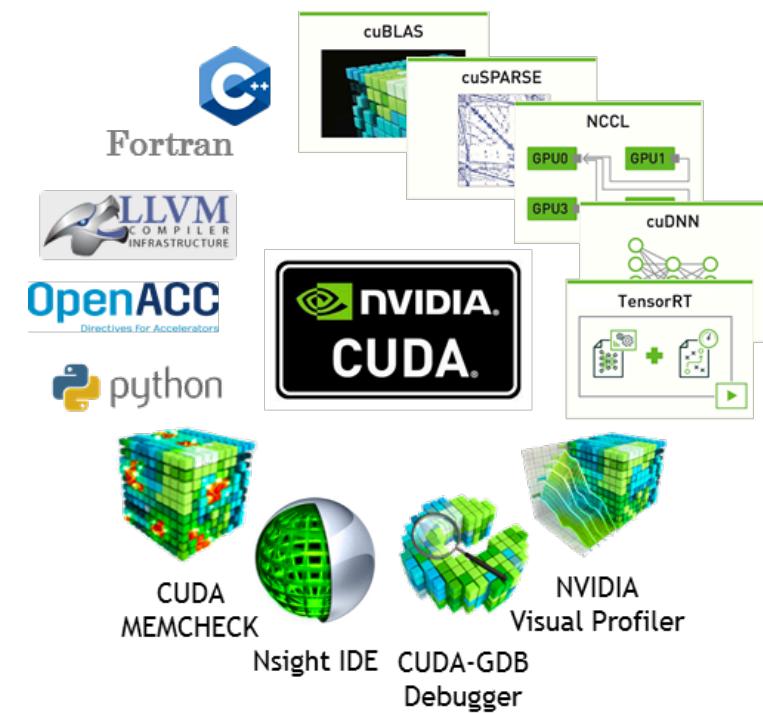
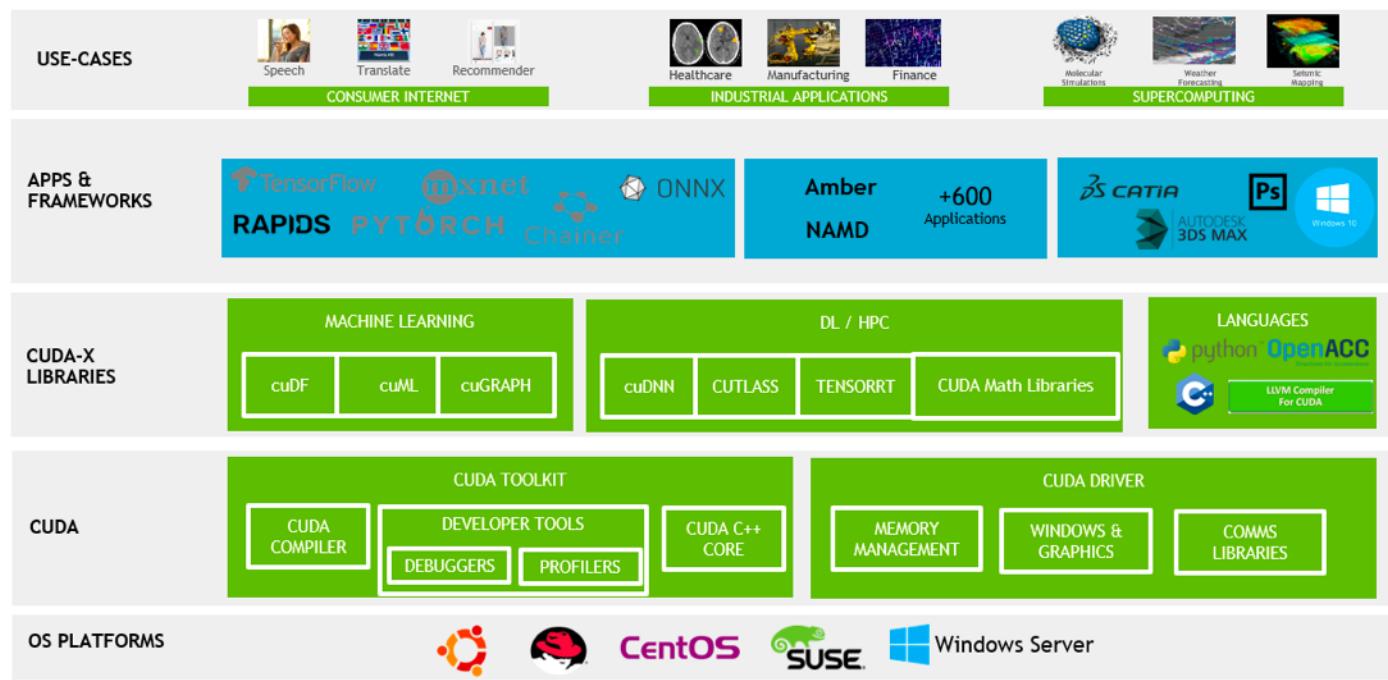
CUDA Core

