Introduction to GPUs, Inference and Model Compression

Ziv Ilan - Solution Architect, NVIDIA

Sergio Perez - Solution Architect, NVIDIA

Harshita Seth - Solution Architect, NVIDIA

About Us

Ziv Ilan, EMEA

- Senior Deep Learning Solutions Architect @ NVIDIA - Supporting delivery of AI / Deep Learning solutions
- Covering inference, model compression, customization, and evaluation

Sergio Perez, EMEA

- Senior Deep Learning Solutions Architect @ NVIDIA - Supporting delivery of AI / Deep Learning solutions
- Covering inference, customization, evaluation and RAG systems

Harshita Seth, EMEA

- Senior Partner Solutions Architect @ NVIDIA -Supporting delivery of AI / Deep Learning solutions
- Covering model compression and evaluation

Agenda of the day

- Intro to GPUs (11:15 to 12:00)
- Model compression overview (13:30 to14:30)
- Practical tutorial about model compression (14:30 to 16:30)

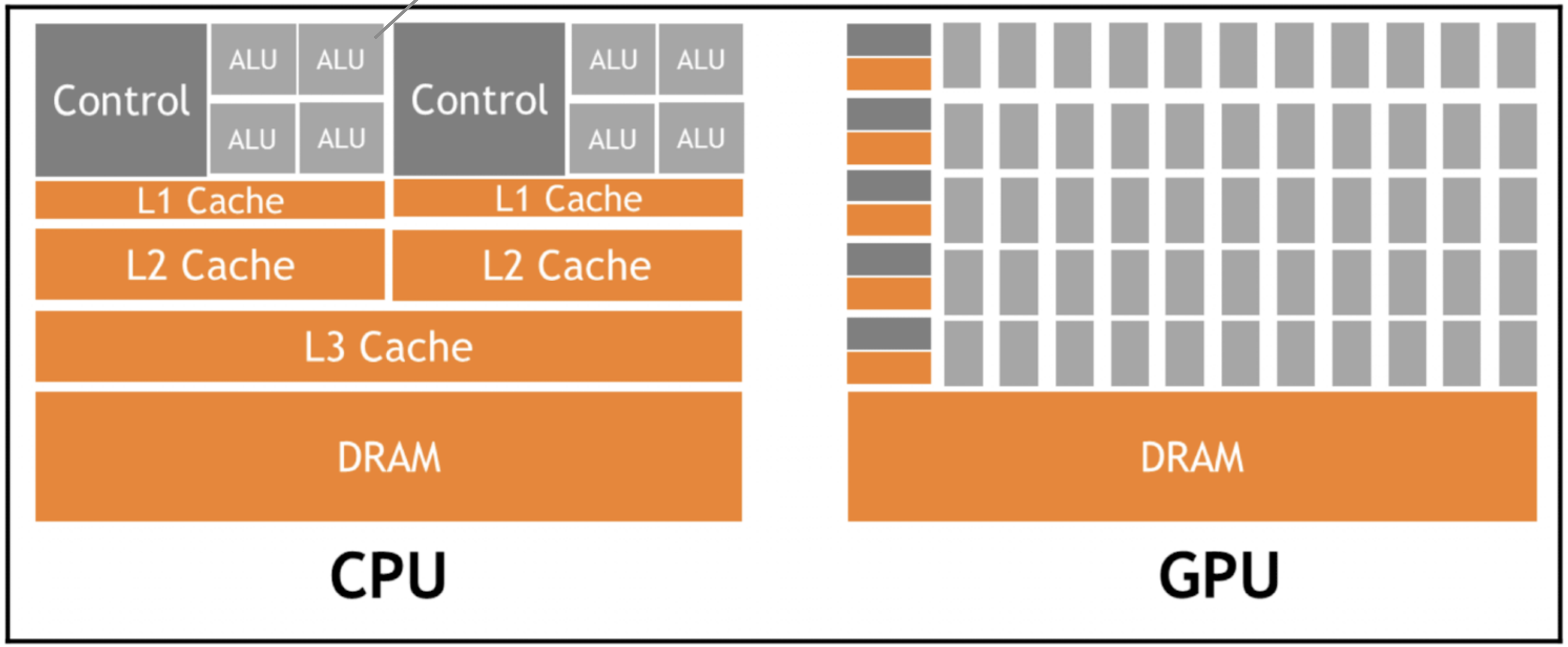
Agenda of introduction

- What’s a GPU?
- Memory versus compute
- Is NVIDIA just a hardware company?
- Training and inferencing LLMs
- Model architectures

Differences between a CPU and a GPU

There are many types of GPUs! Let’s see an example of one

ALU: Arithmetic Logic Unit

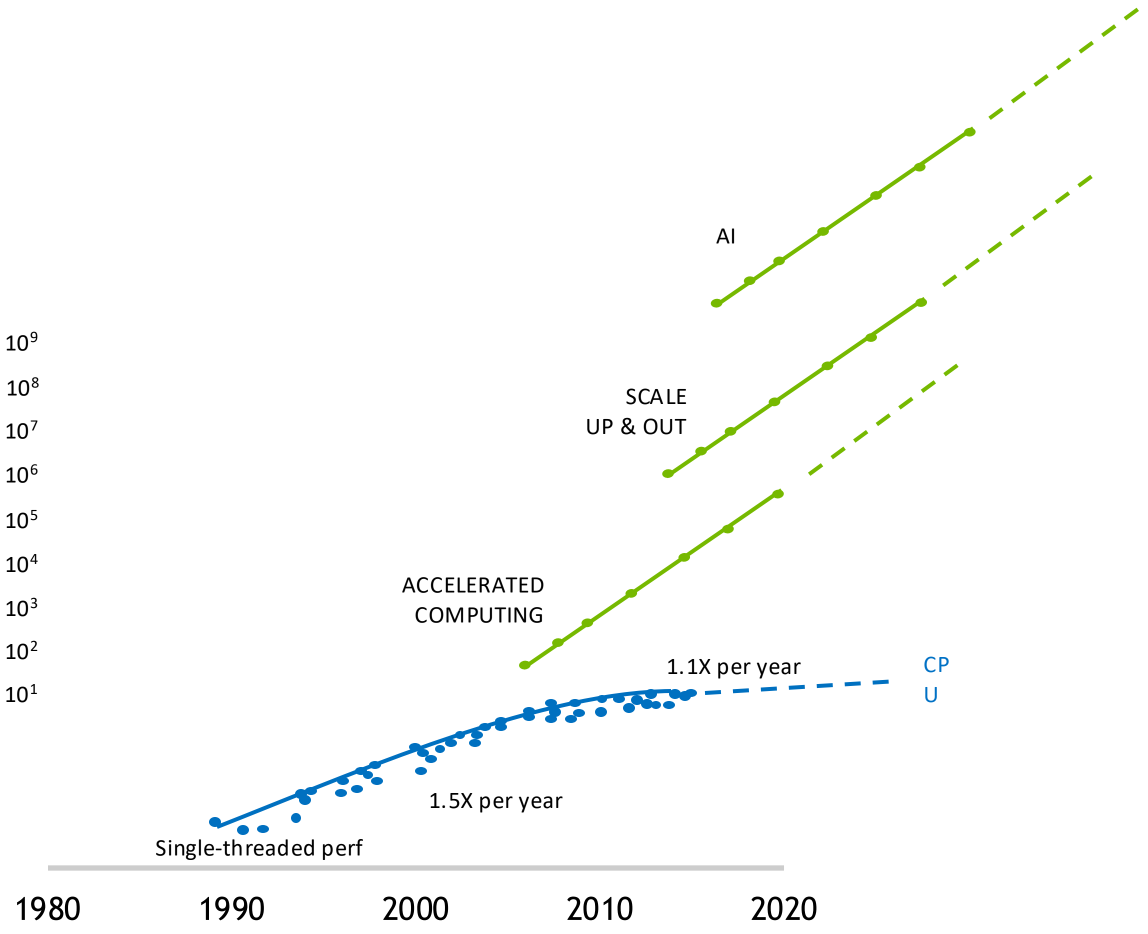


Hide latency to access data ↔ Many operations in parallel

Image from book “Learn CUDA programming: a beginners guide to GPU programming and parallel computing”

