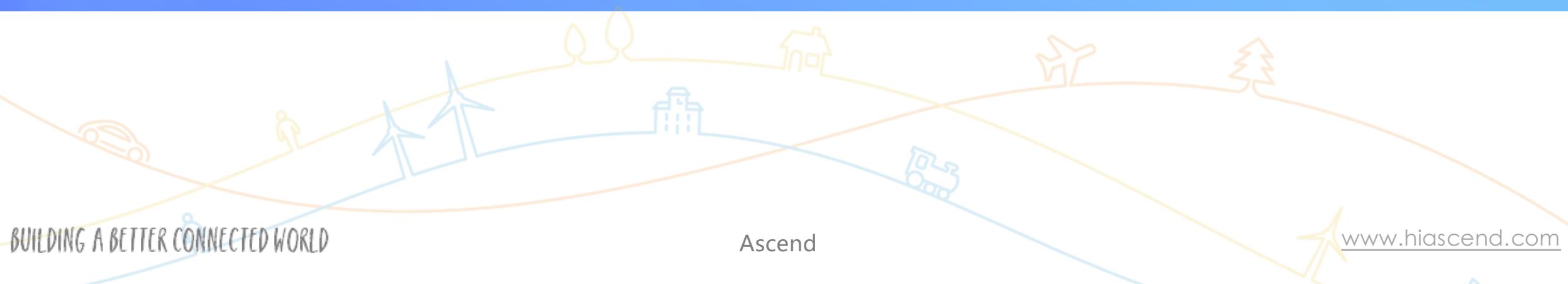


# AI 芯片 - GPU 详解

# Volta-Hopper 架构



ZOMI



BUILDING A BETTER CONNECTED WORLD

Ascend

[www.hiascend.com](http://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 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

## 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 Pascal 架构
- Volta 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

## 1. 从 Fermi 到 Pascal 架构

- Over will – 总体概览
- Volta 伏特架构
- Turing 图灵架构
- Ampere 安培架构
- Hopper 帕斯卡架构

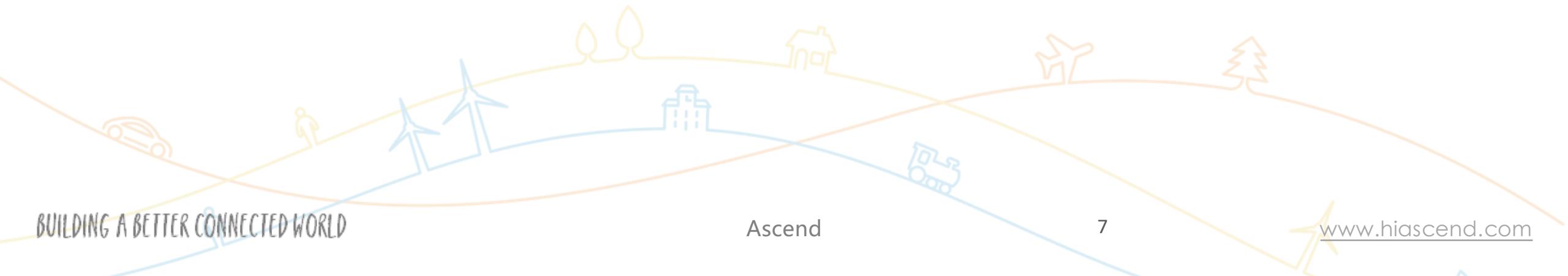
# NVIDIA GPU架构发展

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，NV Link3.0，结构稀疏性矩阵MIG2.0	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	Volta	Turing	Ampere	Hopper
中文名字	伏特	图灵	安培	赫柏
发布时间	2017	2018	2020	2022
核心参数	80个SM，每个SM包括32个FP64+64个INT32+64个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	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	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	V100 TiTan V	T4，2080TI RTX 5000	A100 A30系列	H100

# Volta 架构



BUILDING A BETTER CONNECTED WORLD

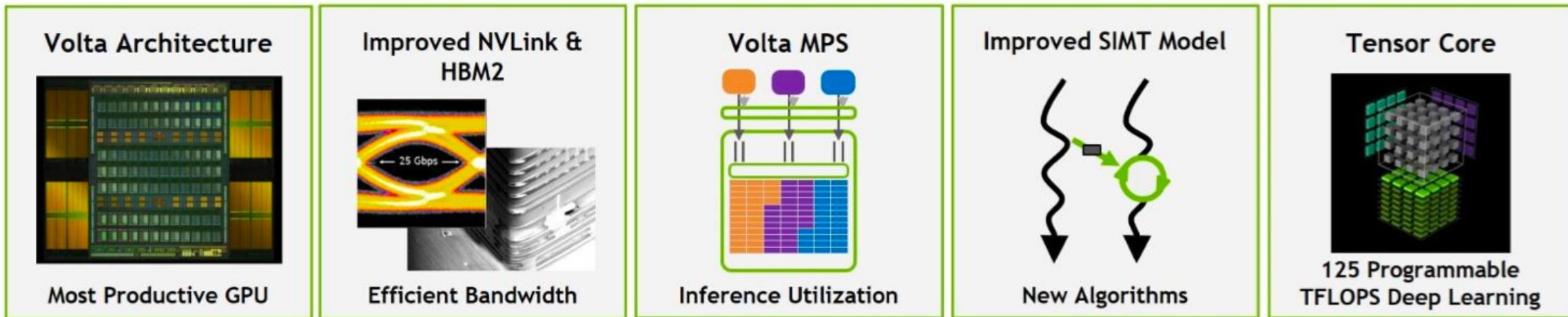
Ascend

7

[www.hiascend.com](http://www.hiascend.com)

