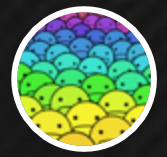
AI 芯片 – GPU 详解

Fermi - Volta 架构



ZOMI

Talk Overview

1. 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

1. **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 Code 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

1. **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 Pascal 架构
   - Volta 到 Hopper 架构
   - Tensor Code 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

1. 从 Fermi 到 Pascal 架构

- Over will – 总体概览
- Fermi 费米架构
- Kepler 开普勒架构
- Maxwell 麦克斯韦架构
- Pascal 帕斯卡架构

NVIDIA GPU架构发展

架构名称	Fermi	Kepler	Maxwell	Pascal	Volta	Turing	Ampere	Hopper
中文名字	费米	开普勒	麦克斯韦	帕斯卡	伏特	图灵	安培	赫柏
发布时间	2010	2012	2014	2016	2017	2018	2020	2022
核心参数	16个SM，每个SM包含32个CUDA Cores，一共512 CUDA Cores	15个SMX，每个SMX包括192个FP32+64个FP64 CUDA Cores	16个SM，每个SM包括4个处理块，每个处理块包括32个CUDA Cores +8个LD/ST Unit + 8 SFU	GP100有60个SM，每个SM包括64个CUDA Cores，32个DP Cores	80个SM，每个SM包括32个FP64 +64 Int32+64 FP32+8个Tensor Cores	102核心92个SM，SM重新设计，每个SM包含64个Int32+64个FP32+8个Tensor Cores	108个SM，每个SM包含64个FP32+64个INT32+32个FP64+4个Tensor Cores	132个SM，每个SM包含128个FP32+64个INT32+64个FP64+4个Tensor Cores
特点&优势	首个完整GPU计算架构，支持与共享存储结合的Cache层次GPU架构，支持ECC GPU架构	游戏性能大幅提升，首次支持GPU Direct技术	每组SM单元从192个减少到每组128个，每个SMM单元拥有更多逻辑控制电路	NVLink第一代，双向互联带宽160 GB/s，P100拥有56个SM HBM	NVLink2.0，Tensor Cores第一代，支持AI运算	Tensor Core2.0，RT Core第一代	Tensor Core3.0，RT Core2.0，NVLink3.0，结构稀疏性矩阵MIG1.0	Tensor Core4.0，NVlink4.0，结构稀疏性矩阵MIG2.0
纳米制程	40/28nm 30亿晶体管	28nm 71亿晶体管	28nm 80亿晶体管	16nm 153亿晶体管	12nm 211亿晶体管	12nm 186亿晶体管	7nm 283亿晶体管	4nm 800亿晶体管
代表型号	Quadro 7000	K80 K40M	M5000 M4000 GTX 9XX系列	P100 P6000 TTX1080	V100 TiTan V	T4，2080TI RTX 5000	A100 A30系列	H100