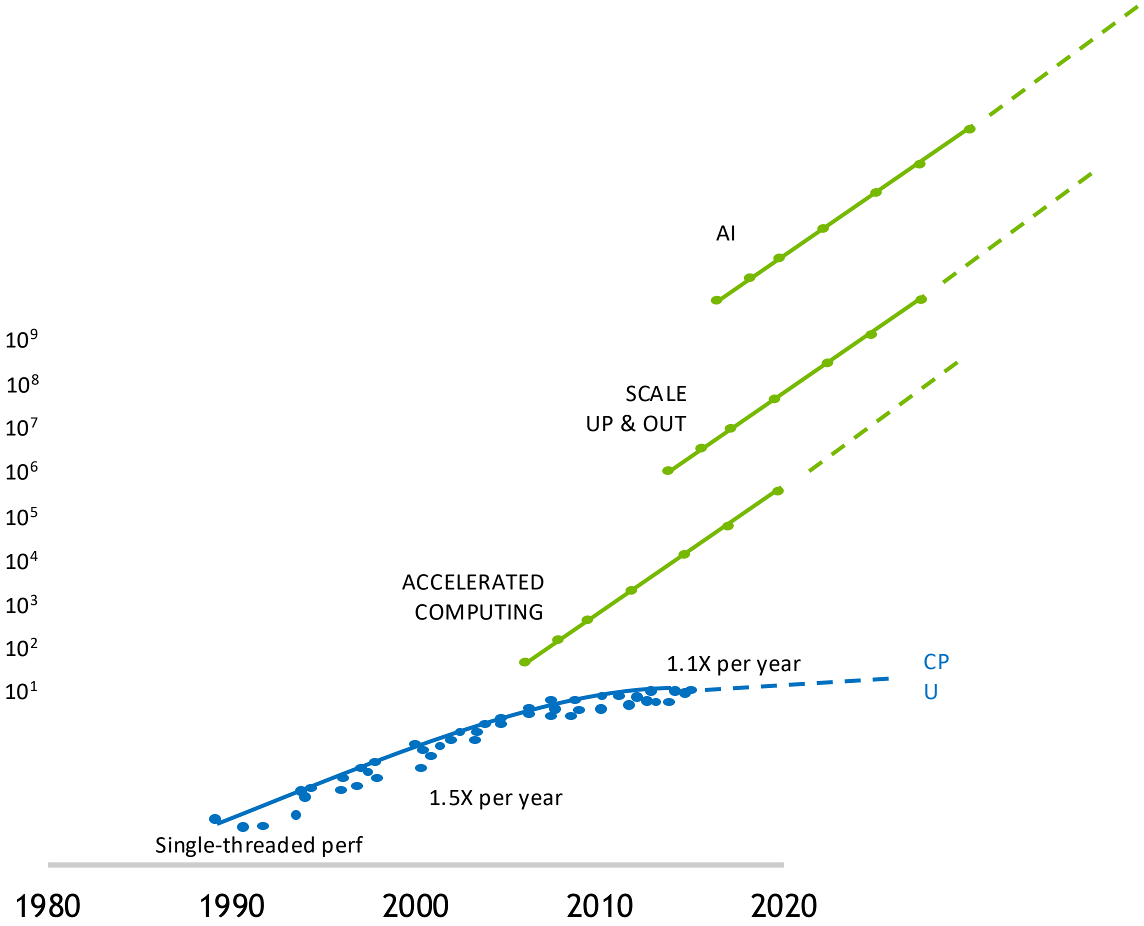
Getting Million-X Speedups to Power AI and Scientific Computing

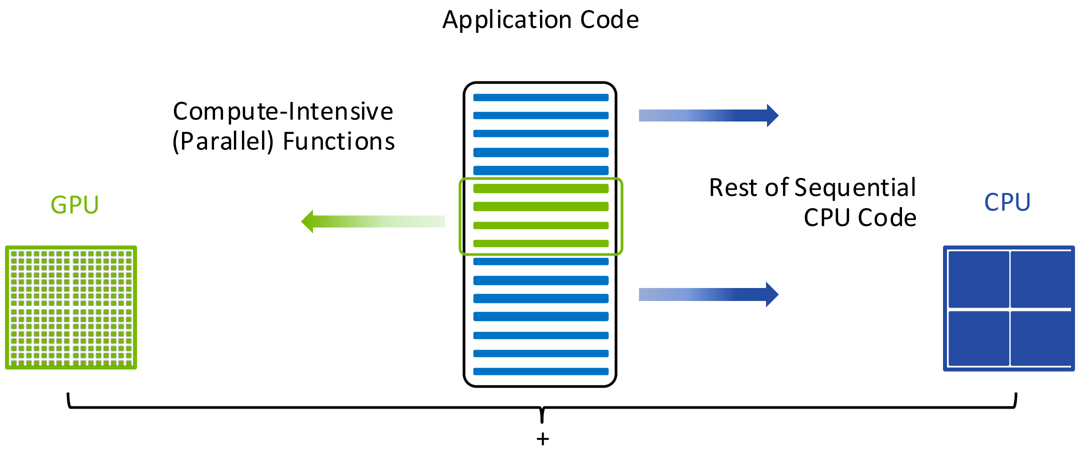
Accelerated Computing + AI Provides the Compute Required