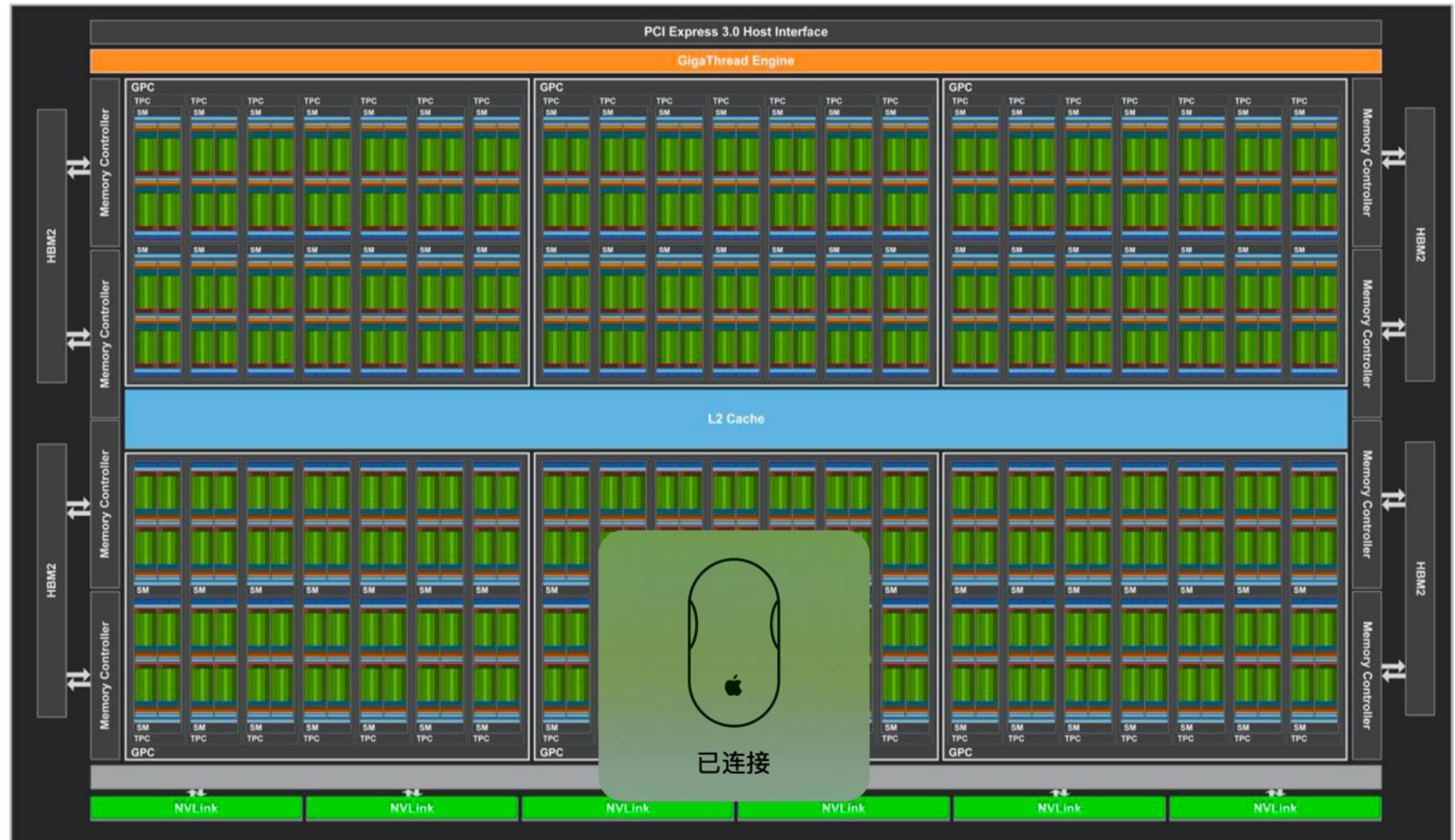
# Volta 伏特架构

- **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。

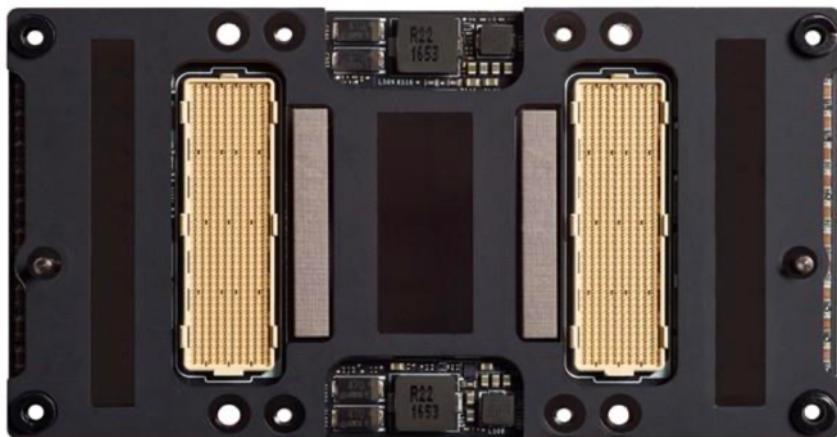


# Volta 伏特架构

- 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  $4 \times 4 \times 4$  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。

$$D = \left( \begin{array}{cccc} A_{0,0} & A_{0,1} & A_{0,2} & A_{0,3} \\ A_{1,0} & A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,0} & A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,0} & A_{3,1} & A_{3,2} & A_{3,3} \end{array} \right) \text{FP16 or FP32} \times \left( \begin{array}{cccc} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,3} \\ B_{1,0} & B_{1,1} & B_{1,2} & B_{1,3} \\ B_{2,0} & B_{2,1} & B_{2,2} & B_{2,3} \\ B_{3,0} & B_{3,1} & B_{3,2} & B_{3,3} \end{array} \right) \text{FP16} + \left( \begin{array}{cccc} C_{0,0} & C_{0,1} & C_{0,2} & C_{0,3} \\ C_{1,0} & C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,0} & C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,0} & C_{3,1} & C_{3,2} & C_{3,3} \end{array} \right) \text{FP16 or FP32}$$

