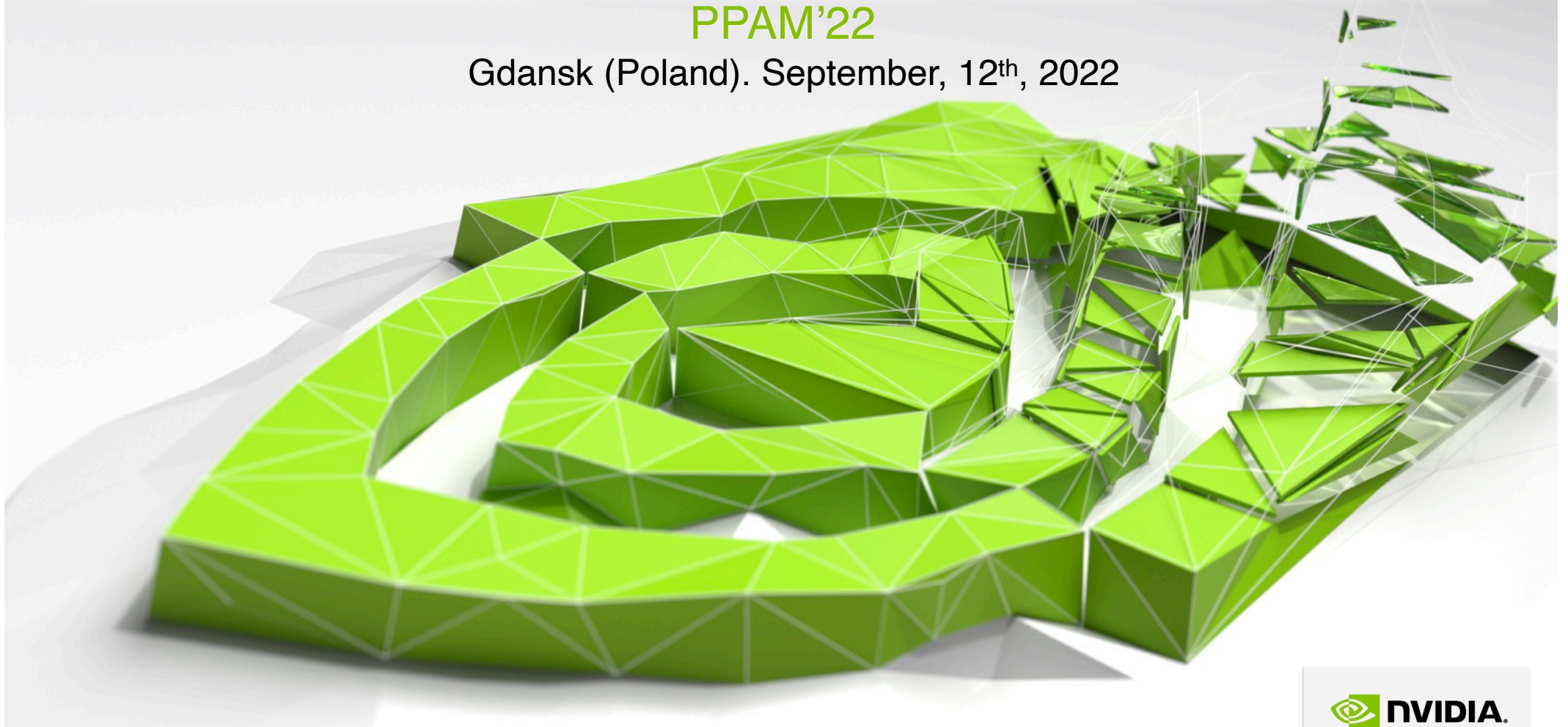


Nvidia Hopper Architecture

PPAM'22

Gdansk (Poland). September, 12th, 2022



Manuel Ujaldón

Full Professor in Computer Architecture @ University of Malaga
DLI Ambassador @ Nvidia Corporation



DEEP
LEARNING
INSTITUTE

UNIVERSITY
AMBASSADOR

Contents

- I. Introduction. [6 slides]
- II. Hardware design. [7]
- III. Major features. [6]
- IV. Performance, scalability, connectivity. [6]
- V. Products, market segments, roadmap. [12]
- VI. Nvidia AI Platform. [6]