Getting Million-X Speedups to Power AI and Scientific Computing

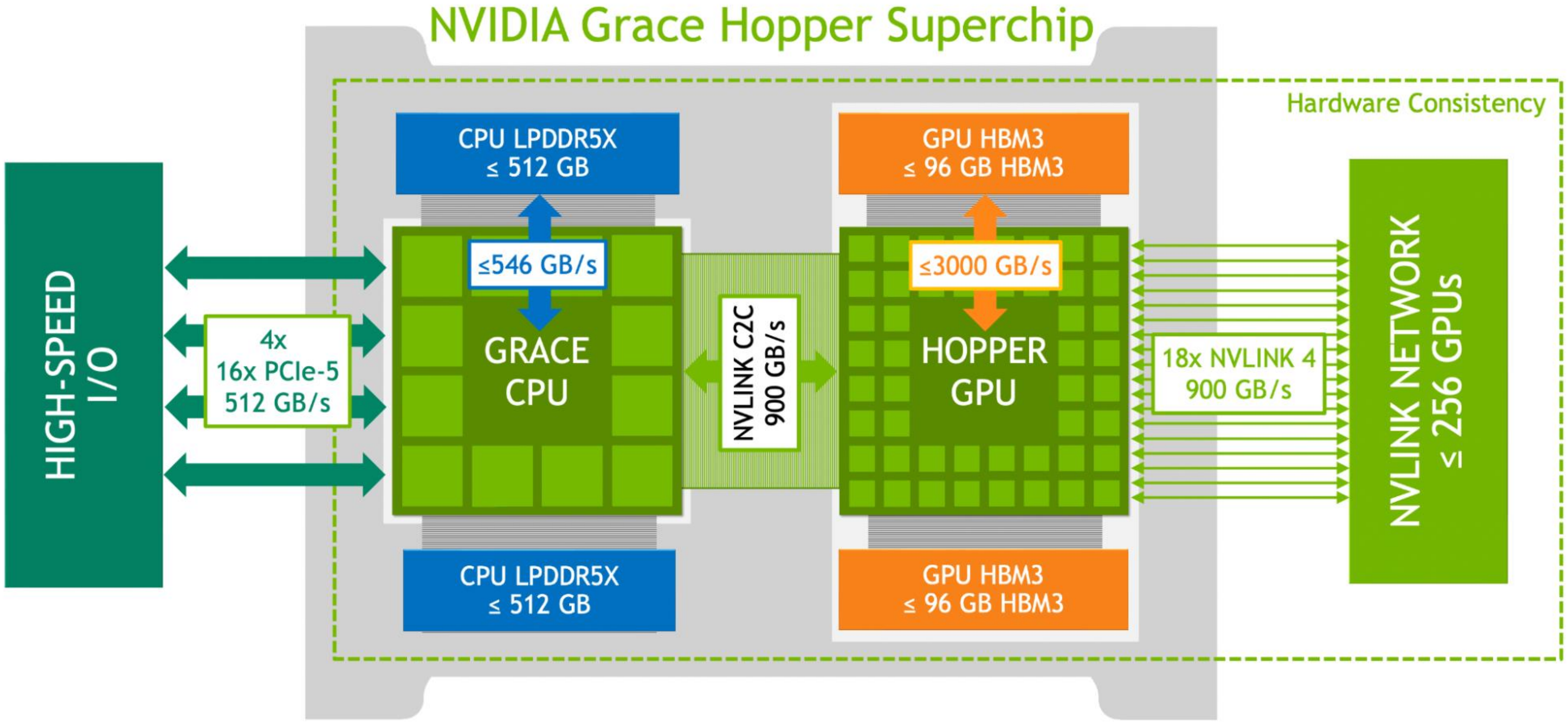
Accelerated Computing + AI Provides the Compute Required





What’s a GPU?

There are many types of GPUs! Let’s see an example of one



Choose the right GPU for your task

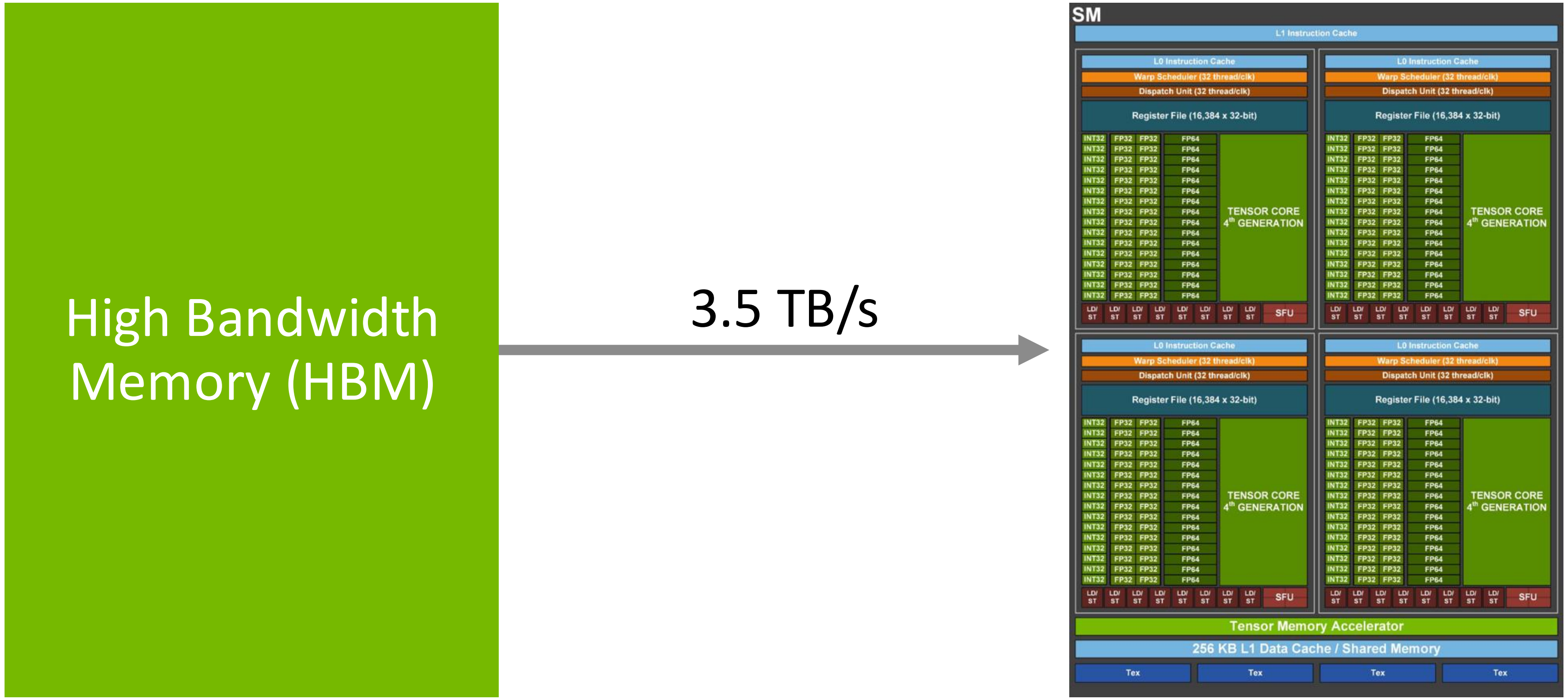
	GPU	DL Training & DA	DL Inference	HPC / AI	Omniverse / Render Farms	Virtual Workstation	Virtual Desktop (VDI)	Mainstream Acceleration	Far Edge Acceleration
Compute	H100	SXM PCIE NVL ●	SXM PCIE NVL ●	SXM PCIE NVL ●				PCIE NVL ●	
	A100	SXM PCIE ◐	SXM PCIE ◐	SXM PCIE ◐				PCIE ◐	
	A30		PCIE ○	PCIE ○				PCIE ○	
Graphics / Compute	L40		●		●	●		●	
	A40				◐	◐		◐	
	A10		○		○	●	◐	○	◐
	A16					○	●		
Small Form Factor Compute/Graphics	L4		●		◐	●	●	●	●
	A2		◐			◐	◐	◐	◐
	T4		○			○	○	○	○