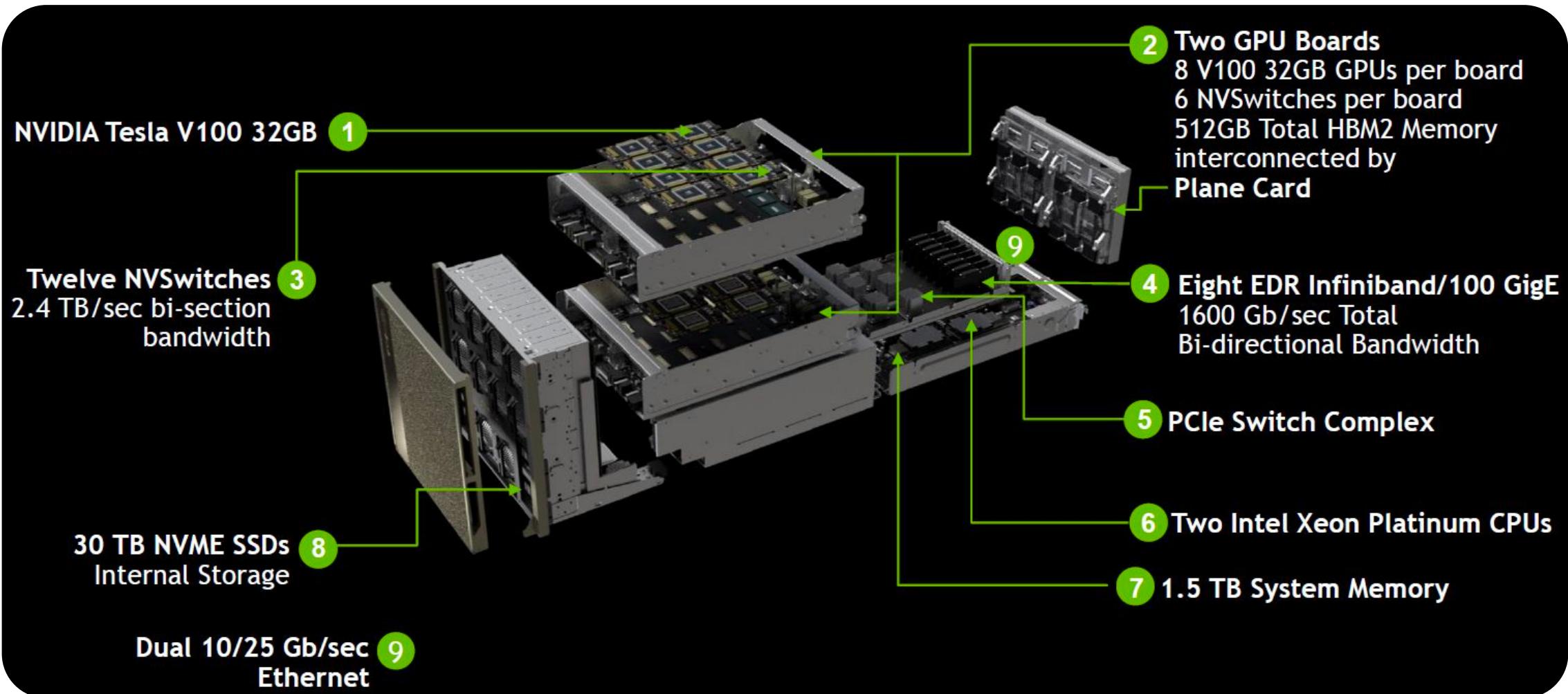
# Volta 伏特架构



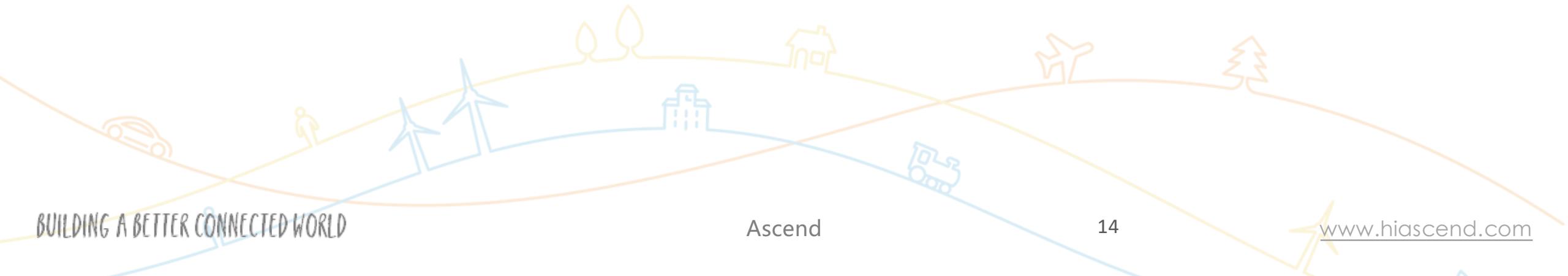
Tesla V100 Powered DGX Station



# Volta 伏特架构



# Turing 架构



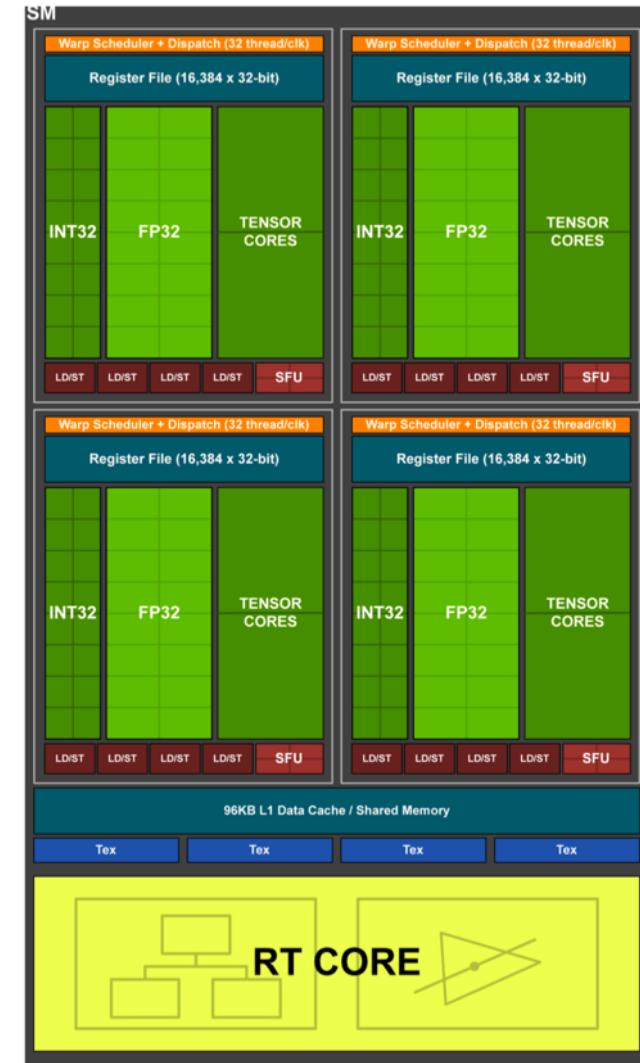
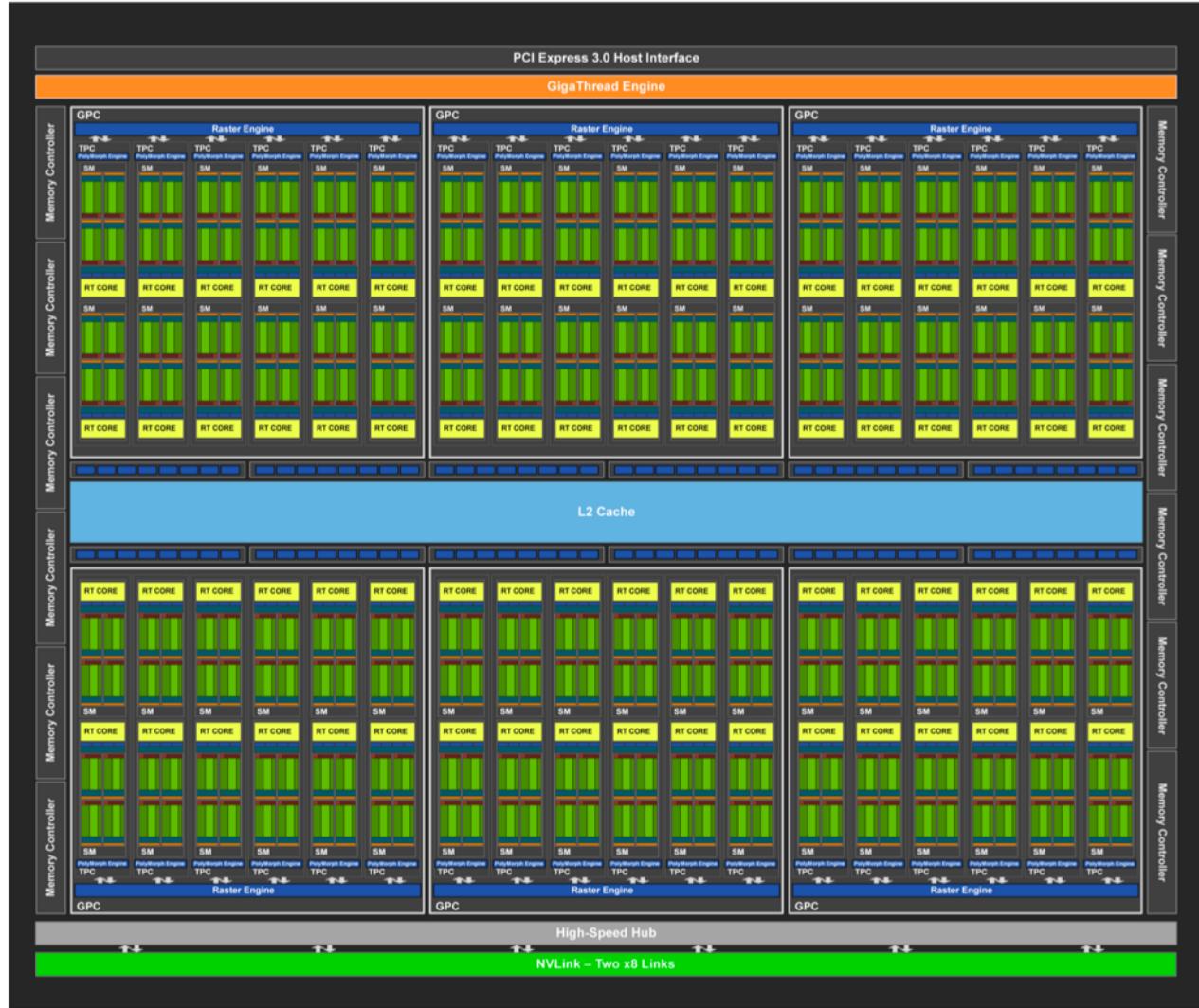
BUILDING A BETTER CONNECTED WORLD