I. Introduction

GTC

Keynote September 20 | Conference & Trainings September 19 - 22, 2022

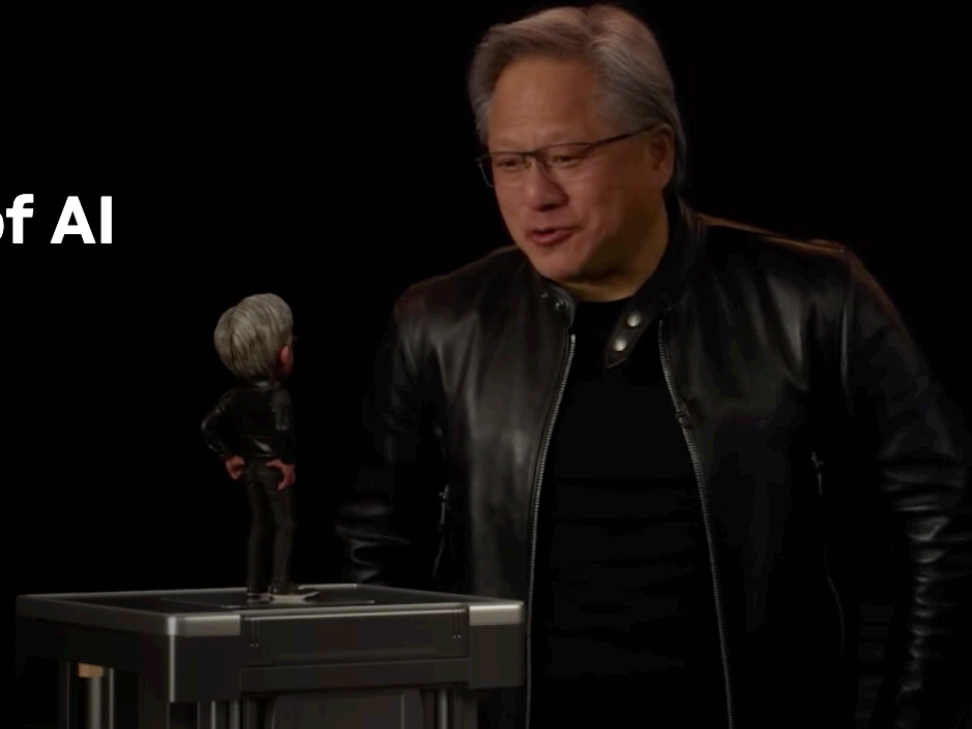
[Register Free](#) | [Log in](#)

[Keynote](#) [Why Attend](#) [Session Catalog](#) [Workshops & Training](#) [Sponsors](#) [Demos](#) [More](#)

The Developer Conference for the Era of AI and the Metaverse

September 19-22, 2022

[Register Free](#)

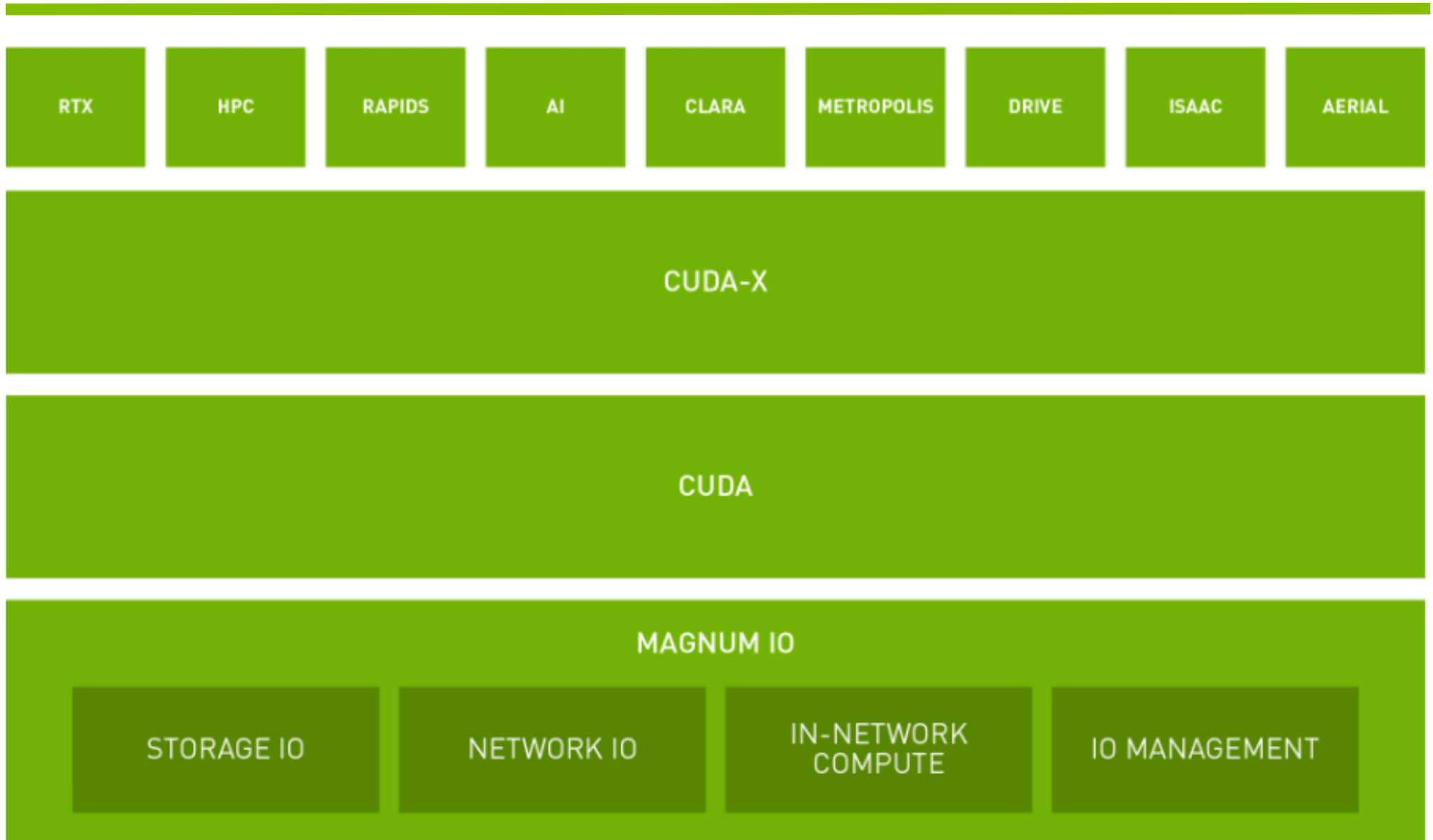


Explore the latest technologies and business breakthroughs.