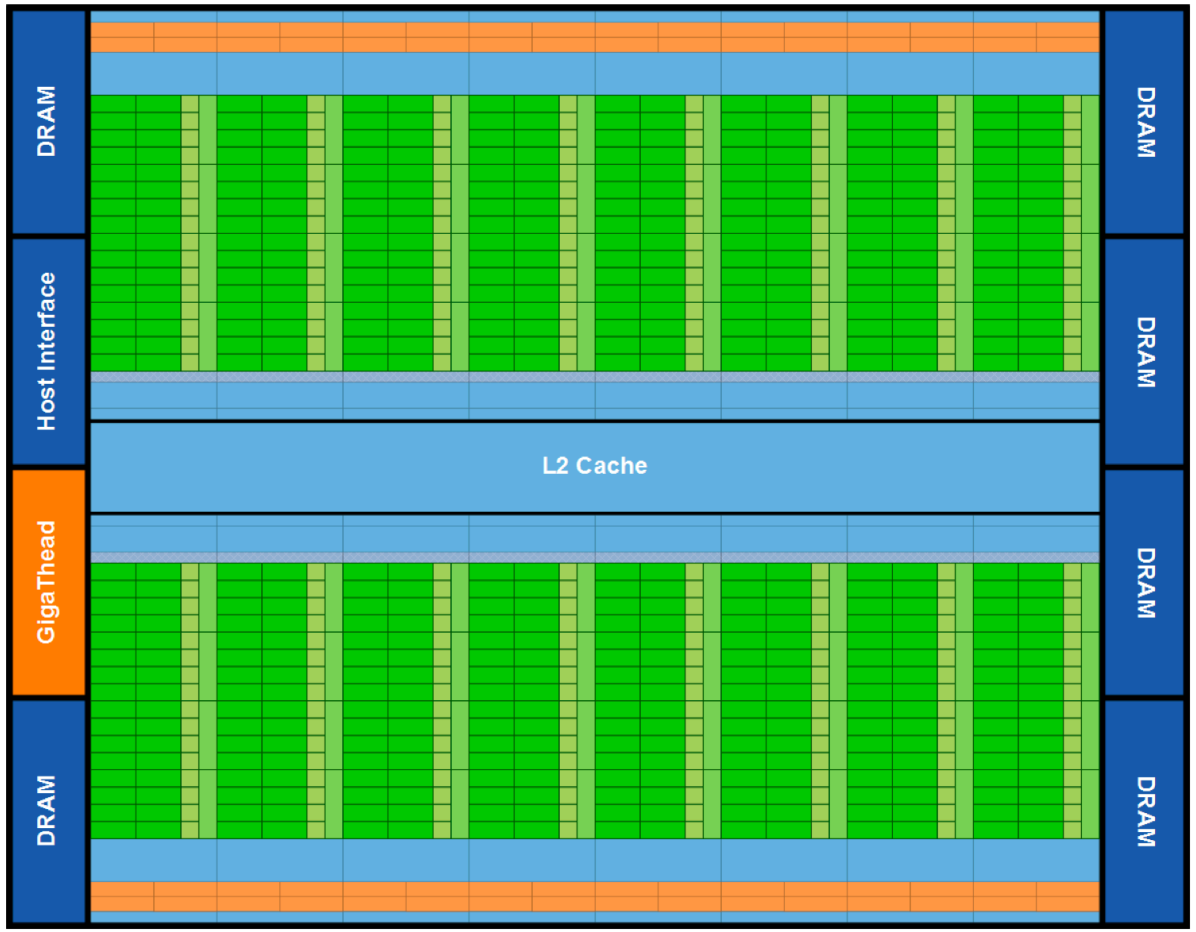
NVIDIA GPU架构发展

架构名称	Fermi	Kepler	Maxwell	Pascal	Volta
中文名字	费米	开普勒	麦克斯韦	帕斯卡	伏特
发布时间	2010	2012	2014	2016	2017
核心参数	16个SM，每个SM包含32个CUDA Cores，一共512 CUDA Cores	15个SMX，每个SMX包括192个FP32+64个FP64 CUDA Cores	16个SM，每个SM包括4个处理块，每个处理块包括32个CUDA Cores+8个LD/ST Unit + 8 SFU	GP100有60个SM，每个SM包括64个CUDA Cores，32个DP Cores	80个SM，每个SM包括32个FP64+64 Int32+64 FP32+8个Tensor Cores
特点&优势	首个完整GPU计算架构，支持与共享存储结合的Cache层次GPU架构，支持ECC GPU架构	游戏性能大幅提升，首次支持GPU Direct技术	每组SM单元从192个减少到每组128个，每个SMM单元拥有更多逻辑控制电路	NVLink第一代，双向互联带宽160GB/s，P100拥有56个SM HBM	NVLink2.0，Tensor Cores第一代，支持AI运算
纳米制程	40/28nm 30亿晶体管	28nm 71亿晶体管	28nm 80亿晶体管	16nm 153亿晶体管	12nm 211亿晶体管
代表型号	Quadro 7000	K80 K40M	M5000 M4000 GTX 9XX系列	P100 P6000 TTX1080	V100 TiTan V

Fermi 架构

Fermi 费米架构

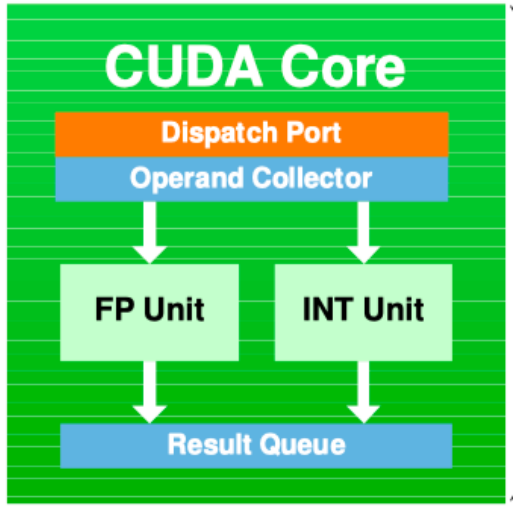
- Fermi 架构最大可支持 16 个 SMs，每个 SM 有 32 个 CUDA Cores，一共 512个 CUDA Cores。
- 整个 GPU 有多个 GPC（图形处理集），单个 GPC 包含一个光栅引擎（Raster Engine），4个 SM。

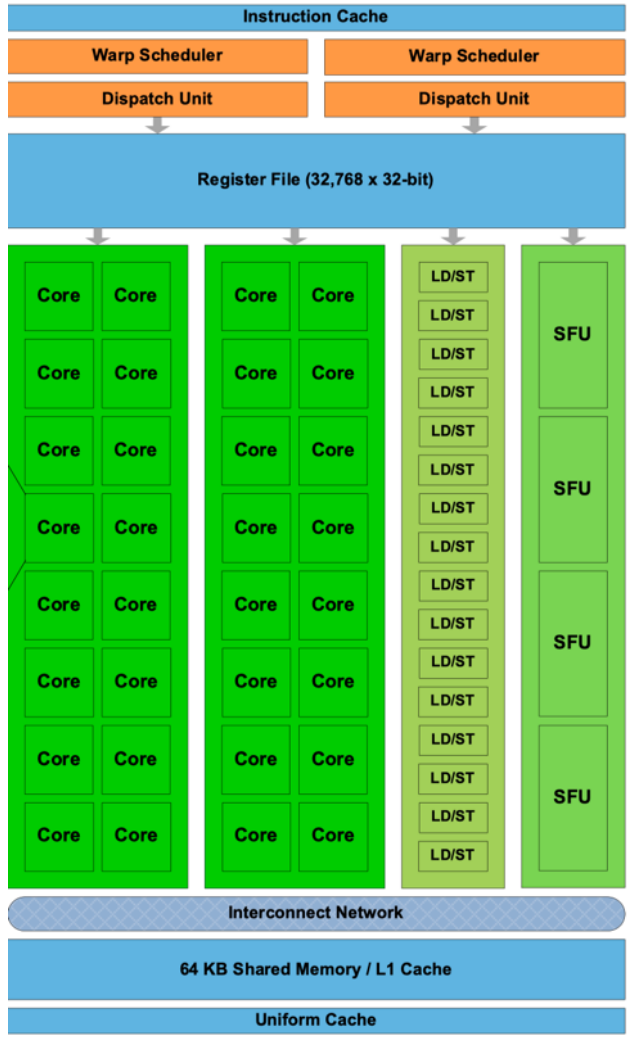


Fermi's 16 SM are positioned around a common L2 cache. Each SM is a vertical rectangular strip that contain an orange portion (scheduler and dispatch), a green portion (execution units), and light blue portions (register file and L1 cache).

Fermi 费米架构

- Fermi 架构最大可支持 16 个 SMs，每个 SM 有 32 个 CUDA Cores，一共 512个 CUDA Cores。
- 整个 GPU 有多个 GPC（图形处理集），单个 GPC 包含一个光栅引擎（Raster Engine），4个 SM。