○ ◐ ● **Price-performance** comparison in each product group (Compute, Graphics & Compute, SFF Compute & Graphics) and workload column

Moving data and computing

A multiprocessor spends time on two operations

A streaming multiprocessor (SM) of the NVIDIA H100, with four sub-cores



High Bandwidth Memory (HBM)

3.5 TB/s

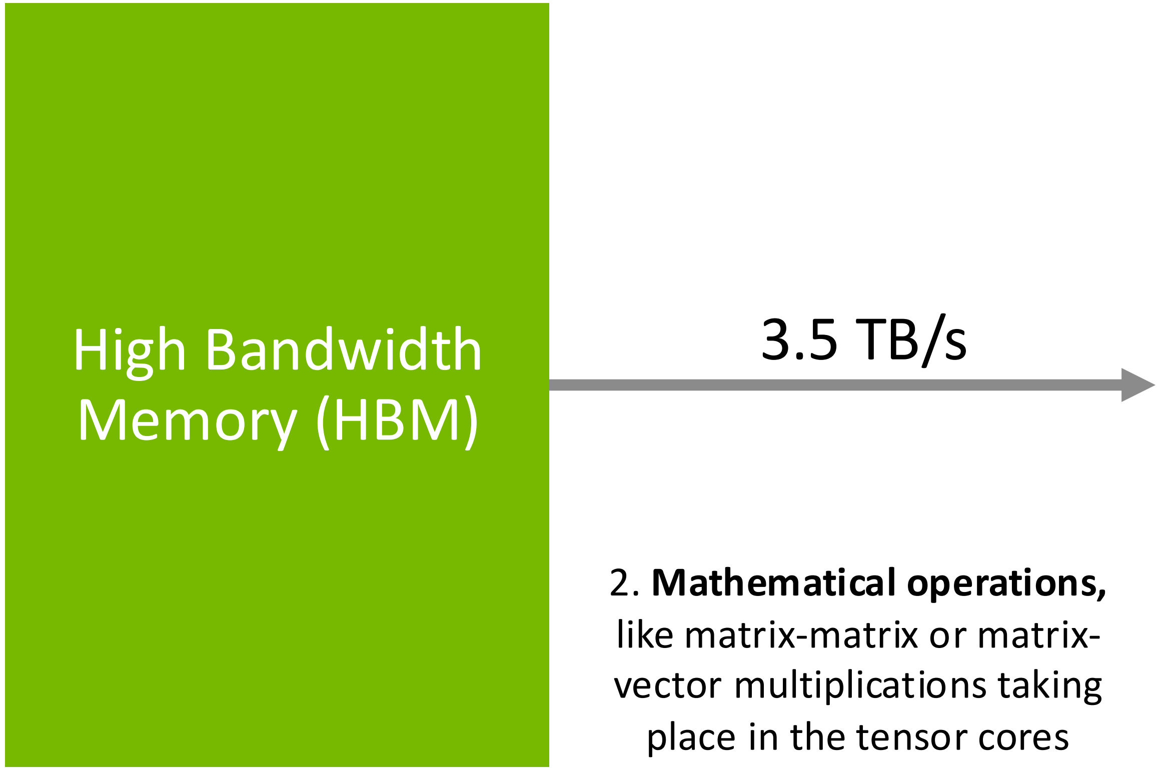
SM
L1 Instruction Cache
L0 Instruction Cache
Warp Scheduler (32 thread/clk)
Dispatch Unit (32 thread/clk)
Register File (16,384 x 32-bit)
INT32
FP32
FP64
TENSOR CORE 4th GENERATION
LD/ST
SFU
Tensor Memory Accelerator
256 KB L1 Data Cache / Shared Memory
Tex

1. **Loading data from GPU memory** to the computing unit’s SRAM and registers at a specified bandwidth

Inspired by https://linden-li.github.io/posts/inference-slides, image from https://resources.nvidia.com/en-us-tensor-core

Moving data and computing

A multiprocessor spends time on two operations



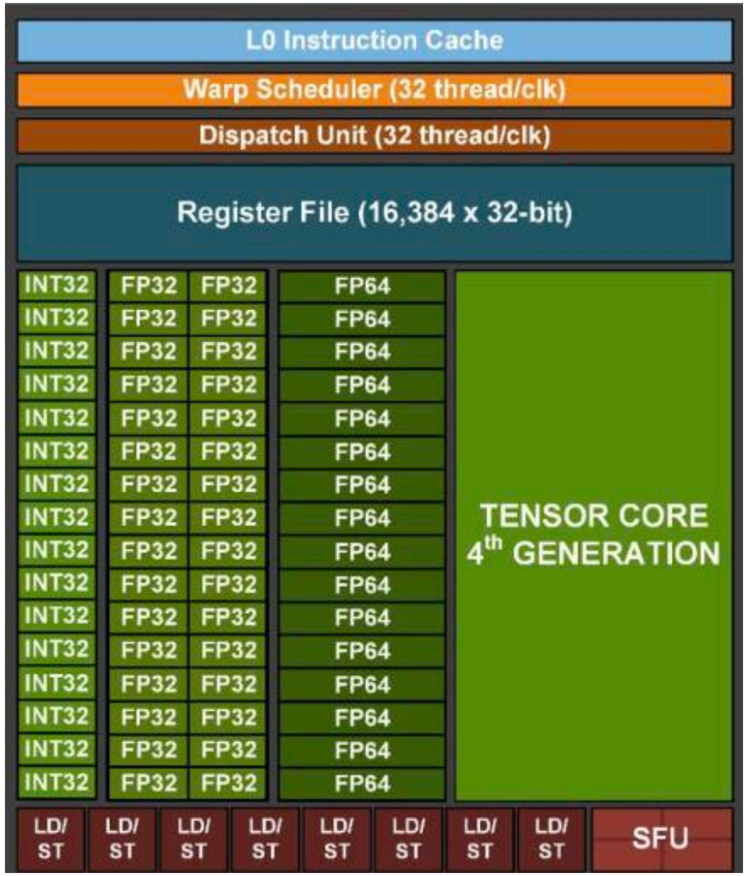
High Bandwidth Memory (HBM)