Ascend

14

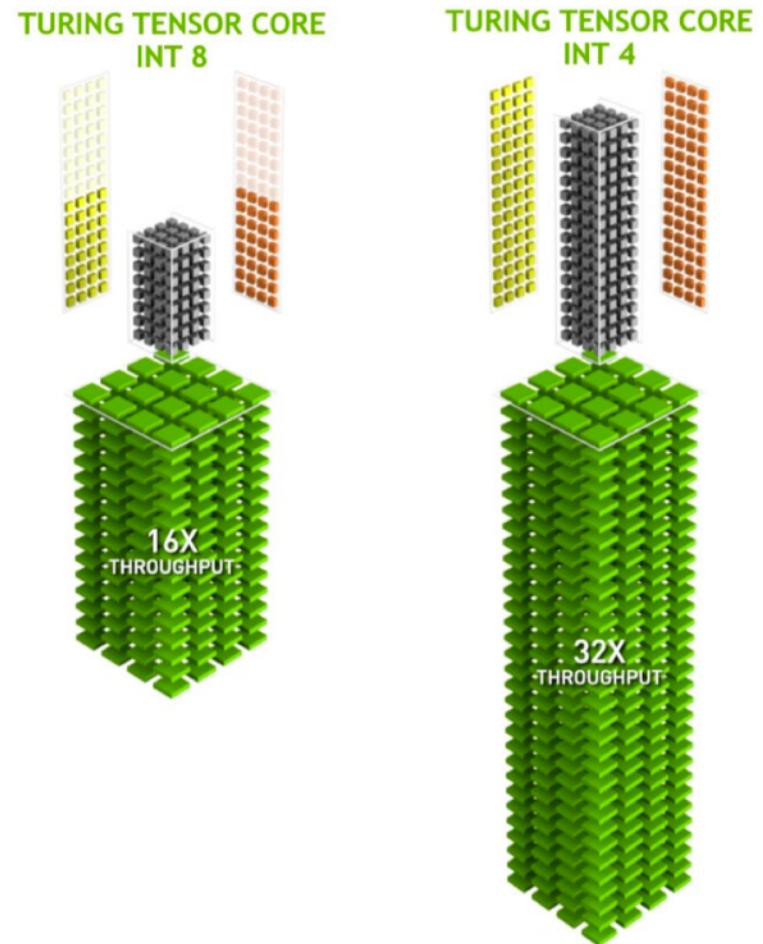
[www.hiascend.com](http://www.hiascend.com)

# Turing 图灵架构



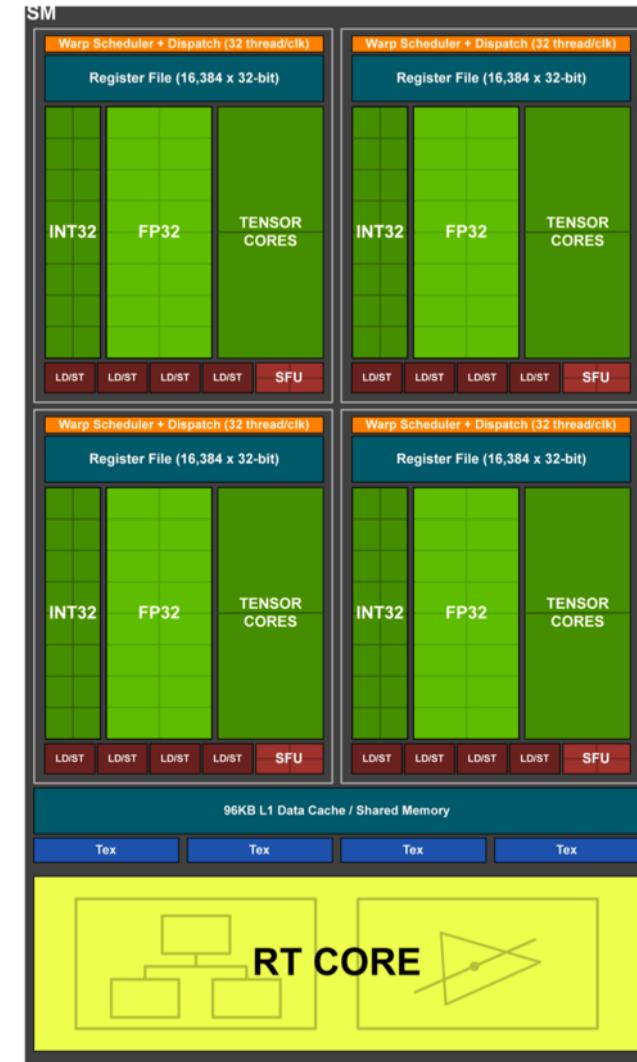
# Turing 图灵架构

- I. 随着深度学习模型的量化部署也渐渐成熟，Turing 架构中的 Tensor Core 增加了对 INT8/INT4/Binary 的支持，为加速 deep learning 的 inference。
2. RT Core (Ray Tracing Core)，主要用来做三角形与光线的求交，并通过BVH结构加速三角形的遍历。由于布置在block之外，相对于普通ALU计算来说是异步。里面分成两个部分，一部分检测碰撞盒来剔除面片，另一部分做真正的相交测试。



# Turing 图灵架构

1. 随着深度学习模型的量化部署也渐渐成熟，Turing 架构中的 Tensor Core 增加了对 INT8/INT4/Binary 的支持，为加速 deep learning 的 inference。
2. RT Core (Ray Tracing Core)，主要用来做三角形与光线的求交，并通过BVH结构加速三角形的遍历。由于布置在block之外，相对于普通ALU计算来说是异步。里面分成两个部分，一部分检测碰撞盒来剔除面片，另一部分做真正的相交测试。



# Turing 图灵架构



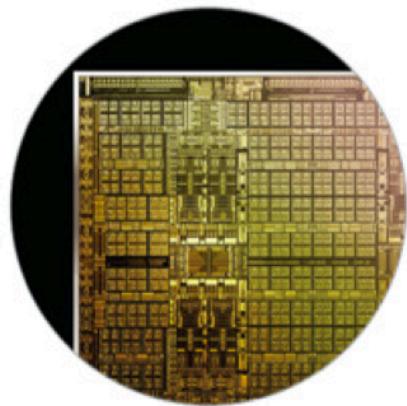
# Turing 图灵架构



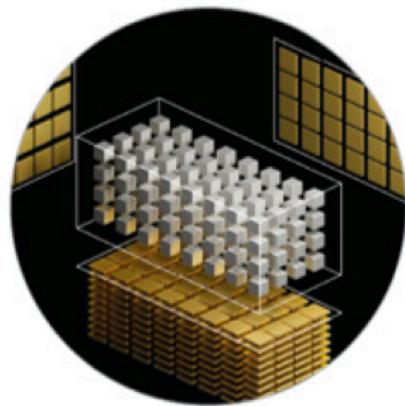
# Ampere 架构

# Ampere 安培架构

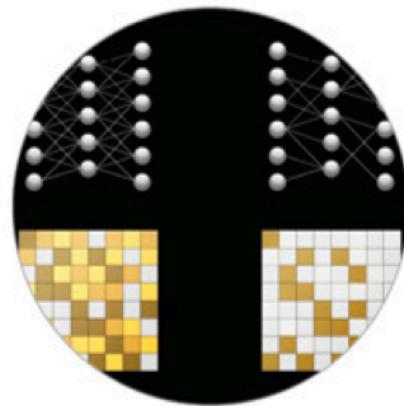
1. NVIDIA Ampere架构：超过540亿个晶体管，使其成为世界上最大的7纳米处理器；
2. Tensor Core3.0：新增 TF32 包括针对AI的扩展，可使FP32精度的AI性能提高20倍；
3. Multi-Instance GPU：多实例GPU，将单个AI00GPU划分为多达七个独立GPU，为不同任务提供不同算力；
4. NVIDIA NVLink2.0：GPU间高速连接速度加倍，可在服务器中提供有效的性能扩展；
5. 结构稀疏性：利用了 AI 数学固有的稀疏特性来使性能提高一倍。