Learn from experts how AI and the evolution of the 3D Internet are profoundly impacting industries—and society as a whole.

Join us for the online conference September 19-22, 2022 and be part of what comes next.

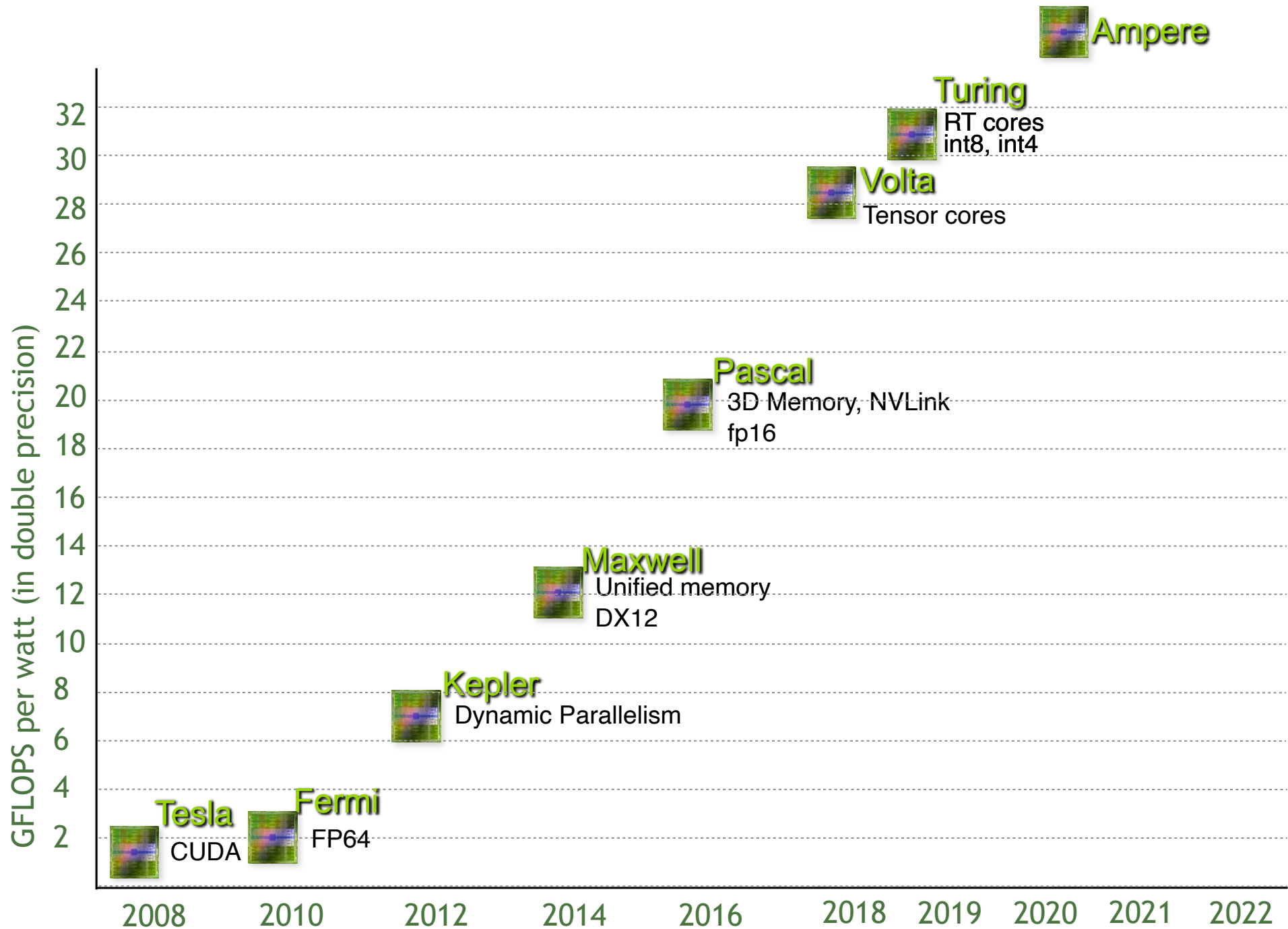
GPUs are everywhere



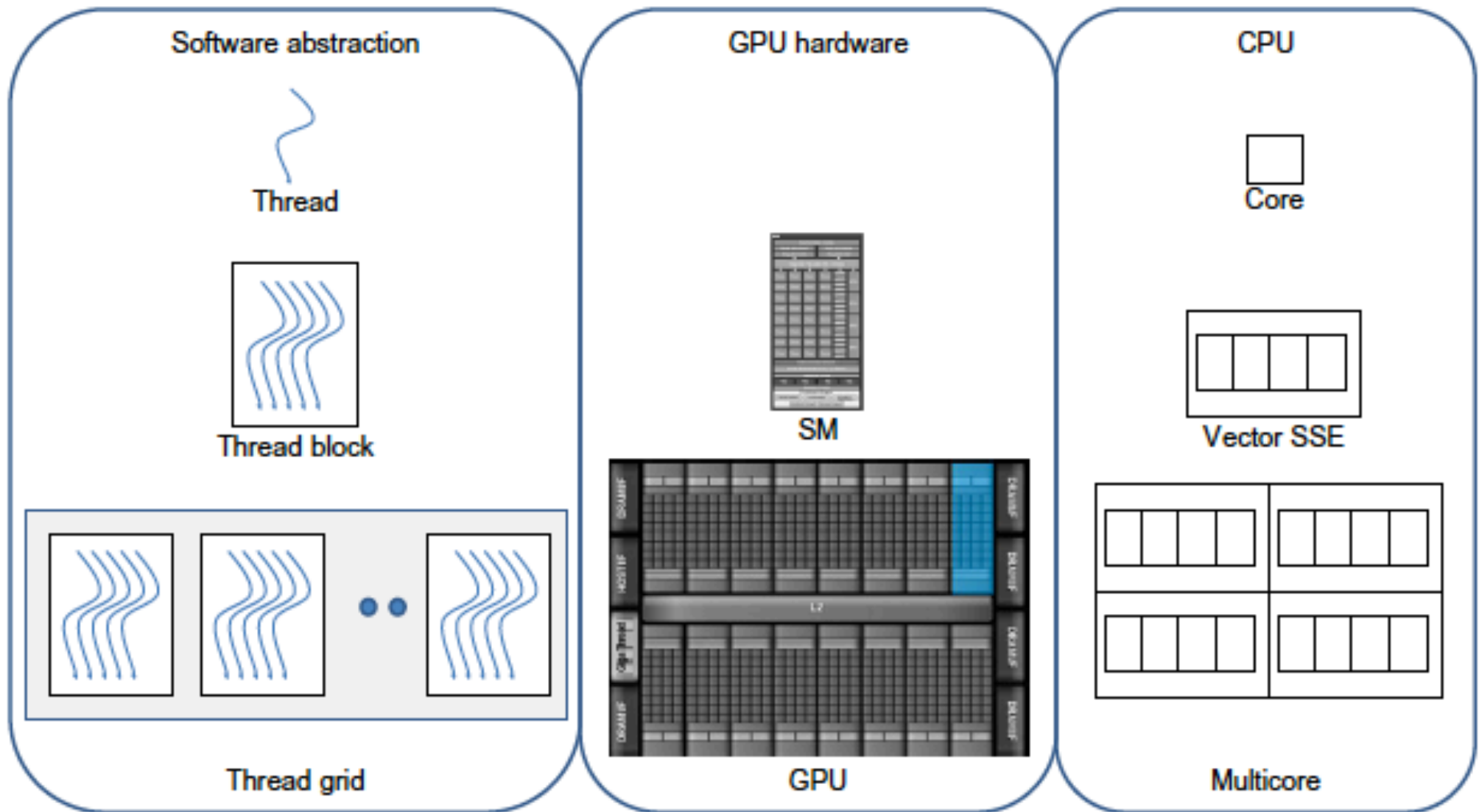
The top layer consists of domain-specific libraries

- RTX: Ray-tracing.
- HPC: High Performance Computing.
- RAPIDS: Data analytics.
- AI: Artificial Intelligence.
- CLARA: Health care and life sciences.
- METROPOLIS: Video analytics and streaming signal AI platform.
