对HPC和AI前所未有的加速

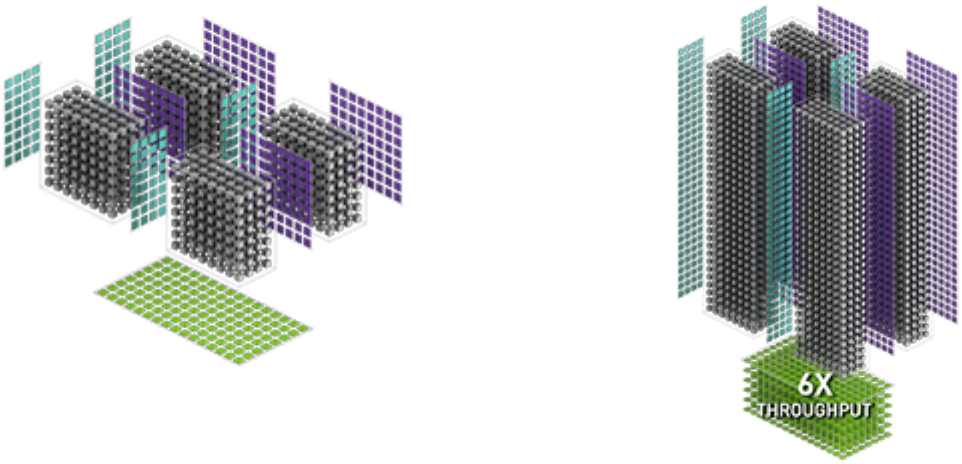
Tensor Core能够进行mixed-precision计算，能在保证准确度的情况动态调整计算精度增加吞吐量。

- Revoluinary AI Training
- Breakthrough AI Inference
- Advanced HPC

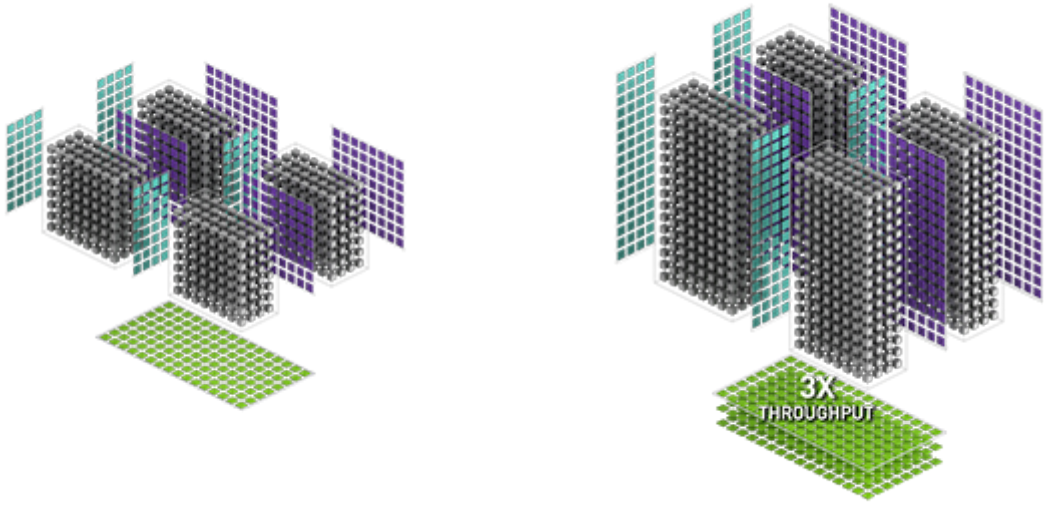
NVIDIA H100 Tensor Cores (4th)

自从Tensor Core技术的引入，Nvidia GPU的峰值性能已经增加了60倍。Hopper架构的第四代Tensor Core通过引入Transfor Engine用一个新的8-bit floating point精度(FP8)达到了6倍于FP16的性能提升(for trillion-parameter model training).