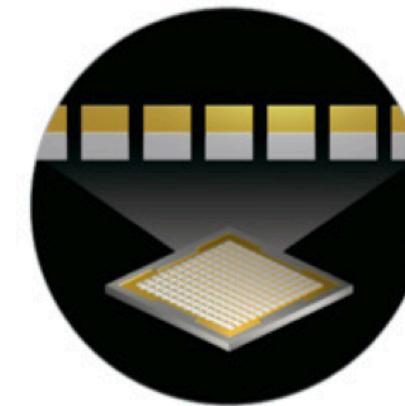
54 BILLION XTORS



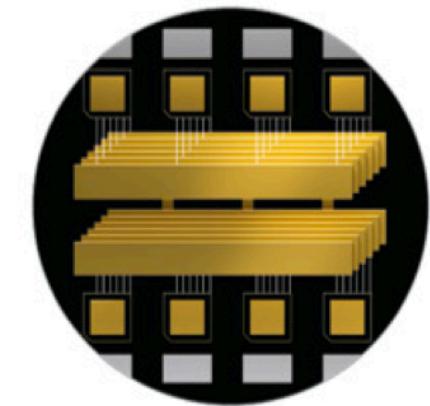
3<sup>rd</sup> GEN  
TENSOR CORES



SPARSITY  
ACCELERATION



MIG



3<sup>rd</sup> GEN  
NVLINK & NVSWITCH

# Ampere 安培架构

- NVIDIA A100基于7nm Ampere GA100 GPU，具有6912 CUDA内核和432 Tensor Core，540亿个晶体管数，108个流式多处理器。采用第三代NVLINK，GPU和服务器双向带宽为4.8 TB/s，GPU间的互连速度为600 GB/s。另外，Tesla A100在5120条内存总线上的HBM2内存可达40GB。

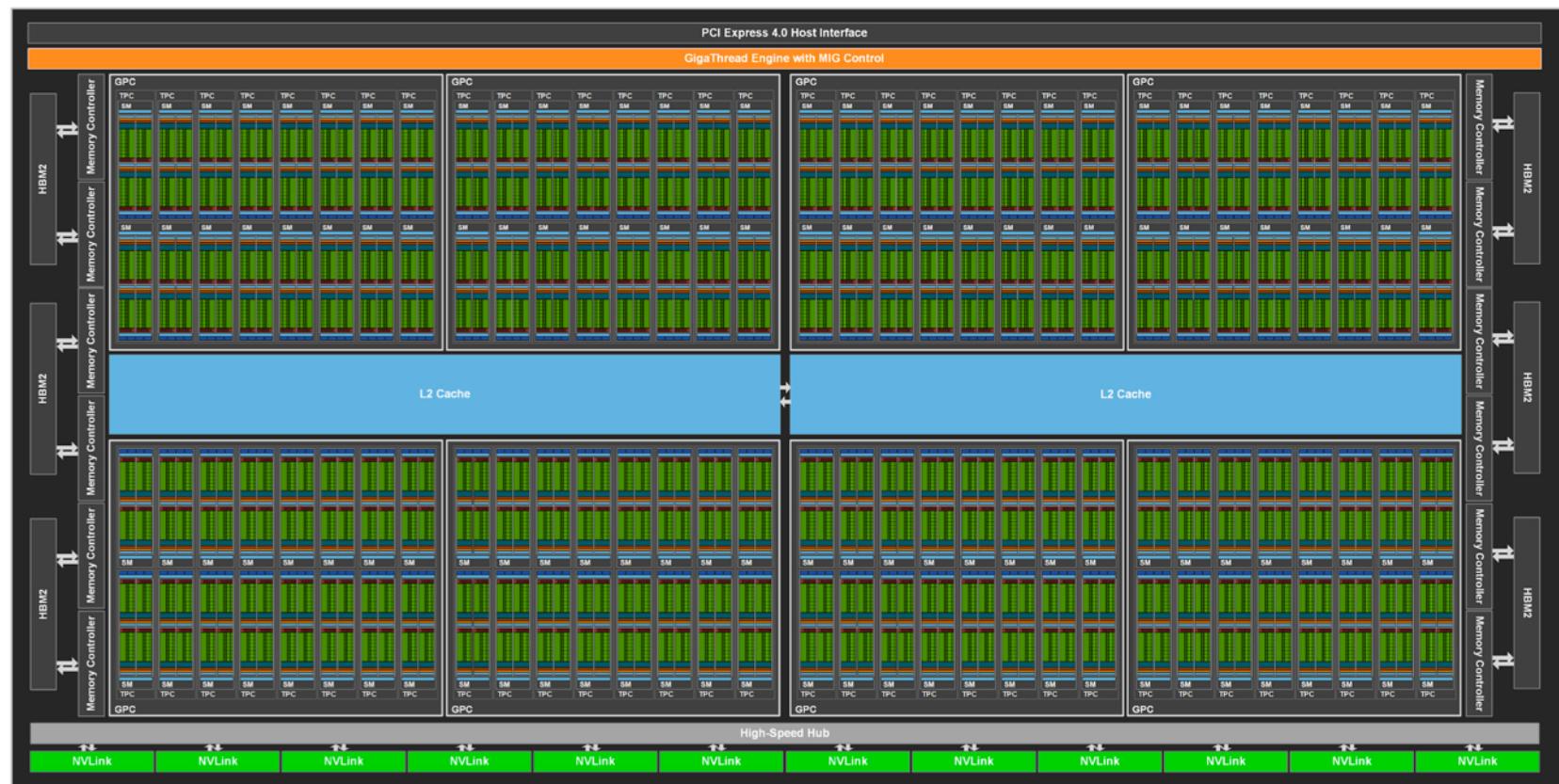
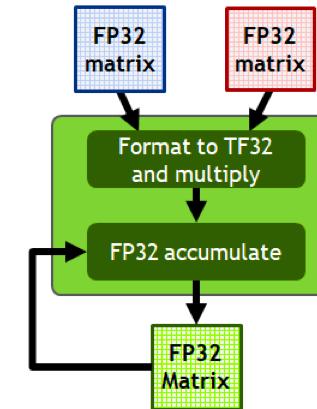
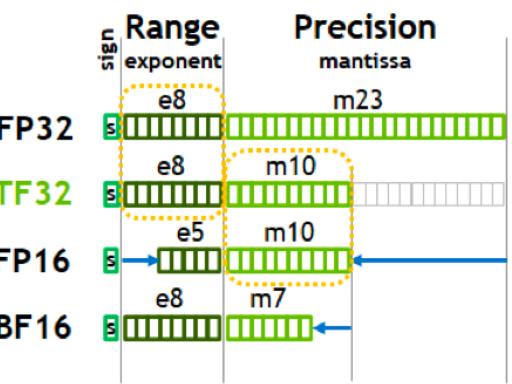


Figure 6. GA100 Full GPU with 128 SMs (A100 Tensor Core GPU has 108 SMs)