- 到了 Volta 架构，CUDA Core 又和 Fermi 架构时期发生了变化。从这里开始就没有以前的 CUDA Core 了，而是变成了单独的 FP32 FPU 和 INT32 ALU。
- 因为 FP32:INT32 是 1:1，所以还是很方便把它们合并成原来的 CUDA Core 去称呼。这样做的好处是每个 SM 现在支持 FP32 和 INT32 的并发执行。



CUDA (Compute Unified Device Architecture)

- 2006年11月，NVIDIA推出了CUDA，通用并行计算平台和编程模型，用于图形处理单元(GPU)上的通用计算。利用NVIDIA GPU中的并行计算引擎(CUDA Core)以比在CPU上更有效的方式解决许多复杂的计算问题。



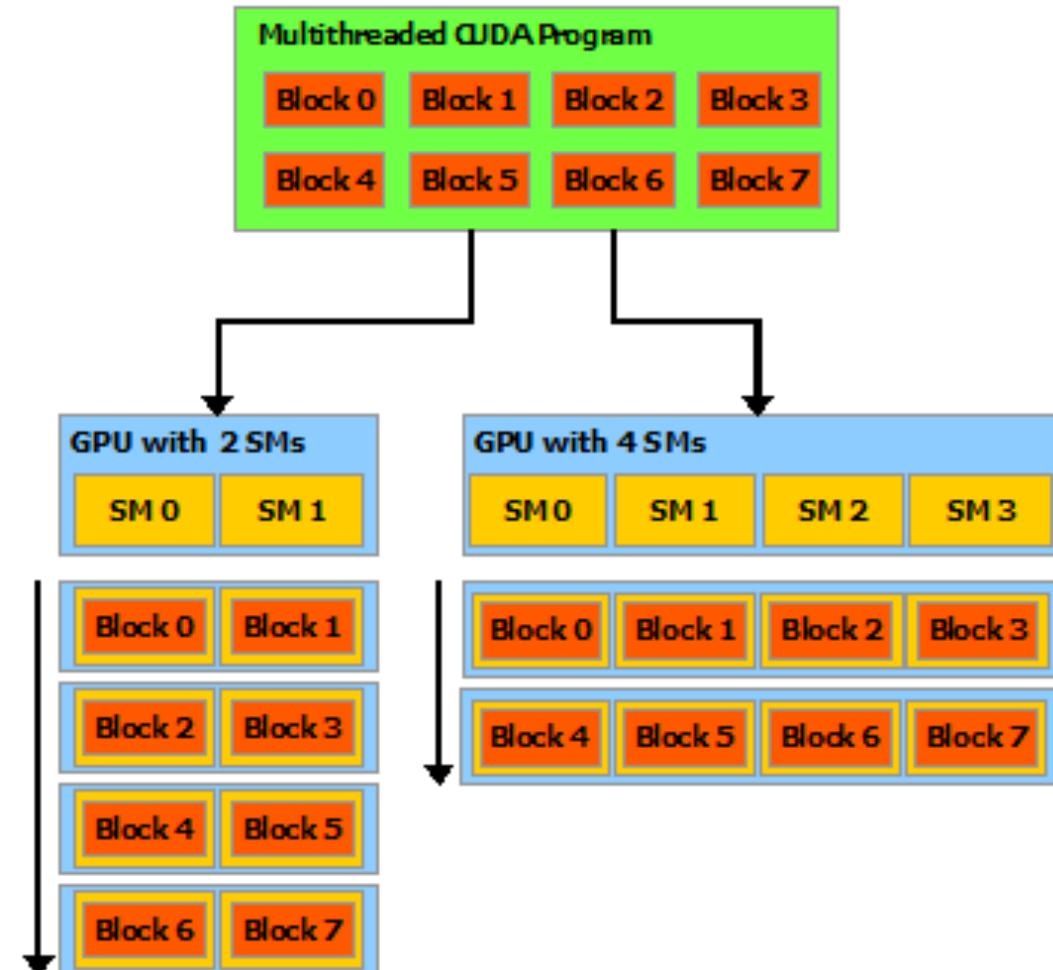
CUDA

- CUDA 并行编程模型作为一款通用接口，为熟悉 C 等标准编程语言的程序员保持较低的学习曲线。

GPU Computing Applications						
Libraries and Middleware						
cuDNN TensorRT	cuFFT, cuBLAS, cuRAND, cuSPARSE	CULA MAGMA	Thrust NPP	VSIPL, SVM, OpenCurrent	PhysX, OptiX, iRay	MATLAB Mathematica
Programming Languages						
C	C++	Fortran	Java, Python, Wrappers		DirectCompute	Directives (e.g., OpenACC)
CUDA-enabled NVIDIA GPUs						
Turing Architecture (Compute capabilities 7.x)	DRIVE/JETSON AGX Xavier	GeForce 2000 Series		Quadro RTX Series	Tesla T Series	
Volta Architecture (Compute capabilities 7.x)	DRIVE/JETSON AGX Xavier					Tesla V Series
Pascal Architecture (Compute capabilities 6.x)	Tegra X2	GeForce 1000 Series		Quadro P Series	Tesla P Series	
Maxwell Architecture (Compute capabilities 5.x)	Tegra X1	GeForce 900 Series		Quadro M Series	Tesla M Series	
Kepler Architecture (Compute capabilities 3.x)	Tegra K1	GeForce 700 Series GeForce 600 Series		Quadro K Series	Tesla K Series	
	EMBEDDED	CONSUMER DESKTOP, LAPTOP		PROFESSIONAL WORKSTATION	DATA CENTER	