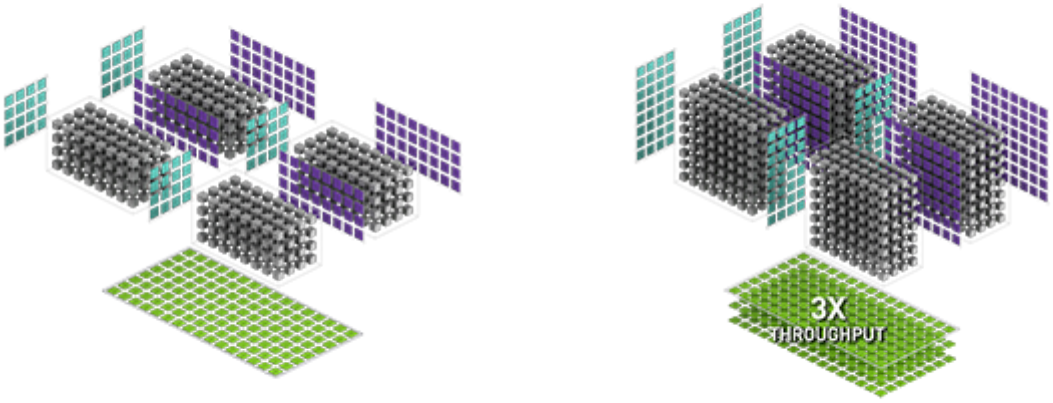
- FP8
  Tranformer AI网络由于大和数学计算，其训练时间可能被拉长到数月。Hopper的新FP8的性能可以达到6倍于Ampere的FP16。FP8被用在Tranformer Engine中，Hopper的Tensor Core实际就是被设计用于加速Tranformer模型的加速。



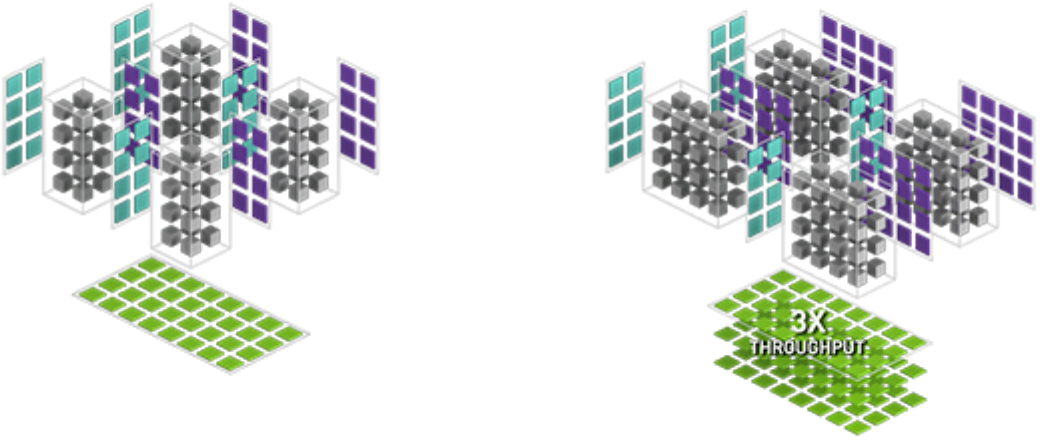
- FP16
  Hopper Tensor Core boost FP16 for deep learning.



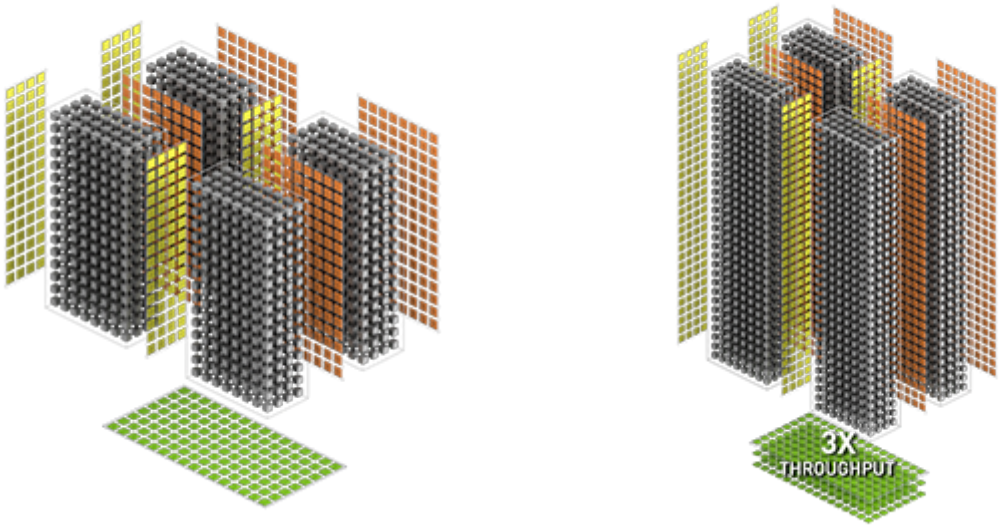
- TF32
  Delivering AI speedup



- FP64
  Accelerating a whole range of HPC application that need double-precision math