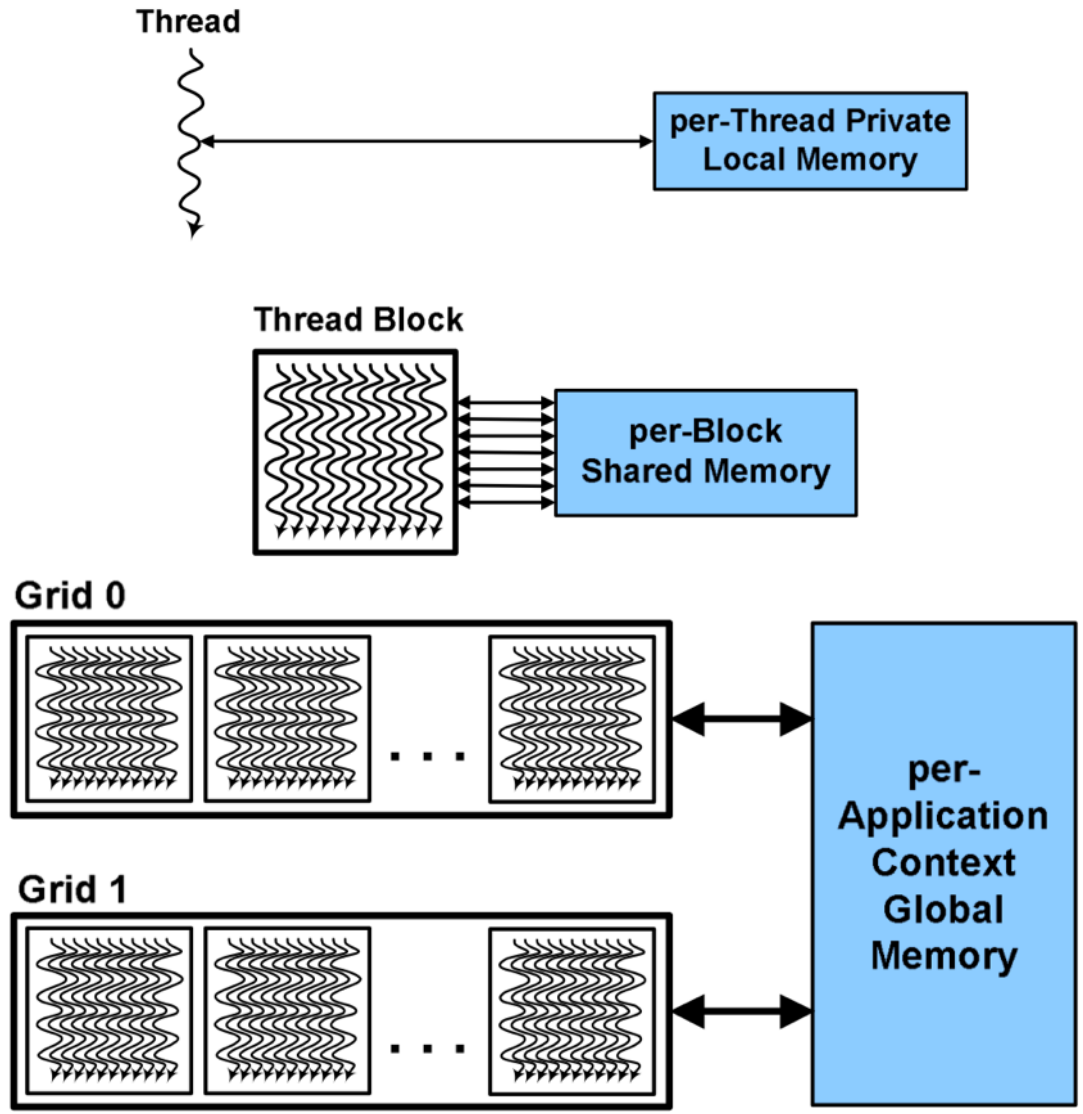




Fermi Streaming Multiprocessor (SM)

Fermi 费米架构

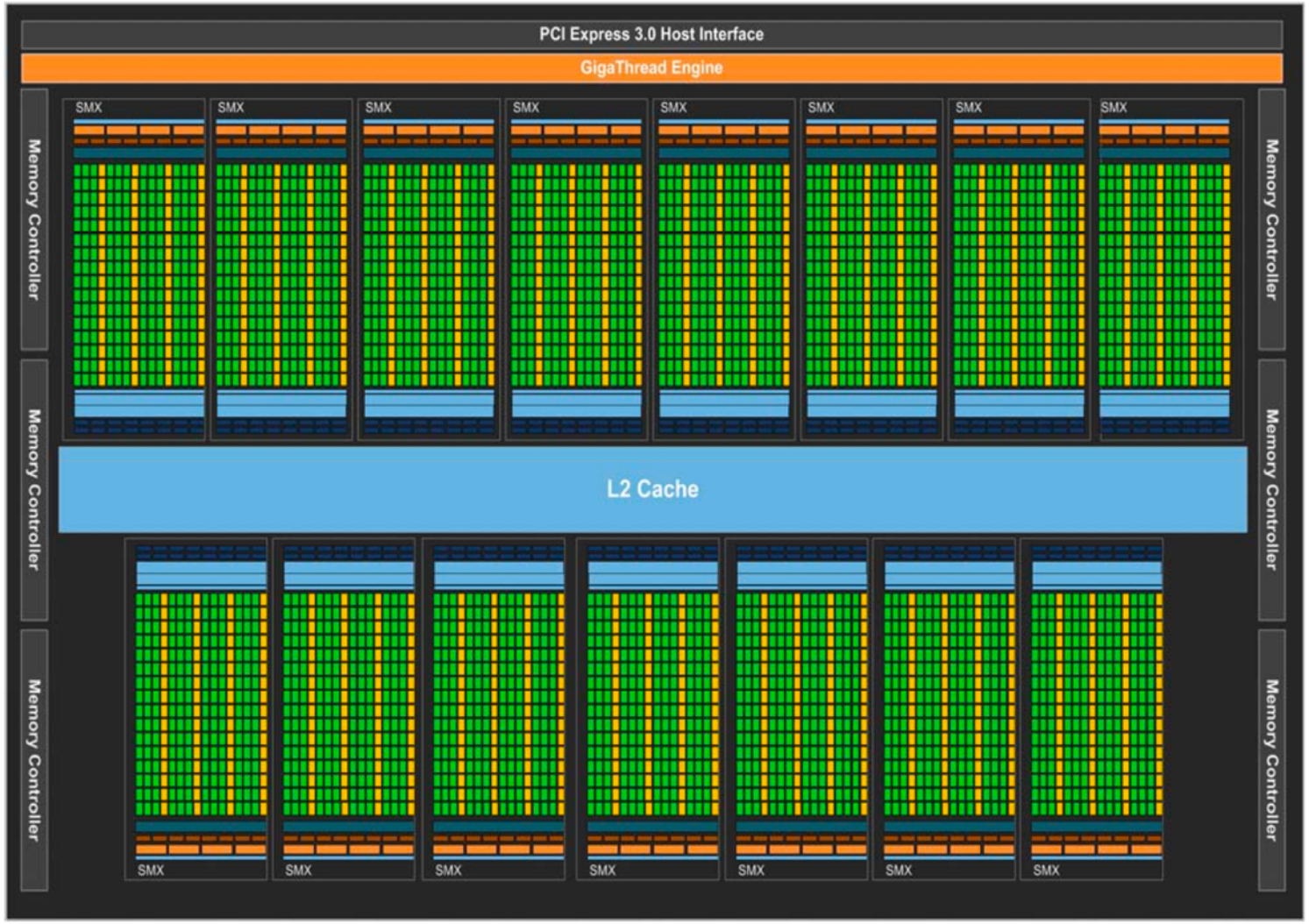
- Fermi 架构最大可支持 16 个 SMs，每个 SM 有 32 个 CUDA Cores，一共 512个 CUDA Cores。
- CUDA 线程 threads 、块 blocks 和网格 grids 的层次结构，具有相应的每个线程专用、每个块共享和每个应用程序全局内存空间。