- DRIVE: Autonomous vehicles.
- ISAAC: Robotics.
- AERIAL 5G: 5G virtual ramp processing.

Overview of CUDA hardware generations Hopper



Comparing the GPU and the CPU: Two methods for building supercomputers



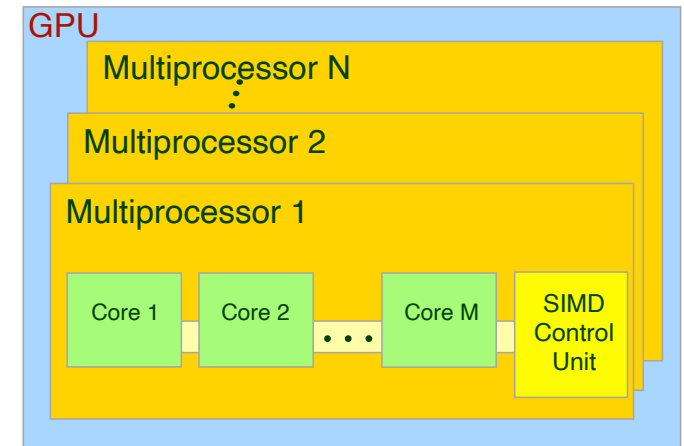
The CUDA hardware: SIMD processors structured, a tale of hardware scalability

- A GPU consists of:

- N multiprocessors (or SMs), each containing M cores (or stream processors).

- Heterogeneous computing:

- GPU: Data intensive. Fine-grain parallelism.
 - CPU: Control/management. Coarse grain parallelism.