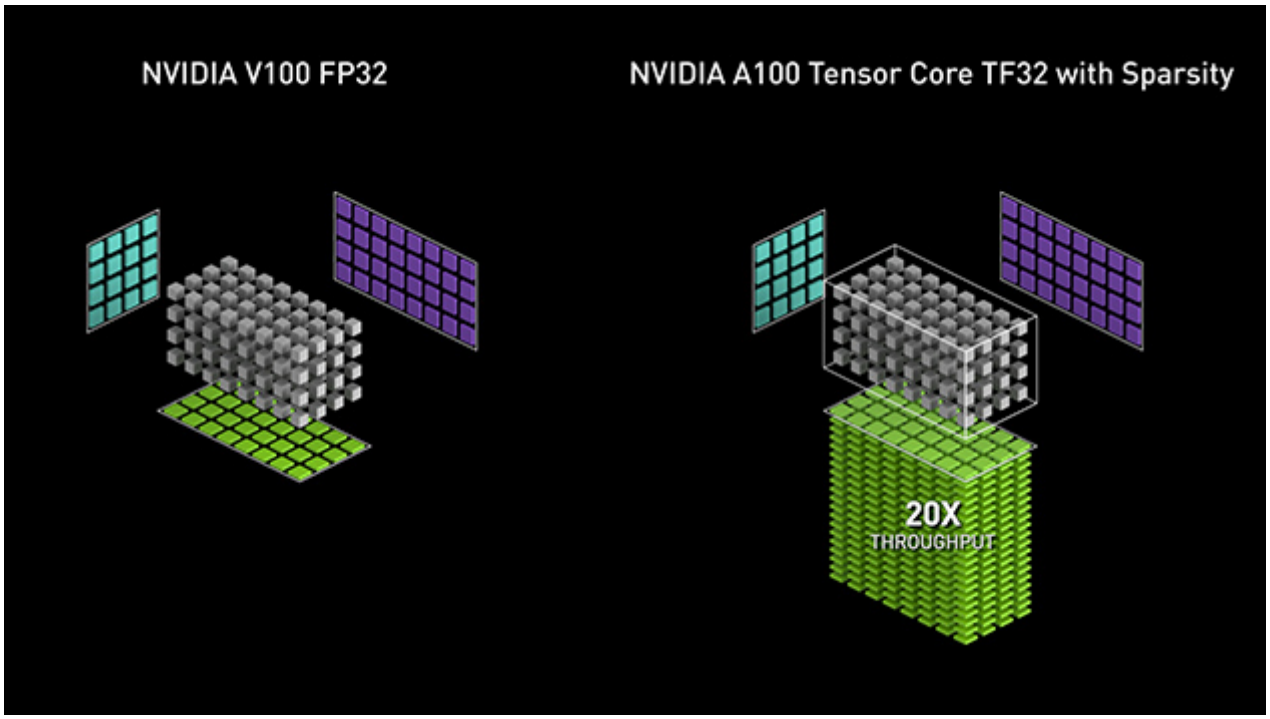


- INT8
  第一次被引入是在Turing架构中，INT8 Tensor Core大大加速了推理的吞吐量



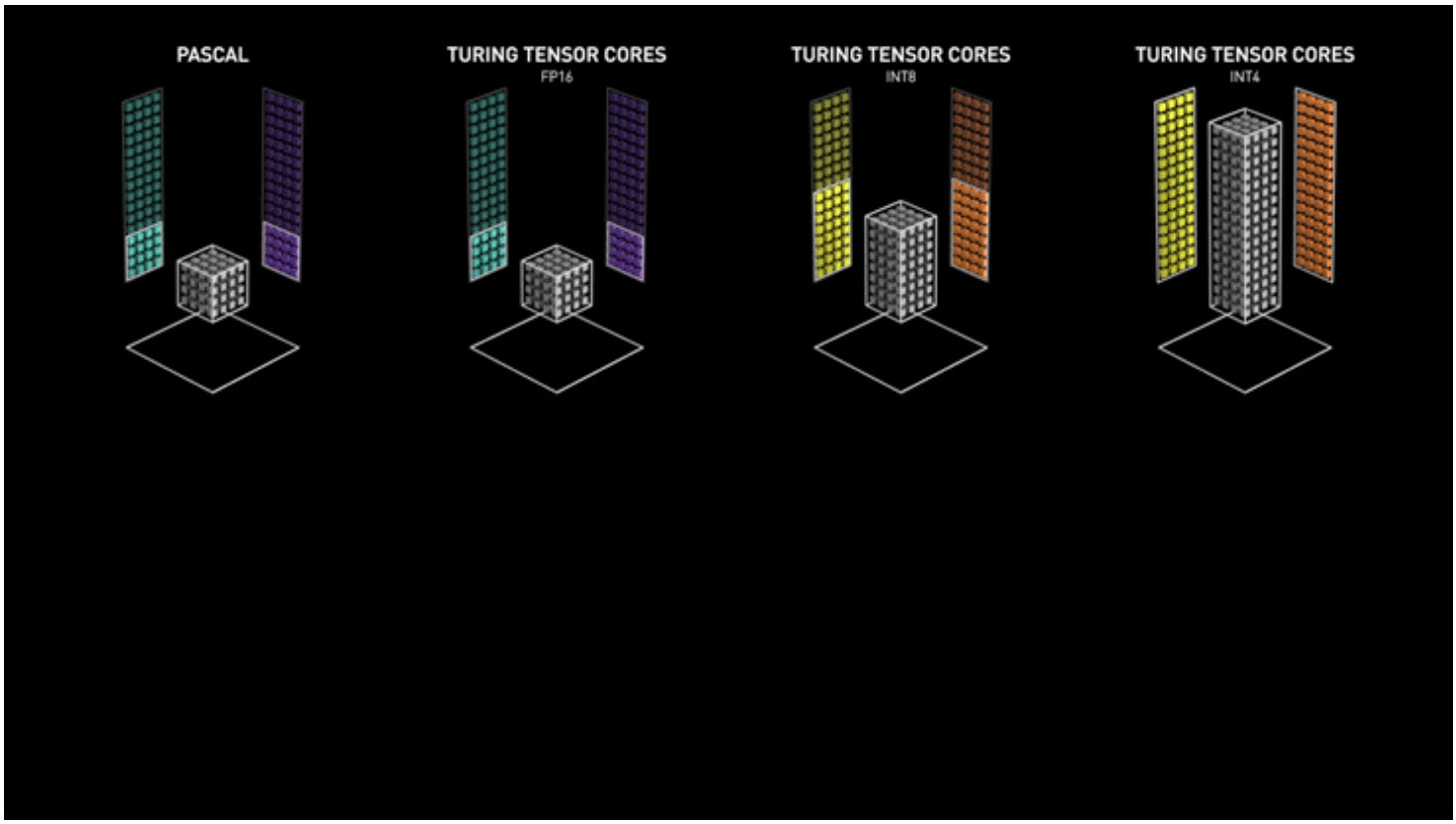
NVIDIA Ampere Architecture Tensor Cores (3th)

Ampere架构Tensor Core的构建是基于前一代的创新，仅引入了新精度的数据类型TF32和FP64来加速和简化AI的使用，并将Tensor Core的功能扩展到了HPC。