Kepler 架构

Kepler 开普勒架构

1. SM改名成了SMX，但是所代表的概念没有大变化；
2. Kepler架构在硬件上直接有双精运算单元的架构；
3. 提出 GPU Direct 技术，可以绕过 CPU/System Memory，完成与本机其他 GPU 或者其他机器 GPU 的直接数据交换。



Kepler 开普勒架构

Fermi 架构

SM（32核）

Kepler 架构

SMX（192核）

Maxwell 架构

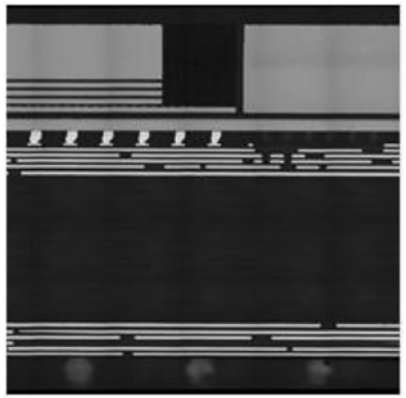
Maxwell 麦克斯韦架构

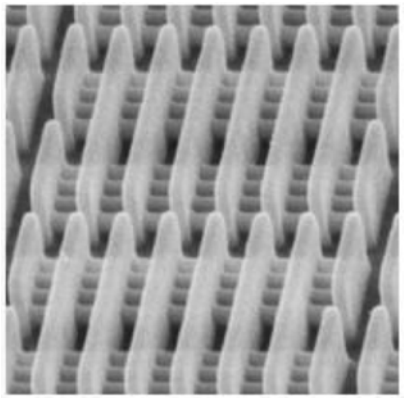
Pascal 架构

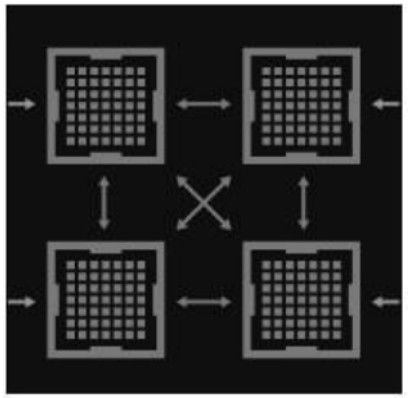
Pascal 帕斯卡架构

1. 深度学习：第一款面向AI 的架构，提供定制化提供 CuDNN 等库；
2. 互联带宽：第一代 NVLink（P100），单机卡间通信带宽提升，多机间 InfiniBand 扩展带宽；
3. 系统内存：内存 GDDR5 换成 HBM2，Global Memory 带宽提升一个数量级；
4. 制造工艺：16nm FinFET 工艺，相同功耗下提升提升一个数量级；
5. 计算核心：CUDA Core 硬件支持 FP16 半精计算；