	G80 (Tesla)	GF100 (Fermi)	GK110 (Kepler)	GM200 (Maxwell)	GP100 (Pascal)	GV100 (Volta)	TU102 (Turing)	A100 (Ampere)	H100 (Hopper)
Time frame	2006-09	2010-11	2012-13	2014-15	2016-17	2018-20	2019-20	2020-22	2022-?
N (multiprocessors)	16-30	14-16	13-15	4-24	56	80	72	108	132
M (fp32 cores/multip.)	8	32	192	128	64	64	64	64	128
# cores	128-240	448-512	2496-2880	512-3072	3584	5120	4608	6912	16896

The new generations (2016-2022)

	Pascal			Volta	Turing	Ampere		Hopper	
Architecture	GP104 (GTX1080)	GP100 (Titan X) (Tesla P100)	GP102 (Tesla P40)	GV100 (Tesla V100)	TU102 (Titan RTX)	A100	GA100	H100	GH100
Time frame	2016	2017	2017	2018	2019	2020	2020	2022	2022
CUDA Compute Capability	6.0	6.0	6.1	7.0	7.5	8.0	8.x	9.0	9.x
N (multiprocs.)	40	56	60	80	72	108	128	114	132
M (cores/multip.)	64	64	64	64	64	64	64	128	128
Number of cores	2.560	3.584	3.840	5.120	4.608	6.912	8.192	14.592	16.896



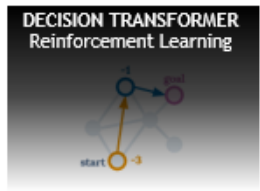
II. Hardware design

Next wave of AI requires performance and scalability

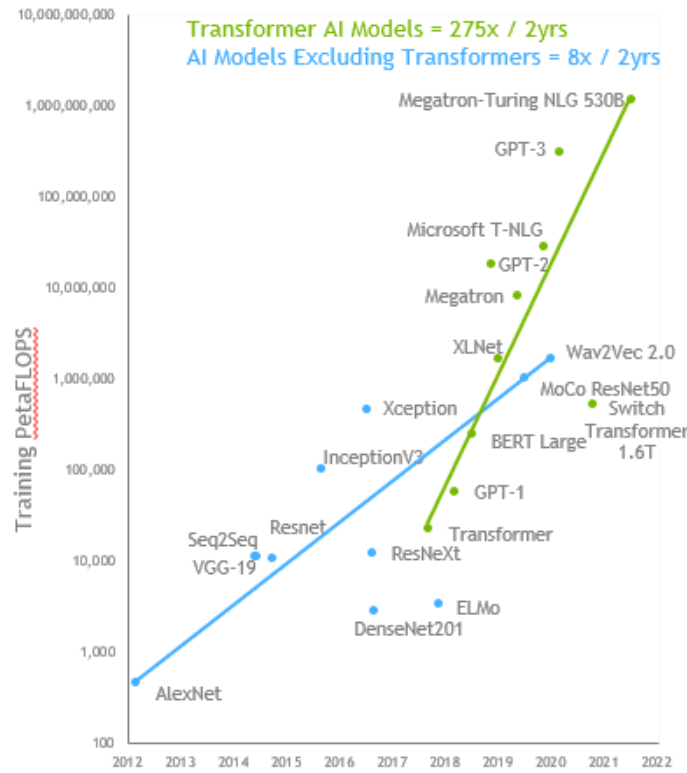
TRANSFORMERS TRANSFORMING AI

70%

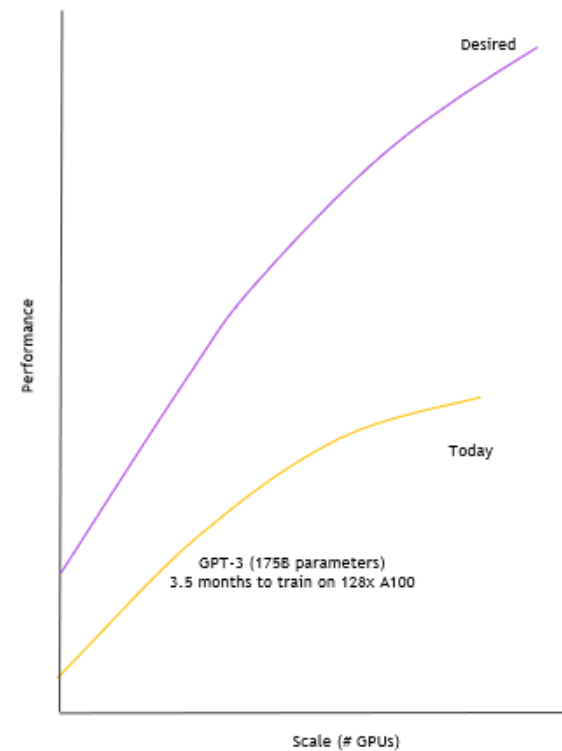
AI Papers
In last 2 years discuss Transformer Models



EXPLODING COMPUTATIONAL REQUIREMENTS



HIGHER PERFORMANCE AND SCALABILITY



MEGAMOLBART: <https://catalog.ngc.nvidia.com/orgs/nvidia/teams/clara/models/megamolbart>