# Ampere 安培架构

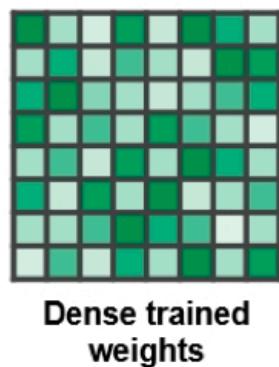
- Ampere 新加入了TF32, BF16, FP64的支持。



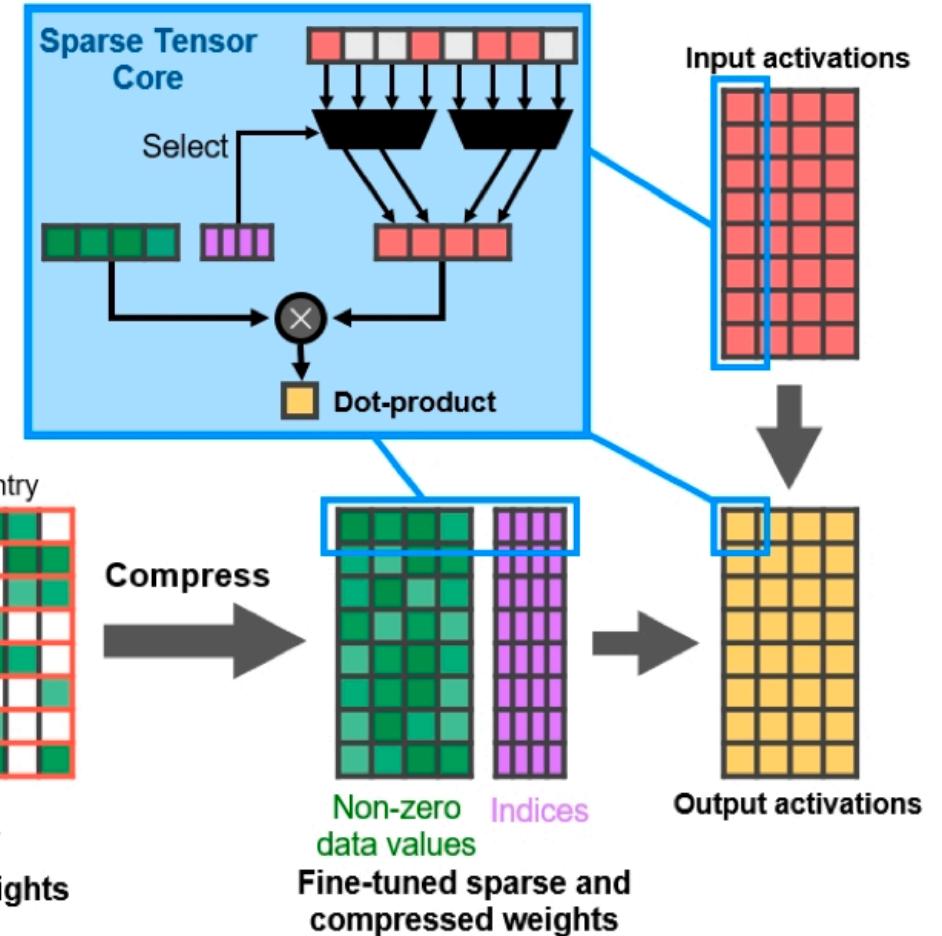
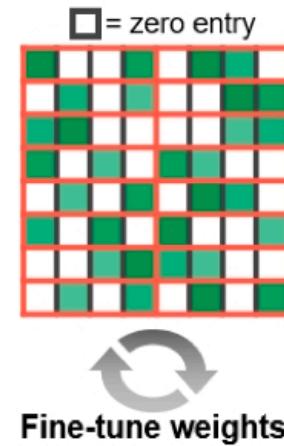
	INPUT OPERANDS	ACCUMULATOR	TOPS	X-factor vs. FFMA	SPARSE TOPS	SPARSE X-factor vs. FFMA
A100	FP32	FP32	19.5	1x	-	-
	TF32	FP32	156	8x	312	16x
	FP16	FP32	312	16x	624	32x
	BF16	FP32	312	16x	624	32x
	FP16	FP16	312	16x	624	32x
	INT8	INT32	624	32x	1248	64x
	INT4	INT32	1248	64x	2496	128x
	BINARY	INT32	4992	256x	-	-
	IEEE FP64	IEEE FP64	19.5	1x	-	-

# Ampere 安培架构

- 细粒度的结构化稀疏，TensorCore支持一个2:4的结构化稀疏矩阵与另一个稠密矩阵直接相乘。

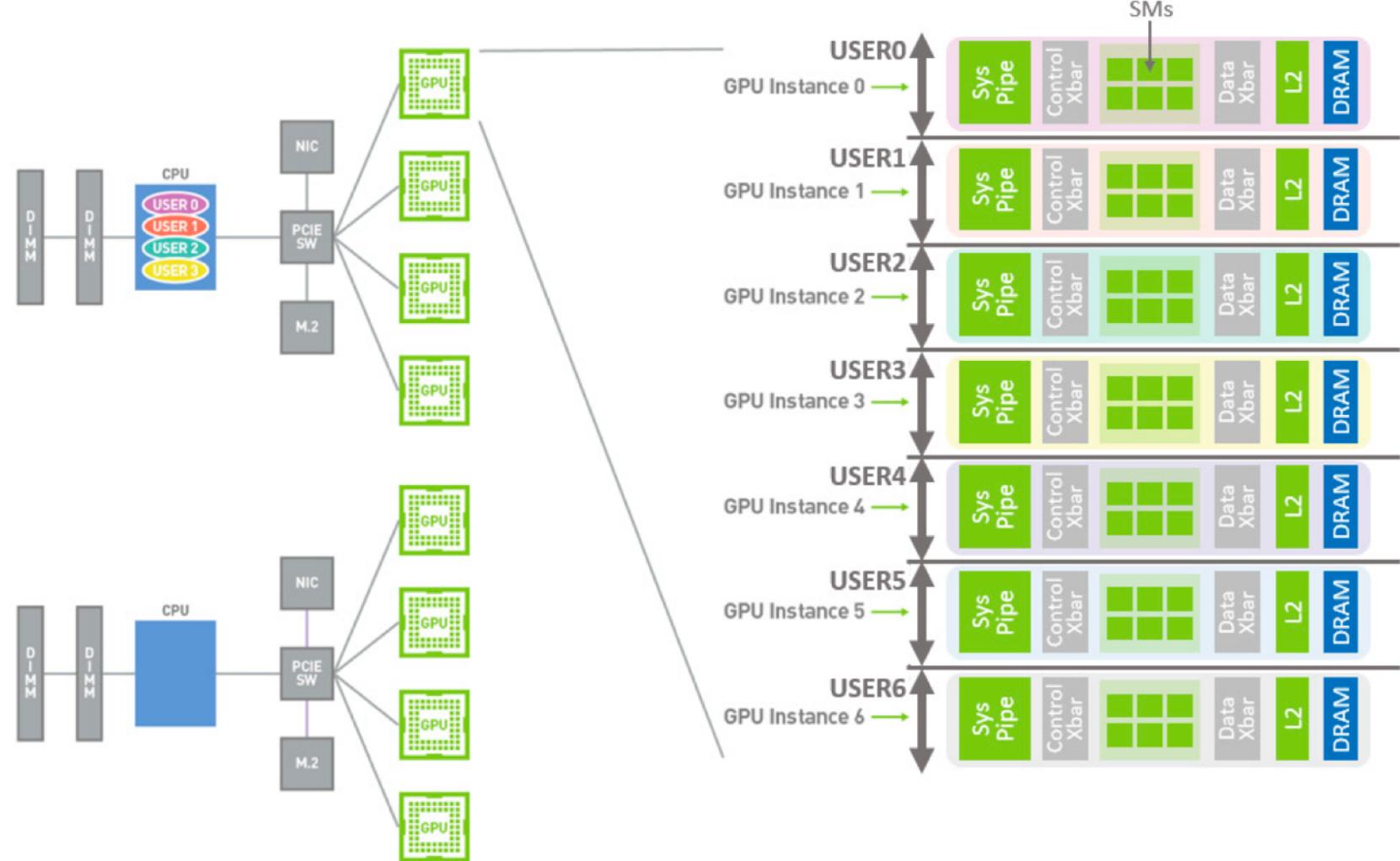


Fine-grained structured pruning  
2:4 sparsity: 2 non-zero out of 4 entries



# Ampere 安培架构

1. 每个AI00可以被分为7个GPU实例被不同的任务使用，用户可以将这些虚拟的GPU实例当成真实的GPU使用。
2. 为云计算厂商提供算力切分和多用户租赁服务。



# Ampere 安培架构

GPUs	<b>8x NVIDIA A100</b>
GPU Memory	320 GB total
Peak performance	5 petaFLOPS AI   10 petaOPS INT8
NVSwitches	6
System Power Usage	<b>6.5kW max</b>
CPU	<b>Dual AMD Rome 7742</b> 128 cores total, 2.25 GHz(base), 3.4GHz (max boost)
System Memory	<b>1TB</b>
Networking	<b>8x Single-Port Mellanox ConnectX-6 200Gb/s HDR Infiniband (Compute Network)</b> <b>1x (or 2x*) Dual-Port Mellanox ConnectX-6 200GB/s HDR Infiniband (Storage Network also used for Eth*)</b>
Storage	OS: 2x 1.92TB M.2 NVME drives Internal Storage: 15TB (4x 3.84TB) U.2 NVME drives
Software	Ubuntu Linux OS ( <b>5.3+ kernel</b> )
System Weight	271 lbs (123 kgs)
Packaged System Weight	315 lbs (143 kgs)
Height	<b>6U</b>
Operating temp range	<b>5 °C to 30 °C (41 °F to 86 °F)</b>