Pascal Architecture


CoWoS with HBM2

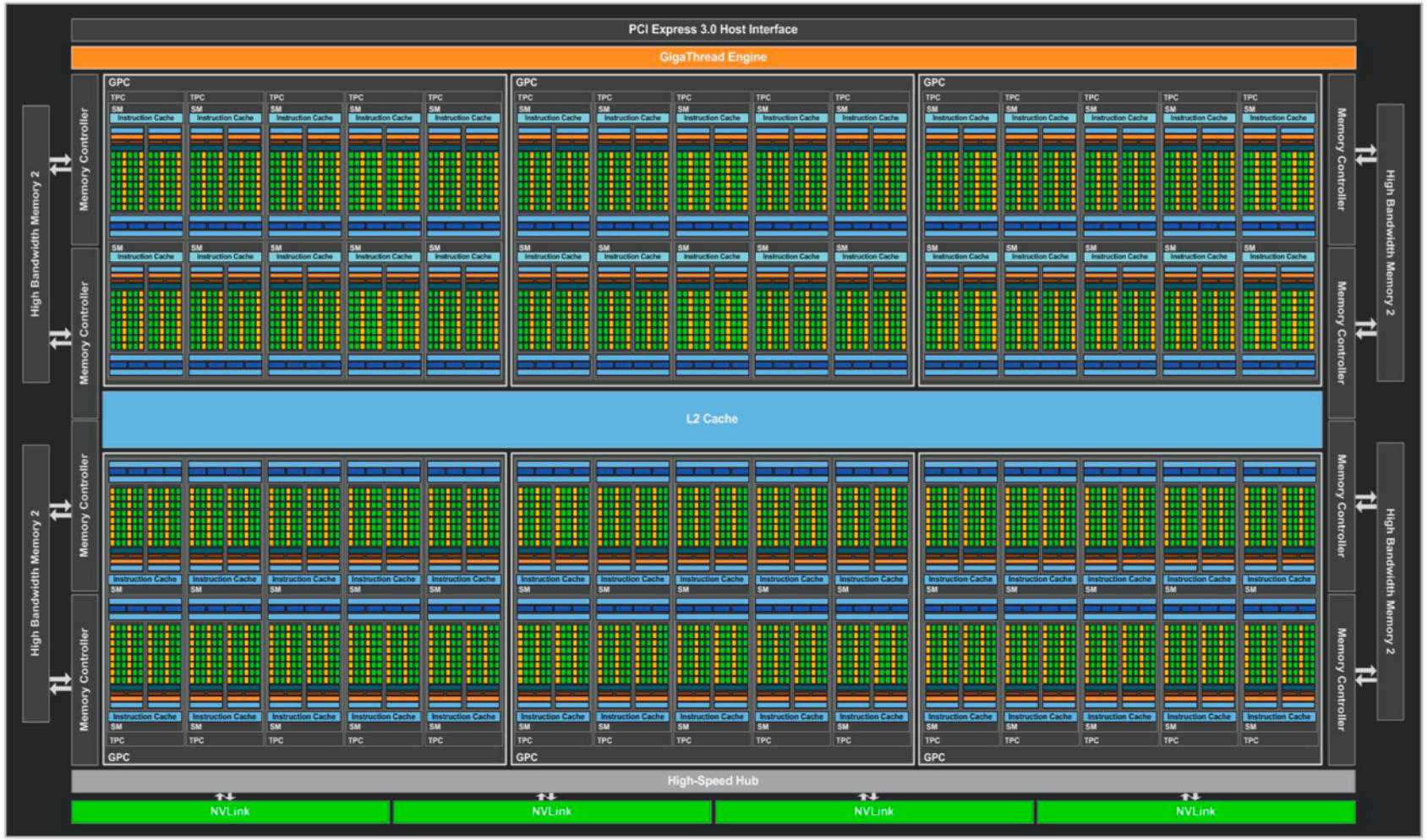

16nm FinFET


NVLink


Unified Memory
Compute Preemption
New AI Algorithms

Pascal 帕斯卡架构

- SM 内部作了进一步精简，整体思路是 SM 内部包含的东西越来越少，但是总体片上 SM 数量每一代都在不断增加。

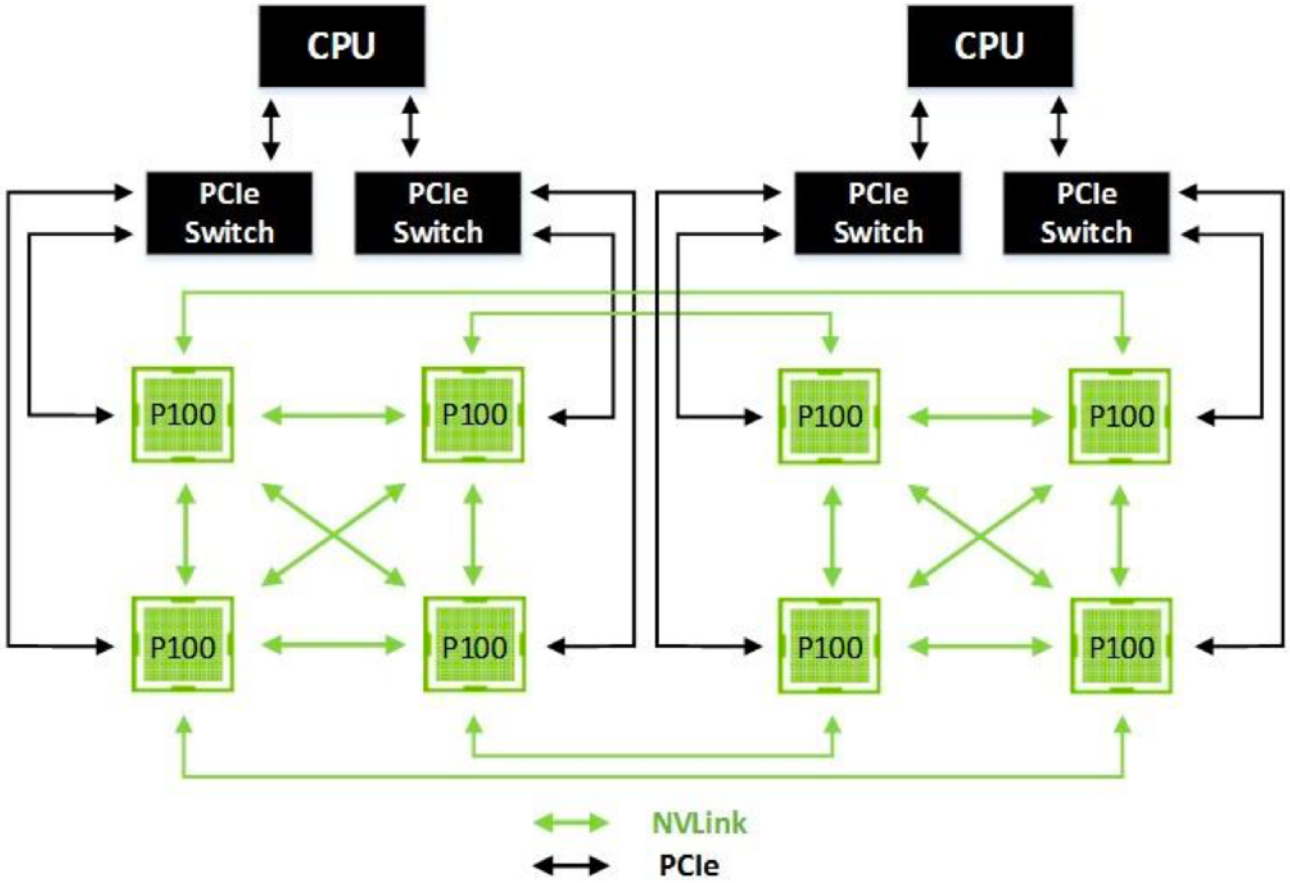


Pascal 帕斯卡架构

1. 单个SM只有64个FP32 CUDA Cores，相比Maxwell的128和Kepler的192，这个数量要少很多，并且64个CUDA Cores分为了两个区块；
2. Register File 保持相同大小，每个线程可以使用更多寄存器，单个SM也可以并发更多 thread/warp/block；
3. 增加 32个FP64 CUDA Cores (DP Unit)，FP32 CUDA Core 具备处理FP16的能力。

Pascal 帕斯卡架构

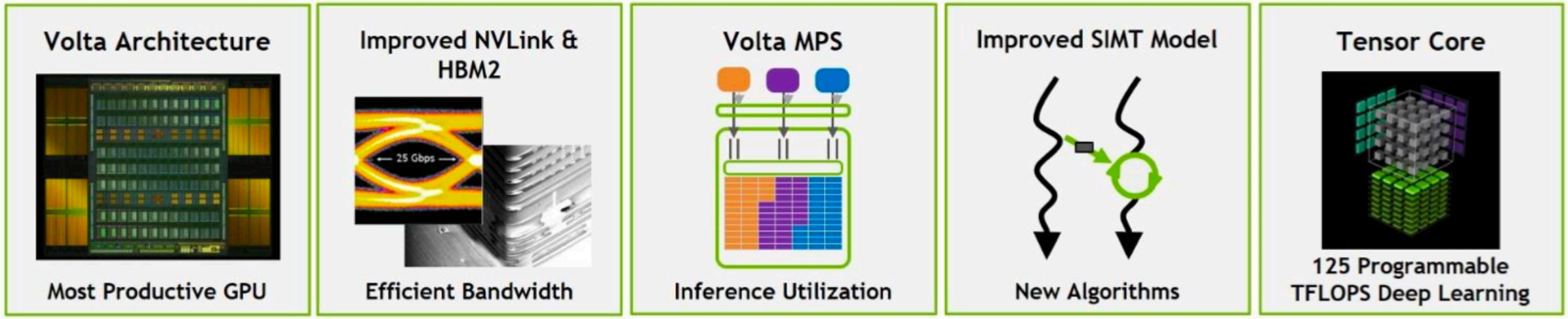
- 多机之间，采用InfiniBand和100 Gb Ethernet通信，单机内单GPU到单机8 GPU，PCIe 带宽成为瓶颈。
- NVIDIA 提供 NVLink 用以单机内多 GPU 内的点到点通信，带宽达到 160GB/s, ~5 倍 PCIe 3 x 16。