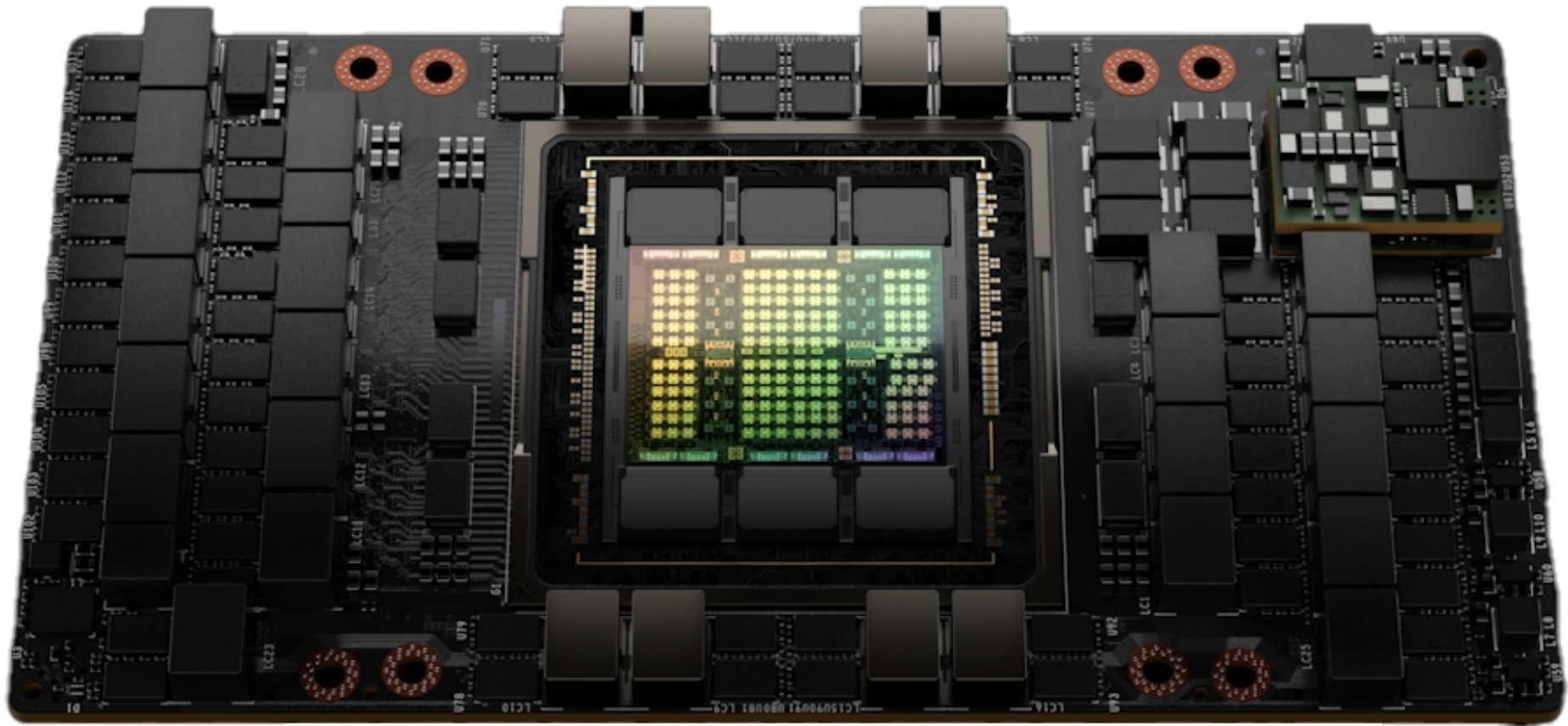
SegFormer: <https://arxiv.org/abs/2105.15203>

Decision Transformer: <https://arxiv.org/pdf/2106.01345.pdf>

SuperGLUE: <https://super.gluebenchmark.com/leaderboard>

Exploding Computational Requirements, source: NVIDIA Analysis and https://github.com/amirgholami/ai_and_memory_wall