\* Optional upgrades

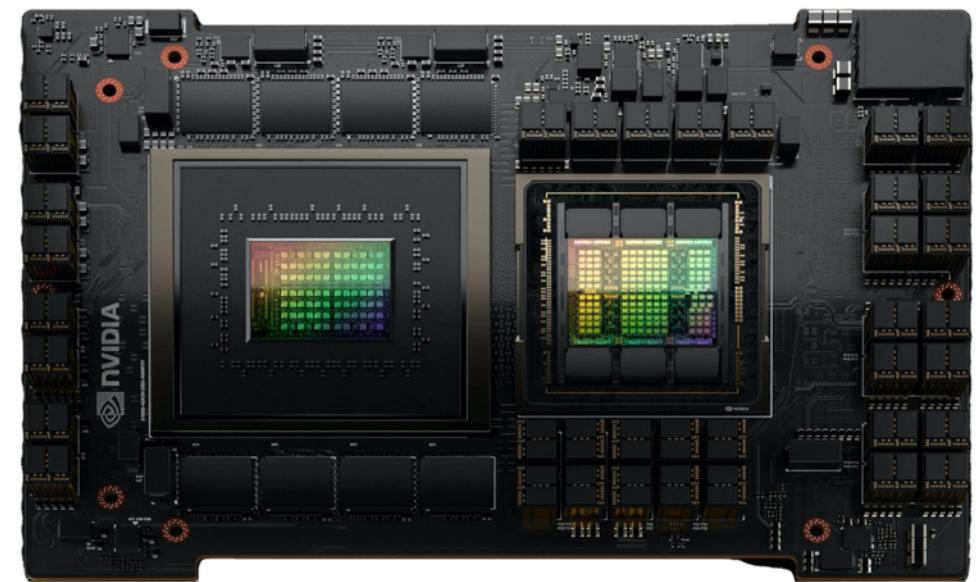
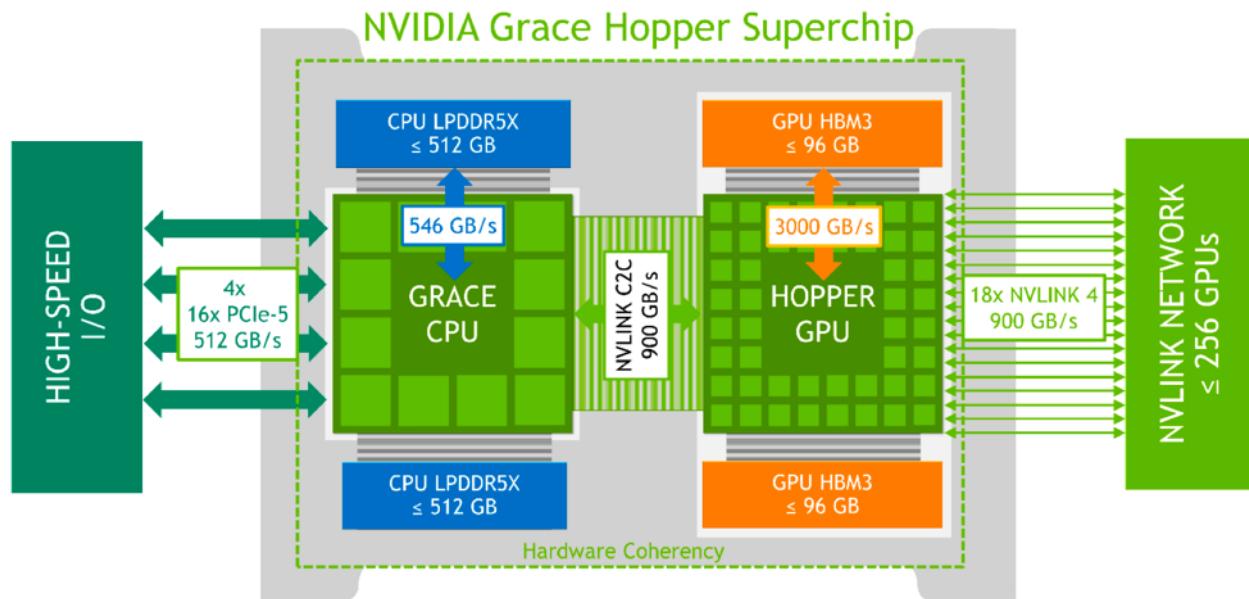
\* Optional upgrades



# Hopper 架构

# Hopper 赫柏架构

- NVIDIA Grace Hopper Superchip 架构将NVIDIA Hopper GPU的突破性性能与NVIDIA Grace CPU 的多功能性结合在一起，在单个超级芯片中与高带宽和内存一致的 NVIDIA NVLink Chip-2-Chip (C2C)互连相连，并且支持新的NVIDIA NVLink 切换系统。



Inside NVIDIA's First GPU-CPU Superchip

# Hopper 赫柏架构

第一个真正的异构加速平台，适用于高性能计算(HPC) 和AI工作负载

- **NVIDIA Grace CPU :**

1. 72 个 Arm Neoverse V2 内核，每个内核 Armv9.0-A ISA 和 4 个 128 位 SIMD 单元；
2. 512 GB LPDDR5X 内存，提供高达 546 GB/s 的内存带宽；
3. 117 MB 的 L3 缓存，内存带宽高达 3.2 TB/s；
4. 64 个 PCIe Gen5 通道；

GEN  
SPEE  
D  
C  
O  
H  
I  
N  
V  
L  
I  
N  
K  
-  
C  
2  
C

4x  
16x PCIe-5  
256 GB/s

## NVIDIA Grace Hopper



- 英伟达 NVLink-C2C :

1. Grace CPU 和 Hopper GPU 之间的硬件一致性互连。
2. 高达 900 GB/s 的总带宽，450 GB/s/dir。
3. 扩展 GPU 内存功能使 Hopper GPU 能够将所有 CPU 内存寻址为 GPU 内存。每个 Hopper GPU 可以在超级芯片内寻址多达 608 GB 的内存。