3.5 TB/s

2. **Mathematical operations,** like matrix-matrix or matrix-vector multiplications taking place in the tensor cores

e.g. 133.8 TFLOPS for FP16

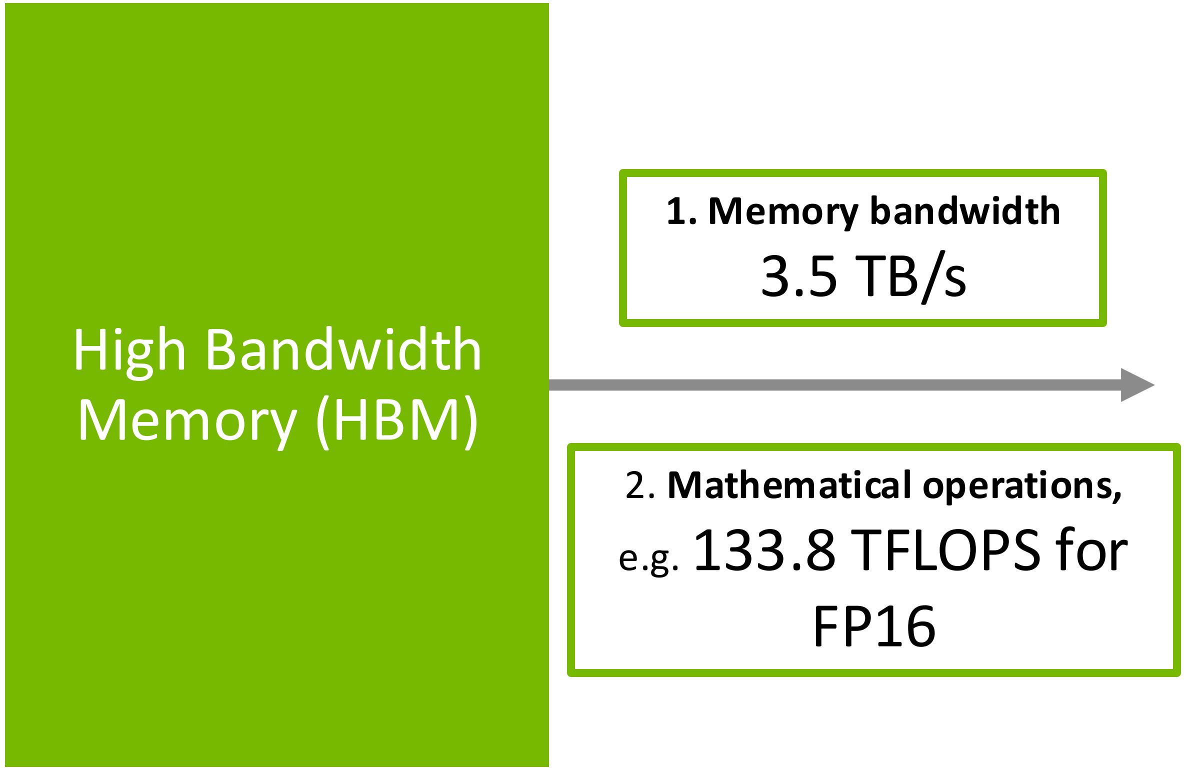
A tensor core of NVIDIA H100


L0 Instruction Cache
Warp Scheduler (32 thread/clk)
Dispatch Unit (32 thread/clk)
Register File (16,384 x 32-bit)
INT32
FP32
FP64
TENSOR CORE 4th GENERATION
LD/ ST
SFU

Inspired by https://linden-li.github.io/posts/inference-slides, image from https://resources.nvidia.com/en-us-tensor-core

Which of the two is faster?

Depends on the kernel under consideration

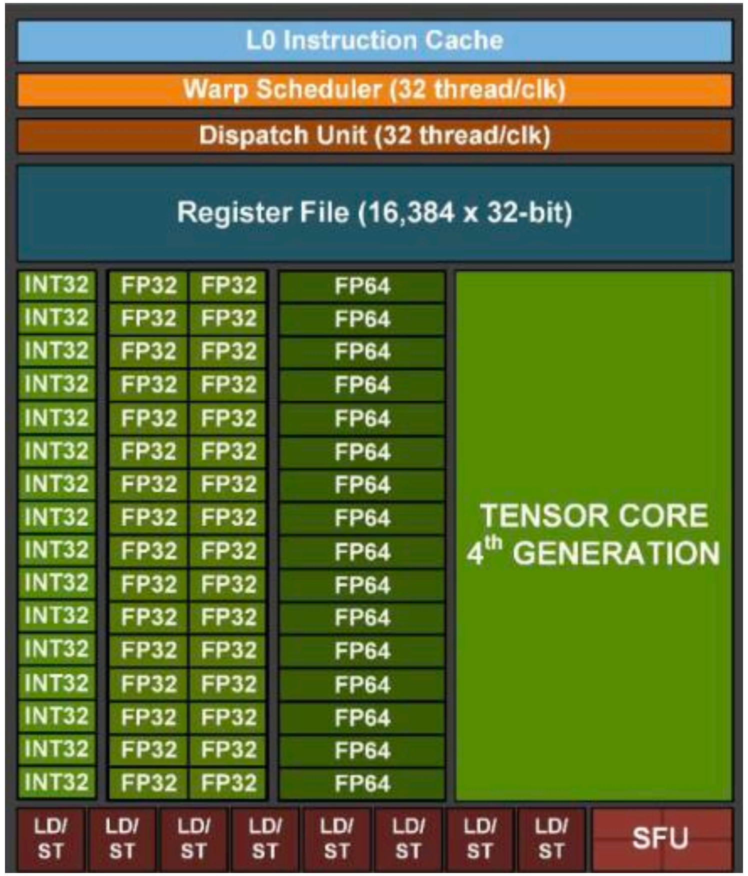


High Bandwidth Memory (HBM)

1. Memory bandwidth
3.5 TB/s

2. **Mathematical operations,**
e.g. 133.8 TFLOPS for FP16

A tensor core of NVIDIA H100


L0 Instruction Cache
Warp Scheduler (32 thread/clk)
Dispatch Unit (32 thread/clk)
Register File (16,384 x 32-bit)
INT32
FP32
FP64
TENSOR CORE 4th GENERATION
LD/ ST
SFU

Inspired by https://linden-li.github.io/posts/inference-slides, image from https://resources.nvidia.com/en-us-tensor-core

A job is memory bandwidth bound if the bandwidth cannot keep up with the computations — the cores are waiting idle

High Bandwidth Memory (HBM)

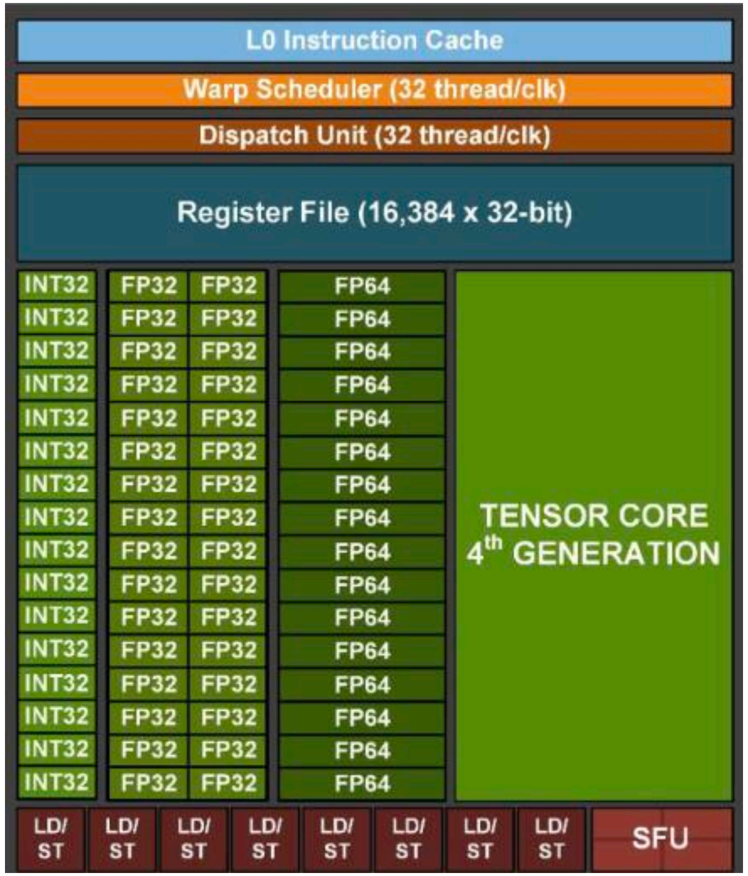
1. Memory bandwidth

3.5 TB/s

2. **Mathematical operations,**

e.g. 133.8 TFLOPS for FP16

A tensor core of NVIDIA H100



Inspired by https://linden-li.github.io/posts/inference-slides, image from https://resources.nvidia.com/en-us-tensor-core