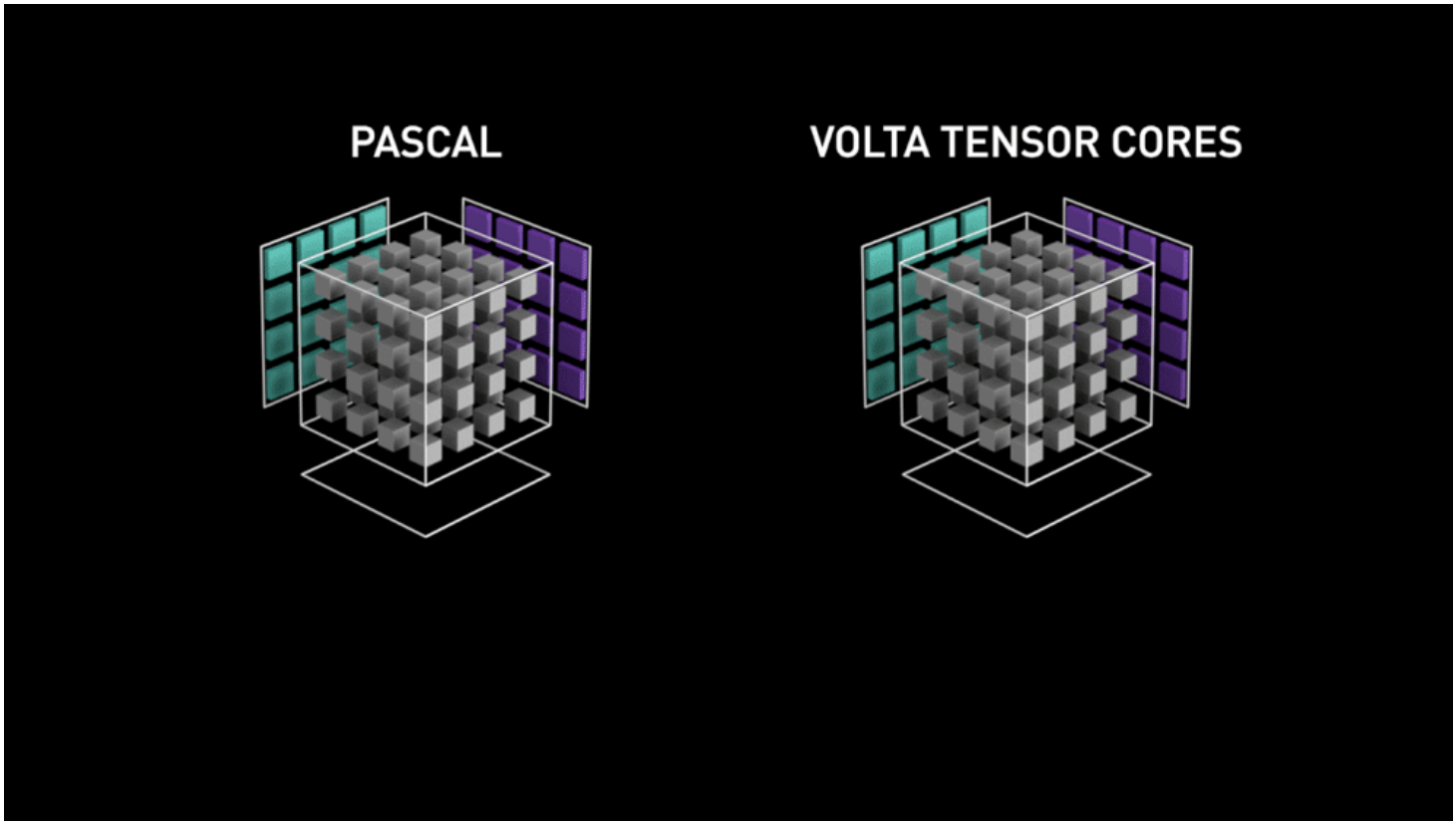
NVIDIA Turing Tensor Cores (2th)

NVIDIA Turing™ Tensor Core technology features multi-precision computing for efficient AI inference. Turing Tensor Cores provide a range of precisions for deep learning training and inference, from FP32 to FP16 to INT8, as well as INT4, to provide giant leaps in performance over NVIDIA Pascal™ GPUs



NVIDIA Volta Tensor Cores (1th)

Designed specifically for deep learning, the first-generation Tensor Cores in NVIDIA Volta™ deliver groundbreaking performance with mixed-precision matrix multiply in FP16 and FP32, up to 12X higher peak teraFLOPS (TFLOPS) for training and 6X higher peak TFLOPS for inference over NVIDIA Pascal.



The Most Powerful End-to-End AI and HPC Data Center Platform

	Hopper	Ampere	Turing	Volta
Supported Tensor Core precisions	FP64, TF32, bfloat16, FP16, FP8, INT8	FP64, TF32, bfloat16, FP16, INT8, INT4, INT1	FP16, INT8, INT4, INT1	FP16
Supported CUDA® Core precisions	FP64, FP32, FP16, bfloat16, INT8	FP64, FP32, FP16, bfloat16, INT8	FP64, FP32, FP16, INT8	FP64, FP32, FP16, INT8

Reference

- https://www.nvidia.com/en-us/data-center/tensor-cores/