- **NVIDIA Hopper GPU :**

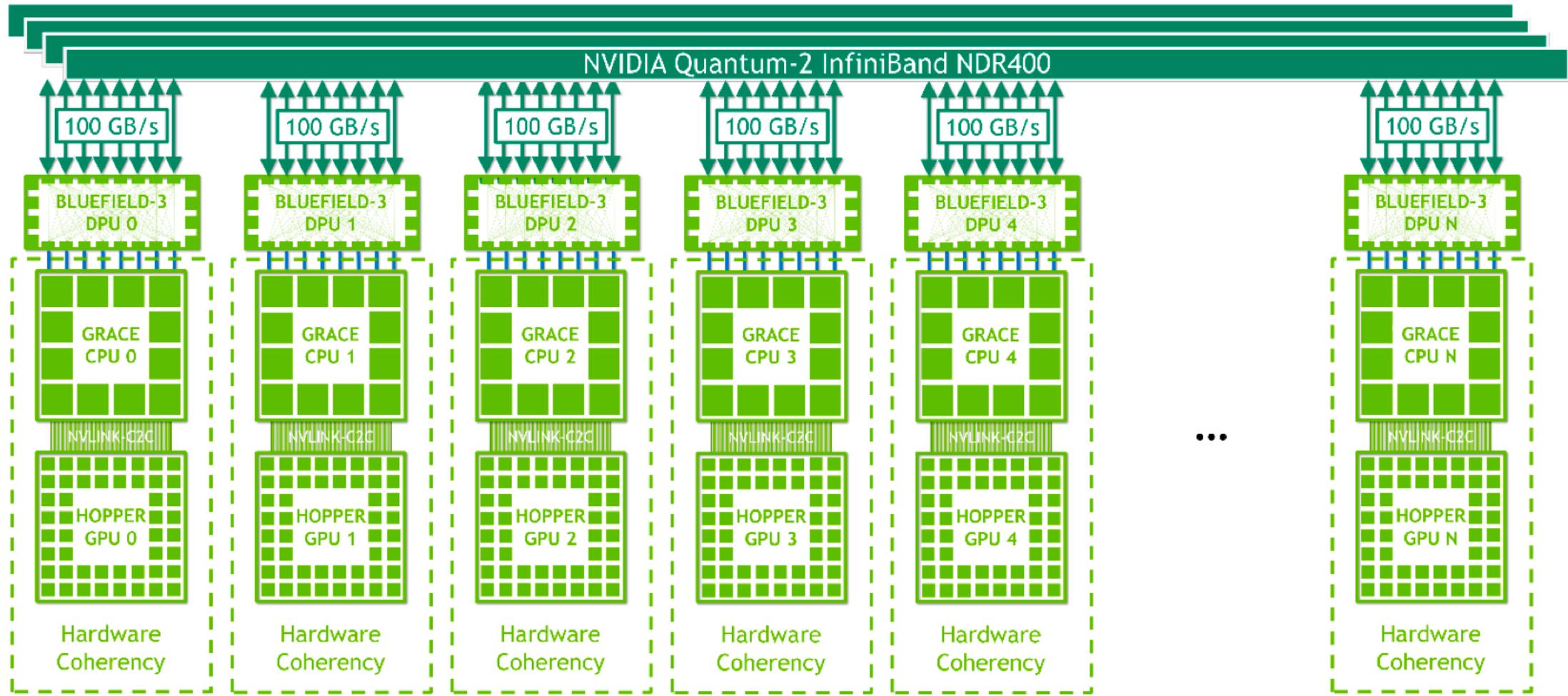
1. 144 个第四代 Tensor Core、Transformer Engine、DPX 和 3 倍高 FP32 和 FP64 的 SM；
2. 96 GB HBM3 内存提供高达 3000 GB/s 的速度。
3. 60 MB 二级缓存；
4. NVLink 4 和 PCIe 5。

HOPPER

## NVIDIA NVLink 切换系统：

1. 使用 NVLink 4 连接多达 256 个 NVIDIA Grace Hopper 超级芯片。
2. 每个连接 NVLink 的 Hopper GPU 都可以寻址网络中所有超级芯片的所有 HBM3 和 LPDDR5X 内存，最高可达 150 TB 的 GPU 可寻址内存。

# Hopper 赫柏架构



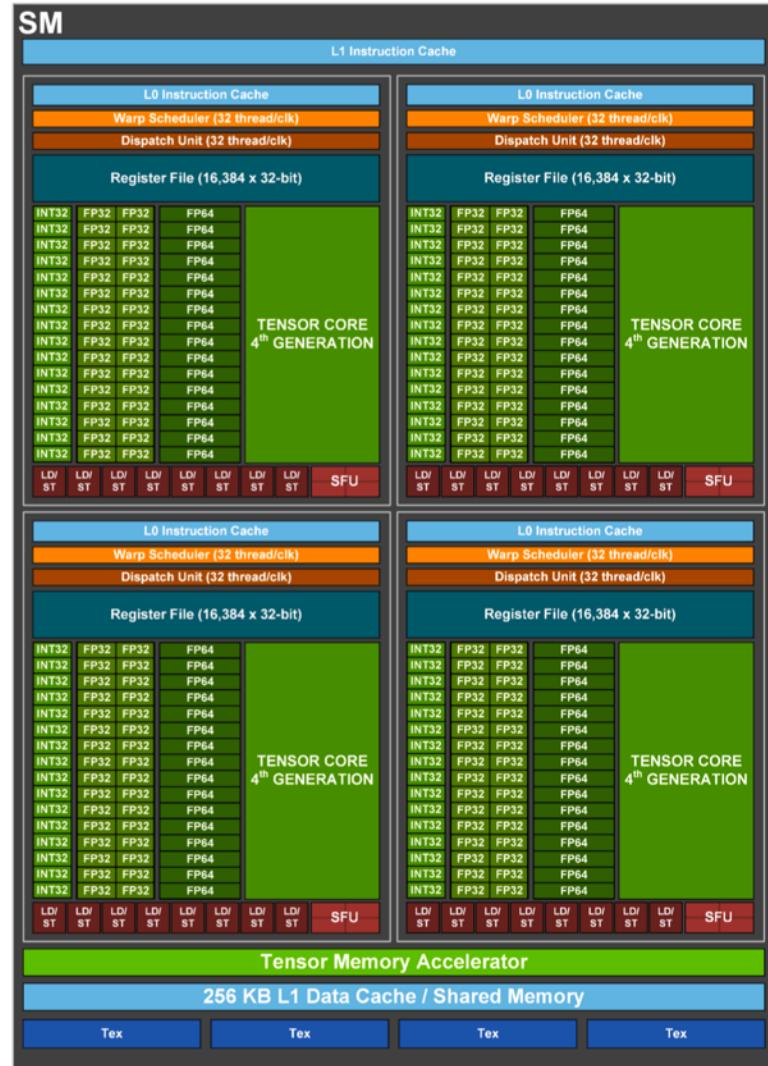
# Hopper 赫柏架构

- GPC 8组， 66组TPC、132组SM，总计有16896个CUDA核心、528个Tensor核心、50MB二级缓存。显存为新一代HBM3，容量80GB，位宽5120-bit，带宽高达3TB/s



# Hopper 赫柏架构

- SM 结构：
  1. 4 个 Warp Scheduler , 4 个 Dispatch Unit (与 A100 一致)
  2. 128 个 FP32 Core (  $4 * 32$  ) ( 相比 A100 翻倍 )
  3. 64 个 INT32 Core (  $4 * 16$  ) ( 与 A100 一致 )
  4. 64 个 FP64 Core (  $4 * 16$  ) ( 相比 A100 翻倍 )
  5. 4 个 TensorCore (  $4 * 1$  )
  6. 32 个 LD/ST Unit (  $4 * 8$  ) ( 与 A100 一致 )
  7. 16 个 SFU (  $4 * 4$  ) ( 与 A100 一致 )
  8. 相比 A100 增加了一个 Tensor Memory Accelerator
- 每个 Process Block：
  1. 1 个 Warp Scheduler , 1 个 Dispatch Unit ( 与 A100 一致 )
  2. 32 个 FP32 Core ( 相比 A100 翻倍 )
  3. 16 个 INT32 Core ( 与 A100 一致 )
  4. 16 个 FP64 Core ( 相比 A100 翻倍 )
  5. 1 个 TensorCore
  6. 8 个 LD/ST Unit ( 与 A100 一致 )
  7. 4 个 SFU ( 与 A100 一致 )



## NVIDIA H100 AT EVERY SCALE

Mainstream Servers to DGX to DGX SuperPOD



CLOUD PROVIDERS

Alibaba Cloud

aws

BAIDU AI CLOUD

Google Cloud

Microsoft Azure

ORACLE  
Cloud Infrastructure

Tencent Cloud

SYSTEM PROVIDERS

Atos

cisco

DELL Technologies

FUJITSU

GIGABYTE<sup>®</sup>

H3C

Hewlett Packard  
Enterprise

inspur

Lenovo

Netrix

SUPERMICRO



# 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> NVIDIA GPU架构演进
3. [https://zhuanlan.zhihu.com/p/258196004?utm\\_id=0](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](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](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>



# 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.