A job is <u>compute bound</u> if the cores cannot keep up with the bandwidth — the bottleneck is in the FLOPS

High Bandwidth Memory (HBM)

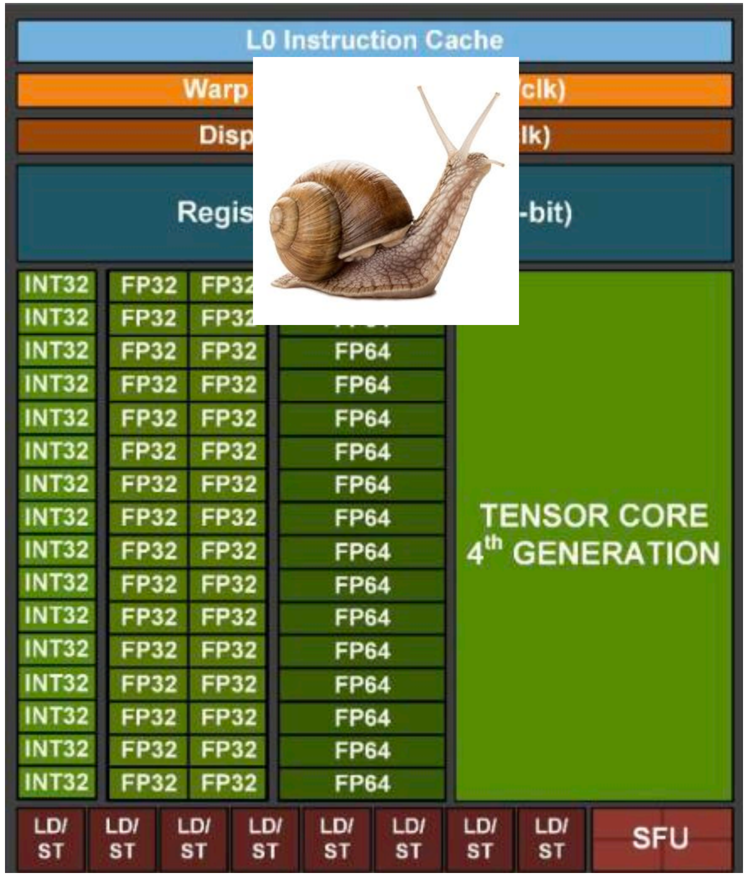
1. Memory bandwidth

3.5 TB/s

2. **Mathematical operations,**

e.g. 133.8 TFLOPS for FP16

A tensor core of NVIDIA H100



Inspired by https://linden-li.github.io/posts/inference-slides, image from https://resources.nvidia.com/en-us-tensor-core

Understanding if your job is memory or compute-bound

Roofline model

Arithmetic intensity of a kernel

$$\frac{\text{number of operations to compute a kernel}}{\text{bytes read from the DRAM memory}}$$

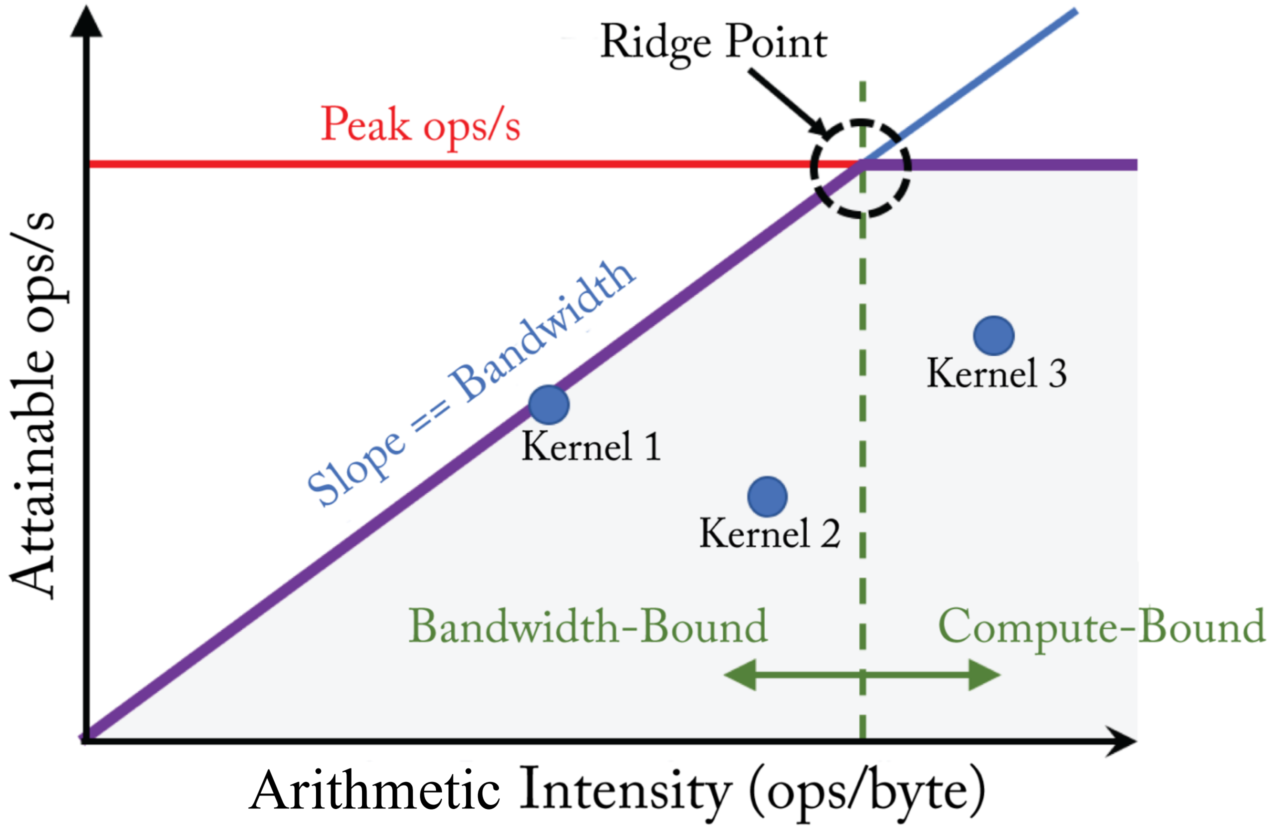


Image from book "Deep Learning Systems: Algorithms, Compilers, and Processors for Large-Scale Production"
"

Understanding if your job is memory or compute-bound

Roofline model

Arithmetic intensity of a kernel

$$\frac{\text{number of operations to compute a kernel}}{\text{bytes read from the DRAM memory}}$$

What can you say about kernel 1, 2 and 3?