Volta 架构

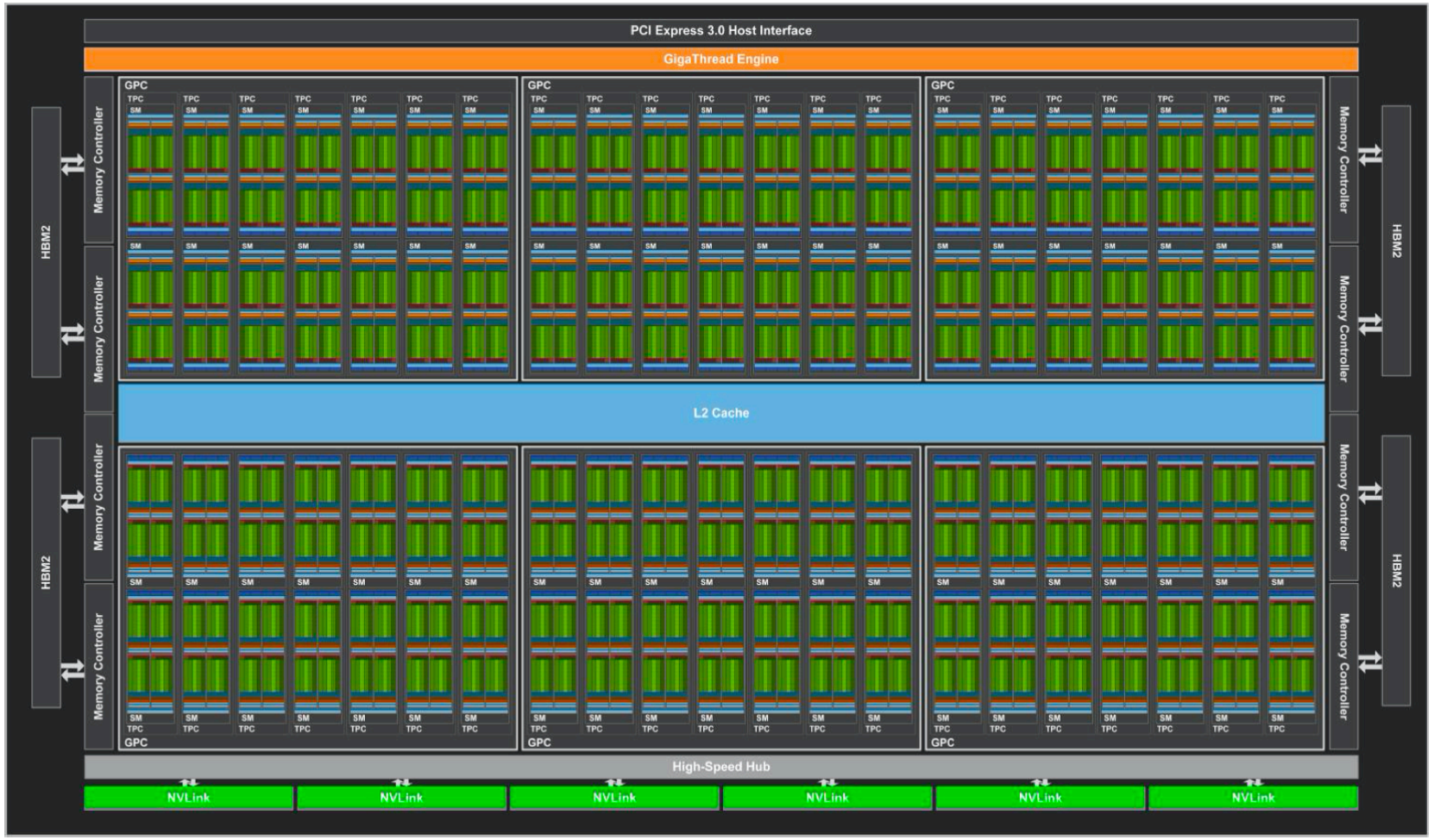
Volta 伏特架构

- **CUDA Core拆分**：分离 FPU 和 ALU，取消 CUDA Core，一条指令可以同时执行不同计算；
- **独立线程调度**：改进SIMT模型架构，使得每个线程都有独立的PC(Program Counter) 和 Stack;
- **Tensor Core**：针对深度学习提供张量计算核心，专门针对卷积计算进行加速；
- **GRF & Cache**：Global memory 访问也能享受 highly banked cache 加速；



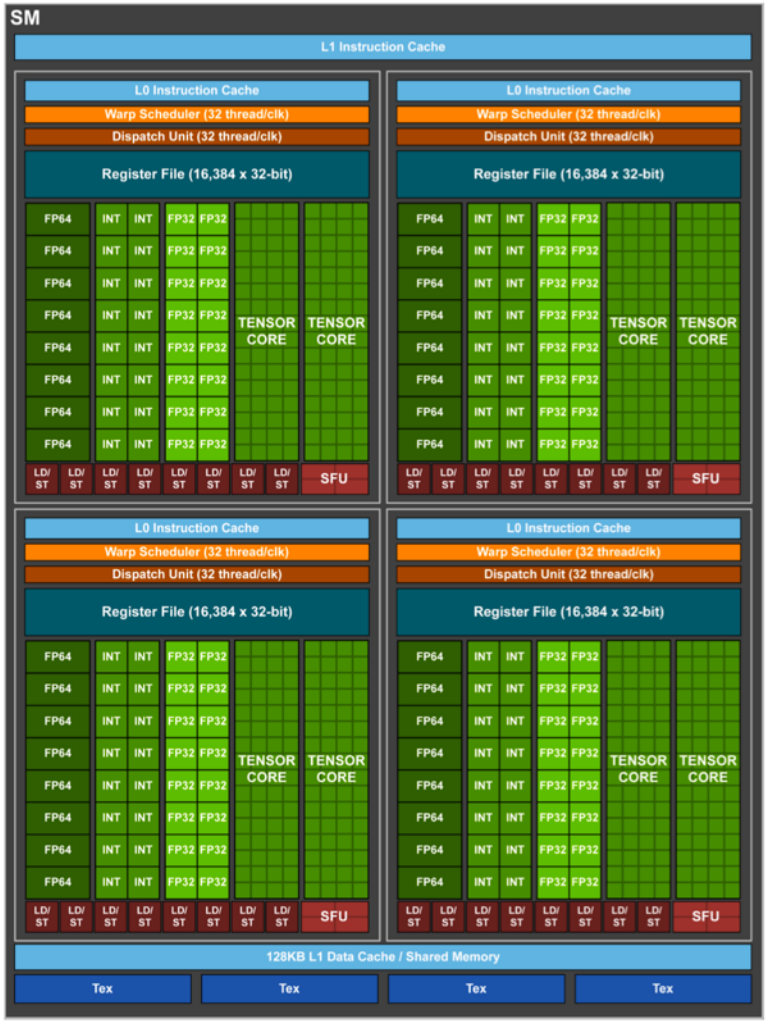
Volta 伏特架构

- SM 中包含：
  1. 4 个 Warp Scheduler
  2. 4 个 Dispatch Unit
  3. 64 个 FP32 Core
  4. 64 个 INT32 Core
  5. 32 个 FP64 Core
  6. 8 个 Tensor Core
  7. 32 个 LD/ST Unit
  8. 4 个 SFU