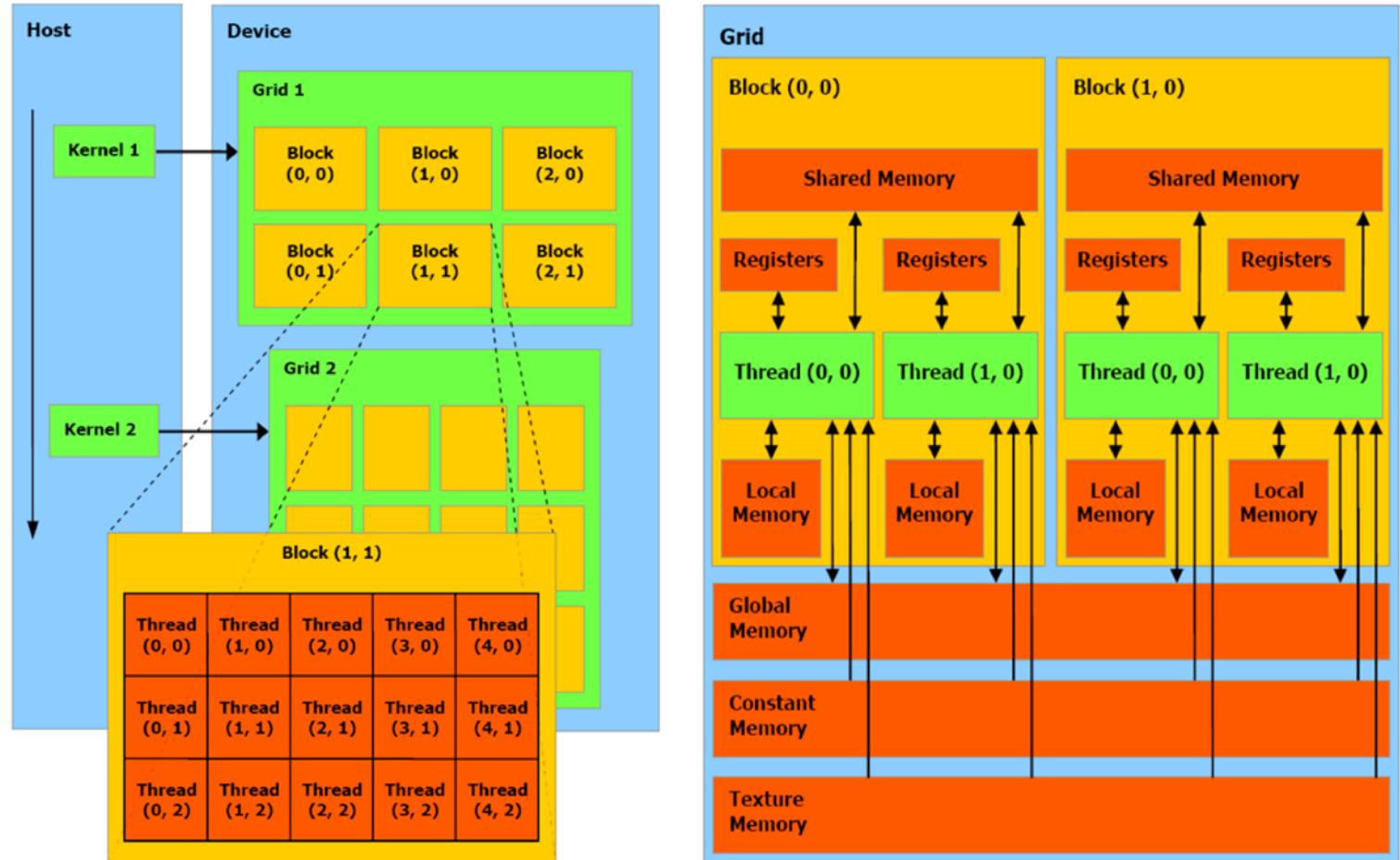
CUDA

- CUDA 将 GPU 架构建模为多核系统，将 GPU 并行线程抽象为**层次化线程结构**（网格状的线程块）。
- 这样可以指导开发者将问题划分为**独立线程块并行解决的子问题**，并进一步将每个子问题划分为可以由块内的**线程并行协作解决**的部分。



CUDA

- 通过允许线程 Thread 在解决每个子问题时，进行协作来保持语言表达能力，同时实现自动可扩展性。实际上，每个线程块 Thread Block 都可以在 GPU 的计算单元内以任何顺序进行调度，以便编译后的 CUDA 程序可以在任意数量的多处理器上执行。



峰值算力

- GPU 算力跟计算核心个数、核心频率、核心单时钟周期能力三个因素有关。GPU 峰值计算能力，公式如下：

$$\text{Peak FLOPS} = F_{clk} * N_{SM} * F_{req}$$

- 其中 F_{clk} 为 GPU 的时钟周期内指令执行数 (FLOPS/Cycle) , F_{req} 为运行频率 (GHz) , N_{SM} 为 GPU SM 数量 (Cores)。

峰值算力

- A100 中 FP32 Tensor Core 指令吞吐 64 FLOPS/Cycle，核心运行频率为 1.41GHz，SM 数量为 108，那么有： $Peak\ FLOPS = 1.41 * 108 * 64 * 2 = 19,491\ GFLOPS \sim 1.95\ TFLOPS$

Table 1. NVIDIA A100 Tensor Core GPU Performance Specs

Peak FP64 ¹	9.7 TFLOPS
Peak FP64 Tensor Core ¹	19.5 TFLOPS
Peak FP32 ¹	19.5 TFLOPS
Peak FP16 ¹	78 TFLOPS
Peak BF16 ¹	39 TFLOPS
Peak TF32 Tensor Core ¹	156 TFLOPS 312 TFLOPS ²
Peak FP16 Tensor Core ¹	312 TFLOPS 624 TFLOPS ²
Peak BF16 Tensor Core ¹	312 TFLOPS 624 TFLOPS ²
Peak INT8 Tensor Core ¹	624 TOPS 1,248 TOPS ²
Peak INT4 Tensor Core ¹	1,248 TOPS 2,496 TOPS ²

1 - Peak rates are based on GPU Boost Clock.

2 - Effective TFLOPS / TOPS using the new Sparsity feature

GPU概念之间的关系

- 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
- HBM —— 高带宽存储器，High Bandwidth Memory