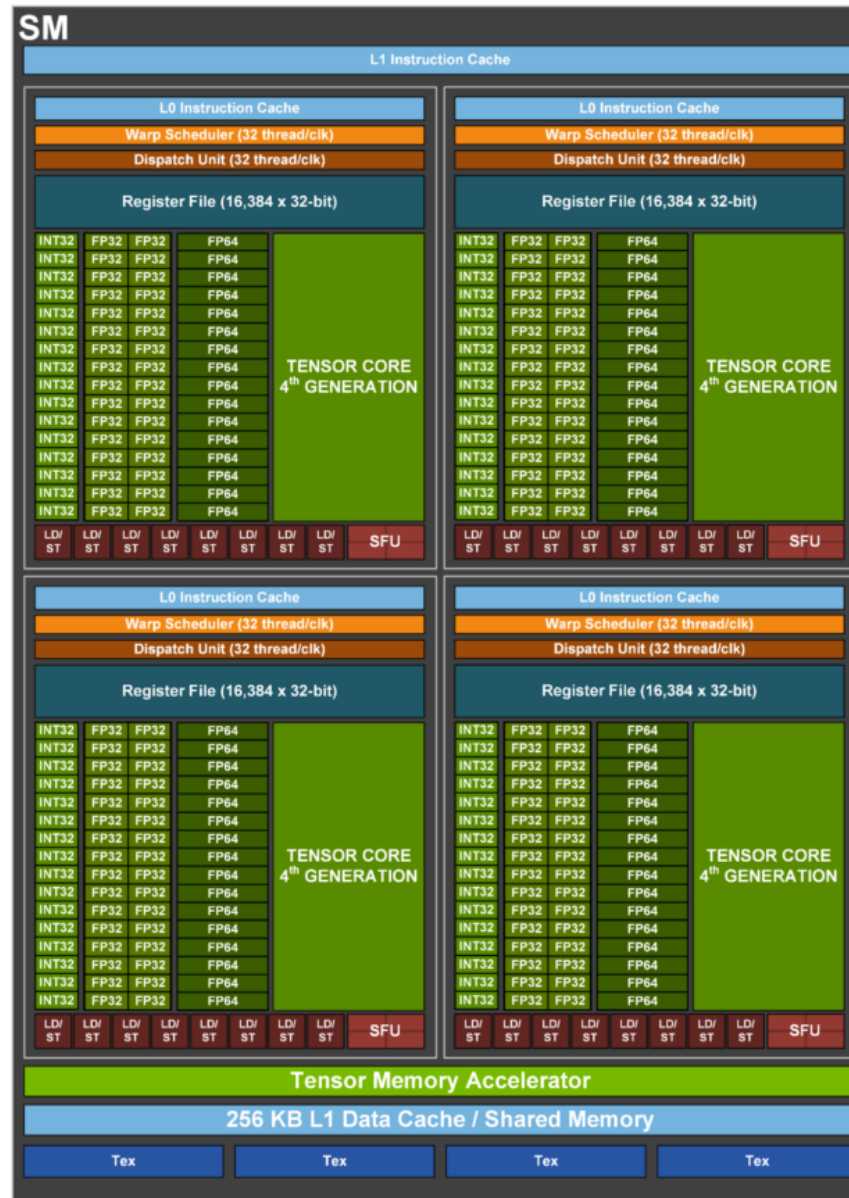
The printed circuit board for Hopper



The GH100 GPU with 144 SMs and 6 HBM3 stacks



GH100 streaming multiprocessor

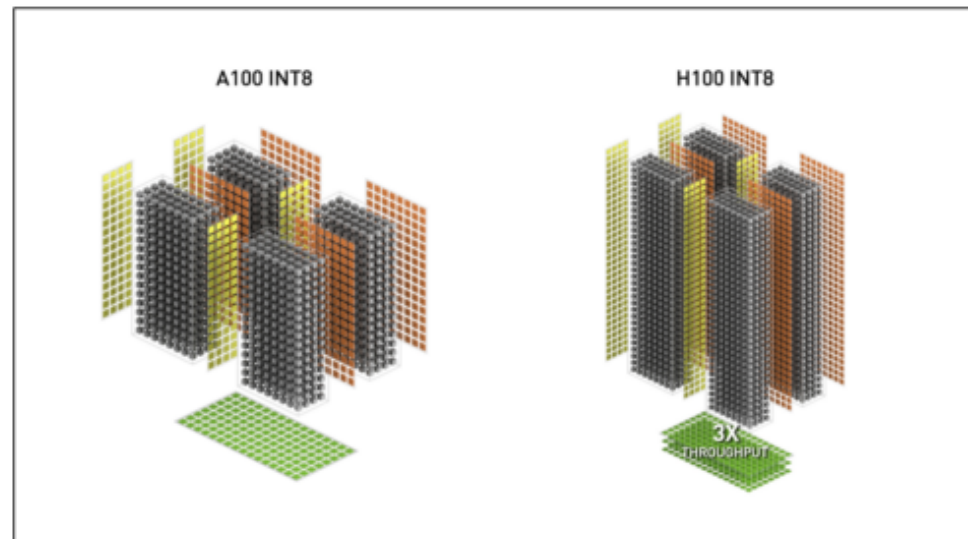
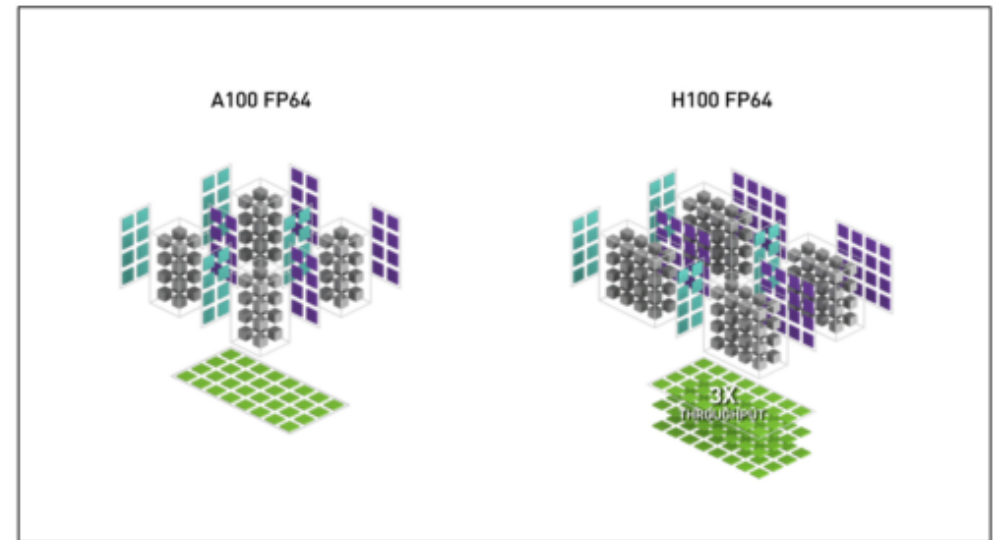
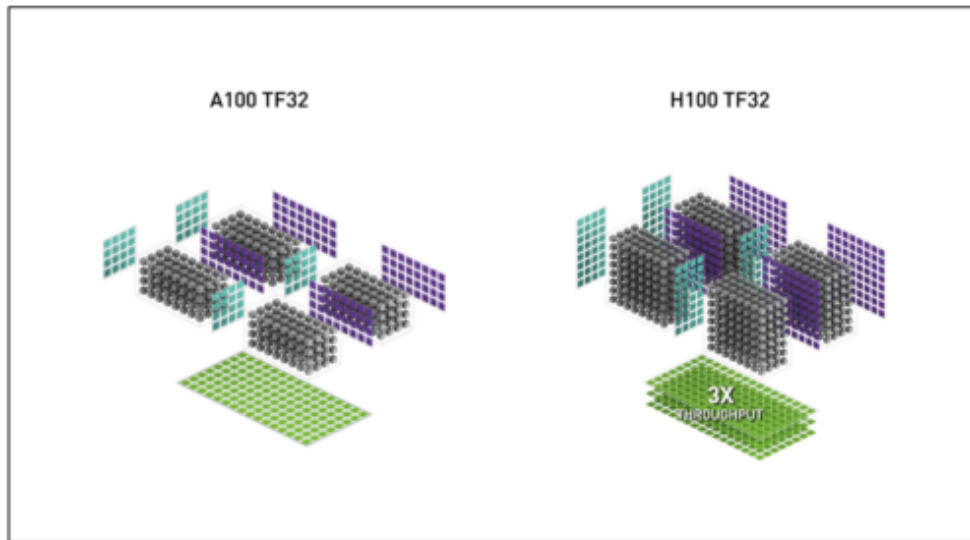