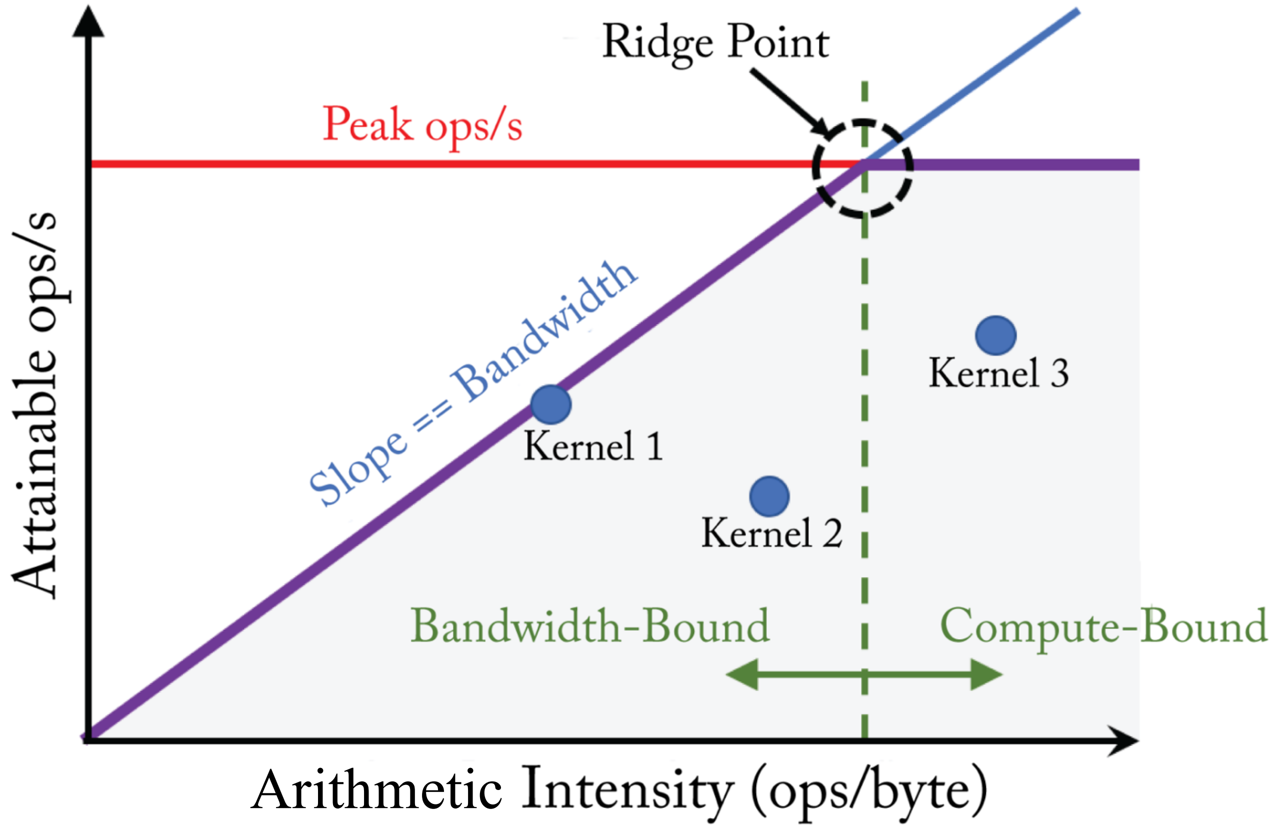
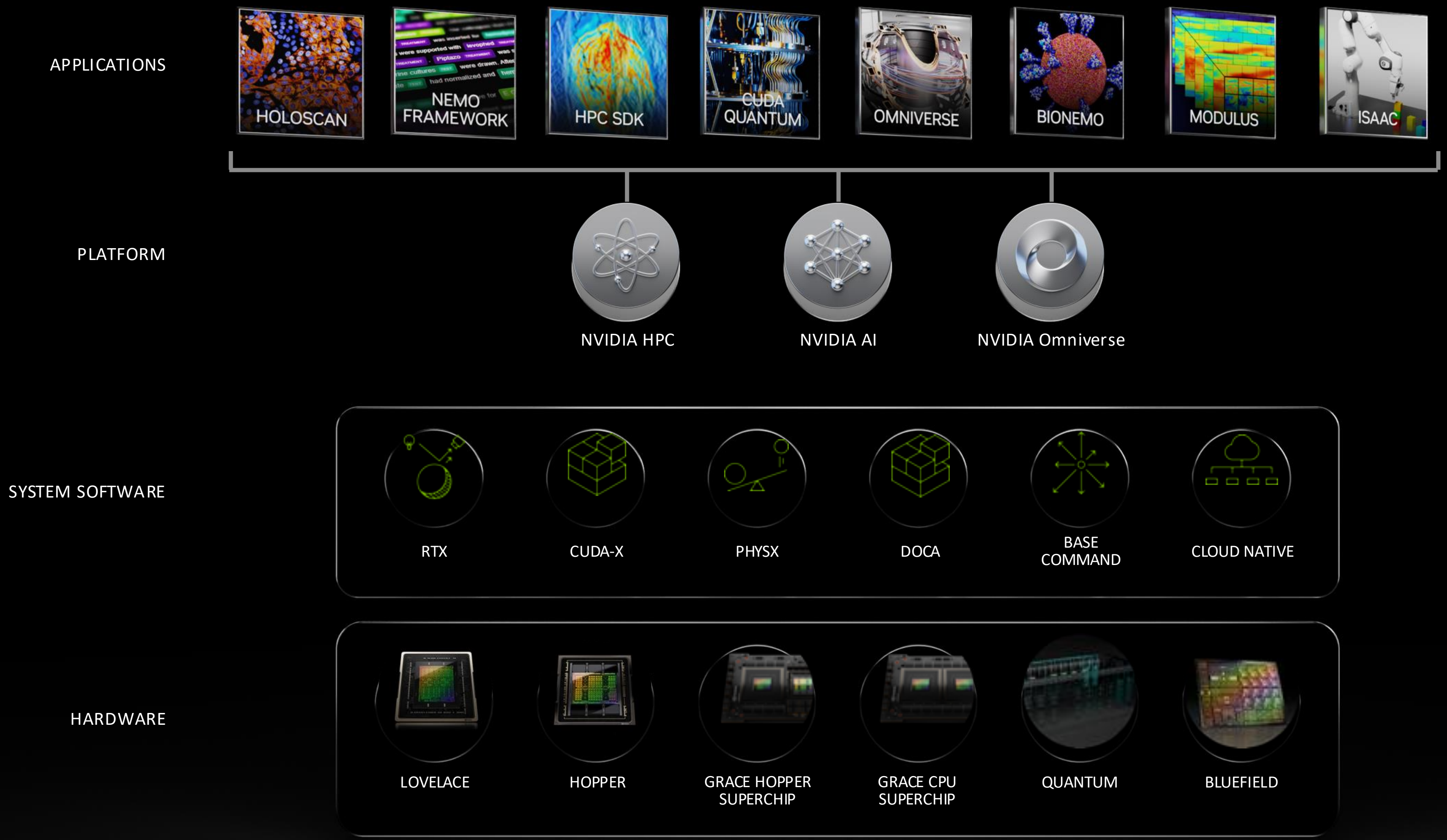


Image from book "Deep Learning Systems: Algorithms, Compilers, and Processors for Large-Scale Production"
"

Is NVIDIA just a hardware company?