- 包含关系为：GPC > TPC > SM > CORE
- SM 中包含 Poly Morph Engine、LI Cache、Shared Memory、CUDA Core等
- CUDA Core 中包含 ALU、FPU、Execution Context、Thread Detach、Command Decode等

GPU概念之间的关系



Reference 引用&参考

1. <https://zhuanlan.zhihu.com/p/620257581> 大佬们，A100显卡上的tensorcore有自己的私有寄存器
2. https://old.hotchips.org/wp-content/uploads/hc_archives/hc29/HC29.21-Monday-Pub/HC29.21.10-GPU-Gaming-Pub/HC29.21.132-Volta-Choquette-NVIDIA-Final3.pdf
3. <https://ieeexplore.ieee.org/author/37086370271>
4. <https://www.computer.org/csdl/proceedings-article/hcs/2019/08875651/1ehCtCN4SUE>
5. <https://resources.nvidia.com/en-us-genomics-ep/ampere-architecture-white-paper>
6. <https://developer.nvidia.com/blog/nvidia-hopper-architecture-in-depth/>
7. <https://zhuanlan.zhihu.com/p/231302709>
8. <https://www.zhihu.com/zvideo/1367794004235542528>
9. <https://www.bilibili.com/read/cv15145865?from=search>
10. <https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html?highlight=matrix%20multiply>
11. <https://nyu-cds.github.io/python-gpu/02-cuda/>
12. https://zhuanlan.zhihu.com/p/258196004?utm_id=0
13. https://blog.csdn.net/qq_42059060/article/details/121274184



BUILDING A BETTER CONNECTED WORLD

THANK YOU

Copyright©2014 Huawei Technologies Co., Ltd. All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.