Volta 伏特架构

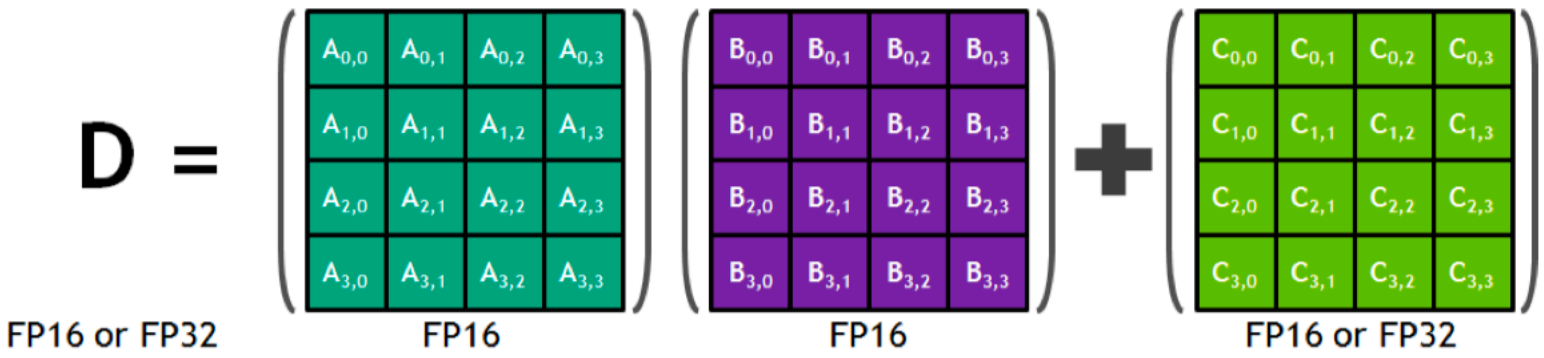
- SM 中包含：
  1. 4 个 Warp Scheduler，4 个 Dispatch Unit
  2. 64 个 FP32 Core（4 * 16）
  3. 64 个 INT32 Core（4 * 16）
  4. 32 个 FP64 Core（4 * 8）
  5. 8 个 Tensor Core （4 * 2）
  6. 32 个 LD/ST Unit（4 * 8）
  7. 4 个 SFU
- FP32 和 INT32 两组运算单元独立出现在流水线中，每个 Cycle 都可以同时执行 FP32 和 INT32 指令。



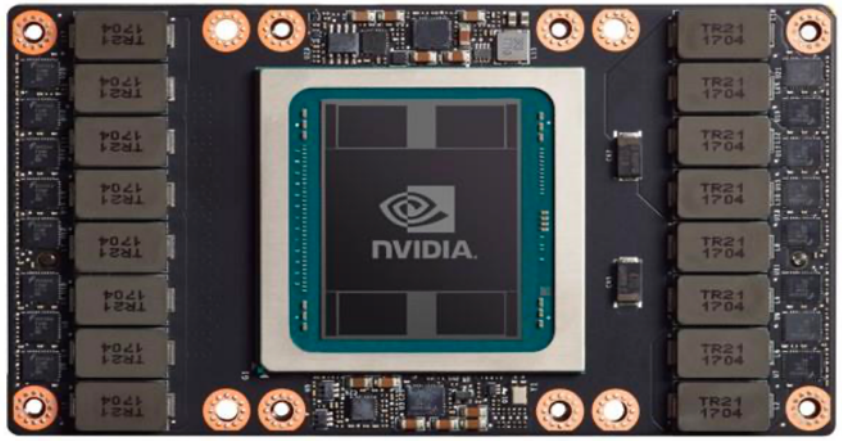
Volta 伏特架构

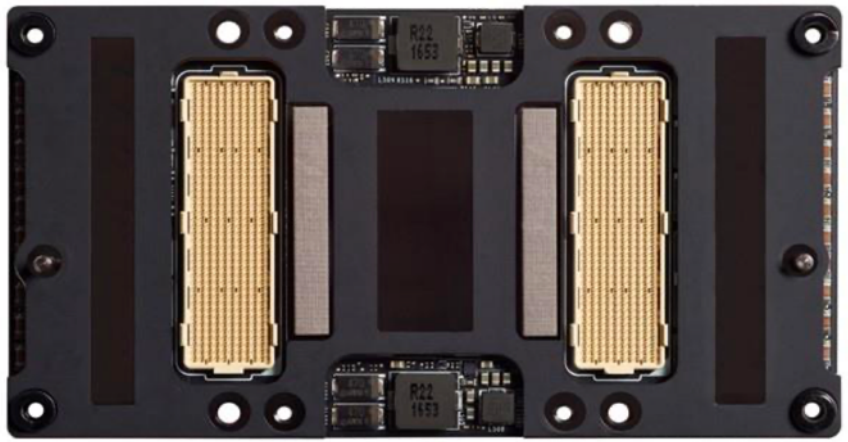
- GPU 并行模式实现深度学习功能过于通用，最常见 Conv/GEMM 操作，依旧要被编码成 FMA，硬件层面还是需要把数据按：~~寄存器-ALU-寄存器-ALU-寄存器~~，方式来回搬运。