NVIDIA Scientific Computing Platform



APPLICATIONS

HOLOSCAN

NEMO FRAMEWORK

HPC SDK

CUDA QUANTUM

OMNIVERSE

BIONEMO

MODULUS

ISAAC

PLATFORM

NVIDIA HPC

NVIDIA AI

NVIDIA Omniverse

SYSTEM SOFTWARE

RTX

CUDA-X

PHYSX

DOCA

BASE COMMAND

CLOUD NATIVE

HARDWARE

LOVELACE

HOPPER

GRACE HOPPER SUPERCHIP

GRACE CPU SUPERCHIP

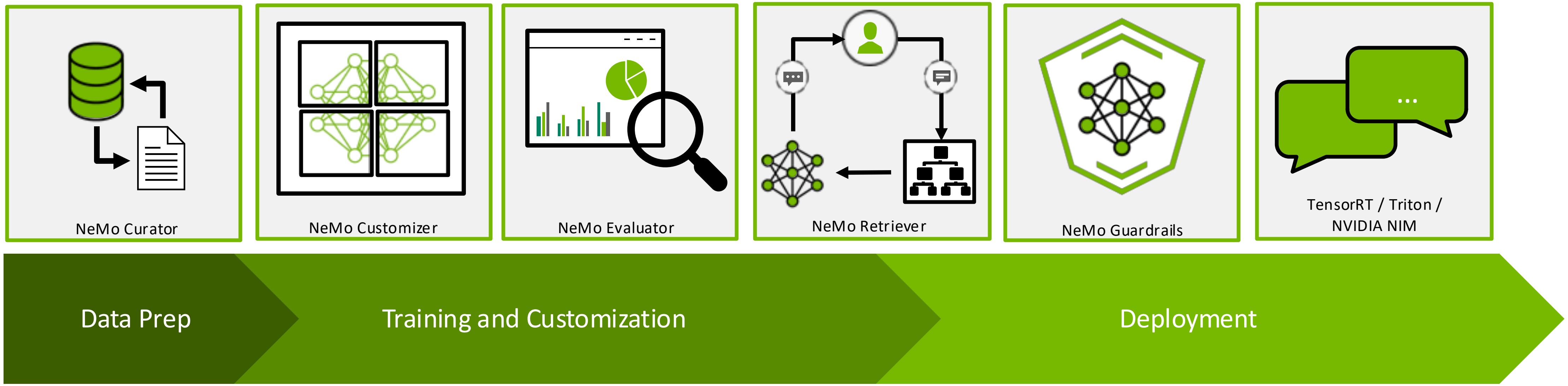
QUANTUM

BLUEFIELD

The focus for today:
Inference and model compression

The LLM cycle of life

Build, customize, and deploy generative AI models with NVIDIA NeMo



NeMo Curator

NeMo Customizer

NeMo Evaluator

NeMo Retriever

NeMo Guardrails

TensorRT / Triton /
NVIDIA NIM

Data Prep

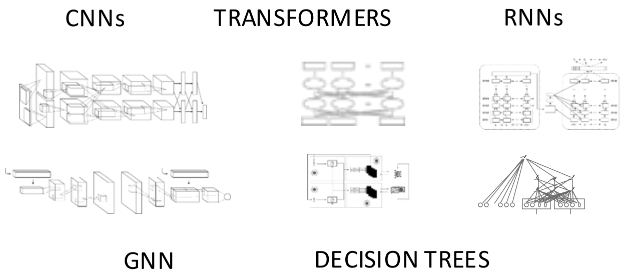
Training and Customization

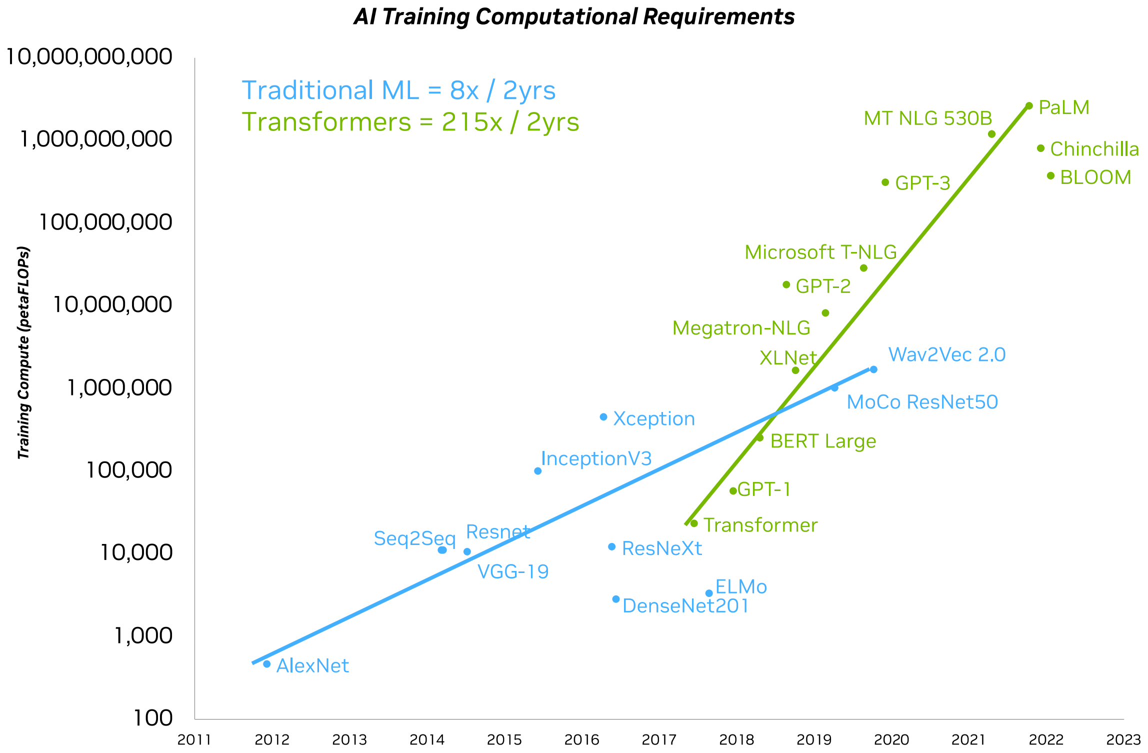
Deployment

NVIDIA Supports AI Model Landscape

Traditional and generative AI / LLM model evolution

- NVIDIA AI Inference Platform supports entire landscape of AI
  - Traditional models for Computer Vision, NLP, recommenders, speech AI
  - Latest LLM transformer models for Generative AI
  - Decade+ of NVIDIA software investment and libraries




AI Training Computational Requirements