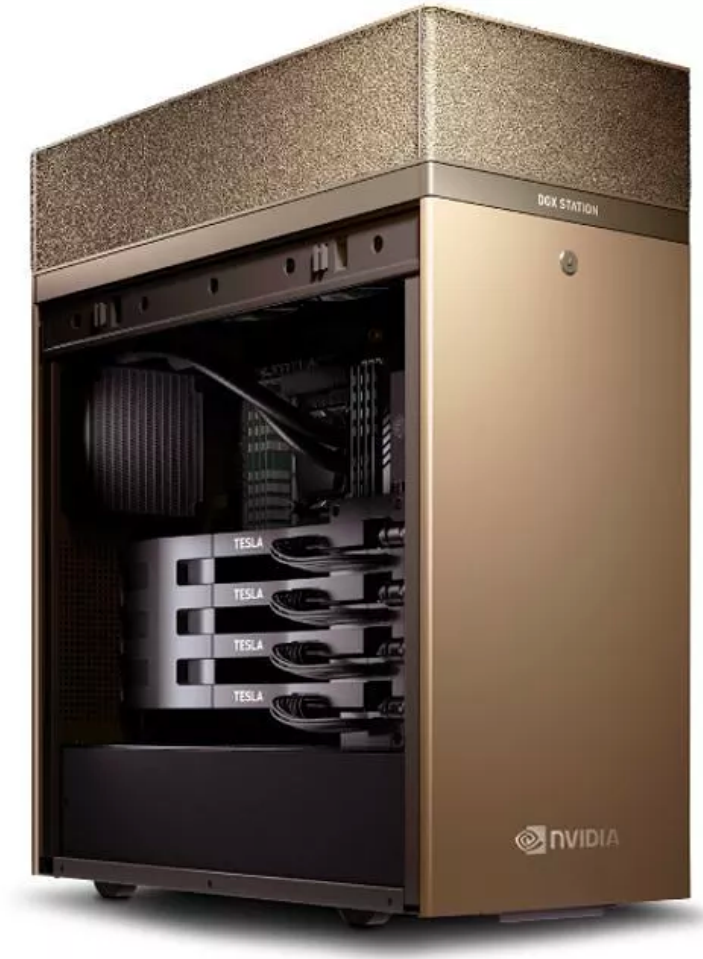
- 每个 Tensor Core 每周期能执行 4x4x4 GEMM，即 64 个 FMA。虽然只支持 FP16 数据，但输出可以是 FP32，相当于 64 个 FP32 ALU 提供算力，能耗上还有优势。



Volta 伏特架构

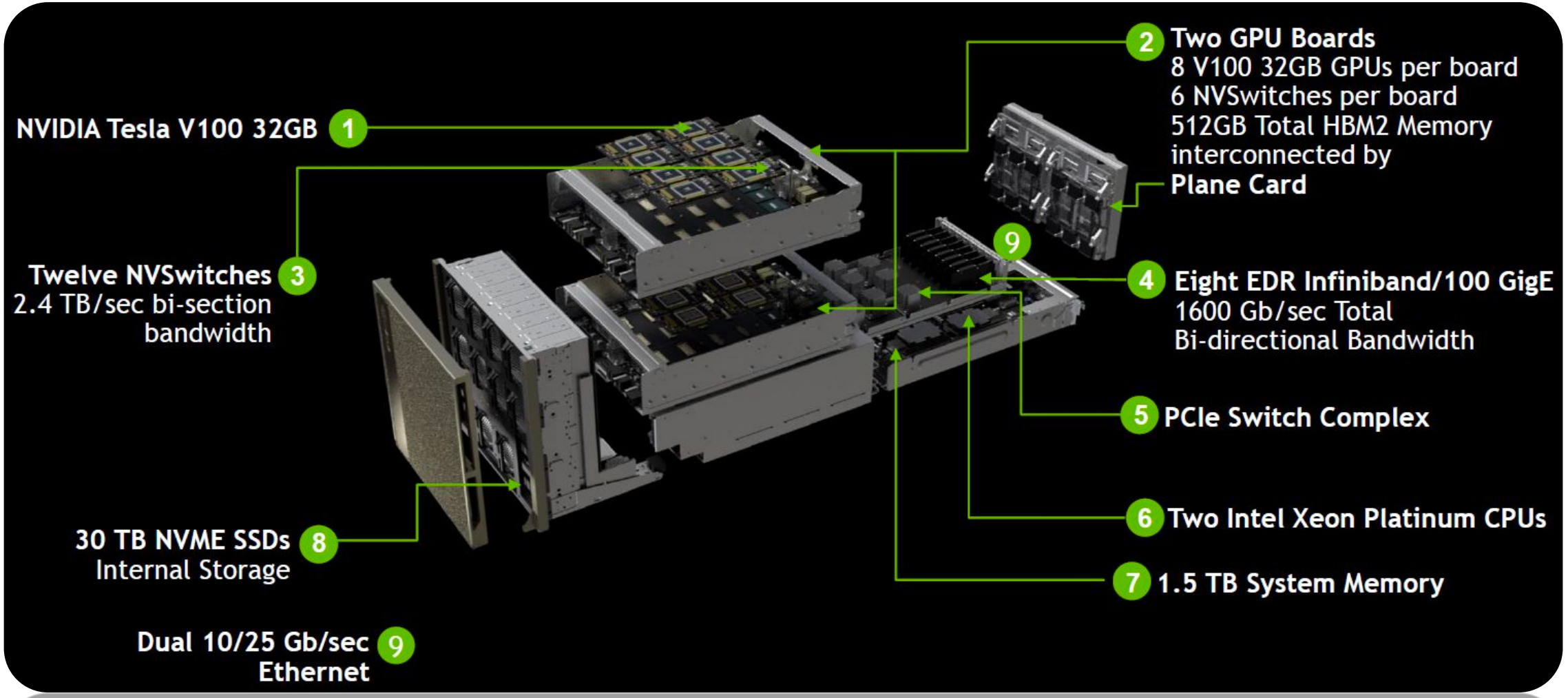






Tesla V100 Powered DGX Station

Volta 伏特架构



Reference 引用&参考

1. https://zhuanlan.zhihu.com/p/413145211 英伟达GPU架构演进近十年，从费米到安培
2. https://blog.csdn.net/daijingxin/article/details/115042353 NVIDA GPU架构演进
3. https://zhuanlan.zhihu.com/p/258196004?utm_id=0 NVIDIA GPU的一些解析（一）
4. https://www.bilibili.com/video/BV1cB4y1Q75r 技术分享：英伟达GPU架构演进(2010-2022)
5. https://www.nvidia.com/en-us/data-center/resources/pascal-architecture-whitepaper/
6. https://images.nvidia.com/aem-dam/en-zz/Solutions/design-visualization/technologies/turing-architecture/NVIDIA-Turing-Architecture-Whitepaper.pdf
7. https://resources.nvidia.com/en-us-tensor-core
8. https://www.hpctech.co.jp/catalog/gtc22-whitepaper-hopper_v1.01.pdf
9. https://www.microway.com/download/whitepaper/NVIDIA_Maxwell_GM204_Architecture_Whitepaper.pdf
10. https://developer.nvidia.com/maxwell-compute-architecture
11. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/tesla-product-literature/NVIDIA-Kepler-GK110-GK210-Architecture-Whitepaper.pdf
12. https://github.com/g-truc/sdk/blob/master/documentation/hardware/nvidia/2012%20-%20Kepler/NVIDIA-Kepler-GK110-Architecture-Whitepaper.pdf
13. https://www.dell.com/learn/aw/en/awbsdt1/shared-content~data-sheets~en/documents~nvidia-fermi-compute-architecture-whitepaper-en.pdf