Computational and memory resources in last 3 flagship GPUs

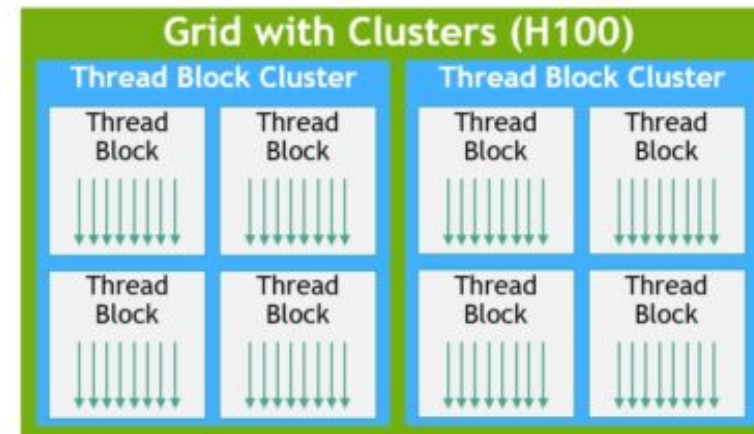
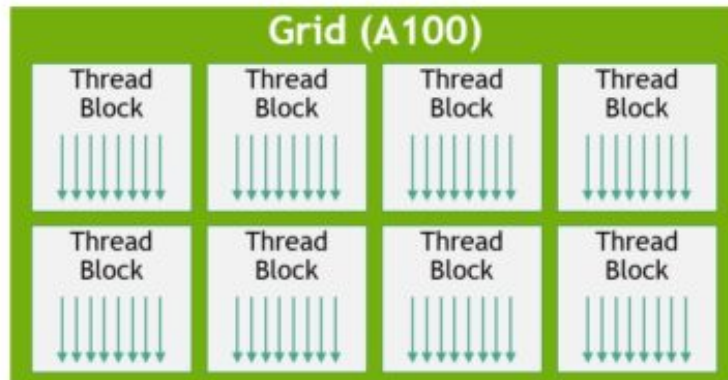
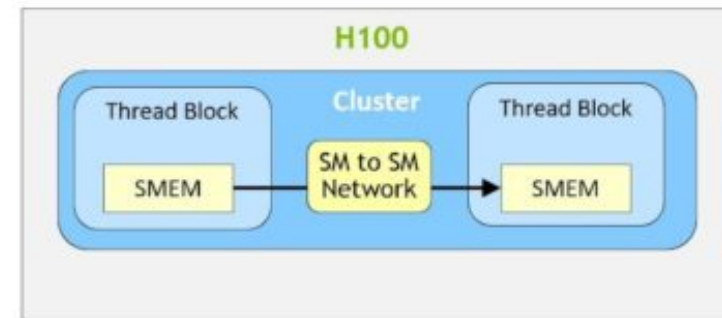
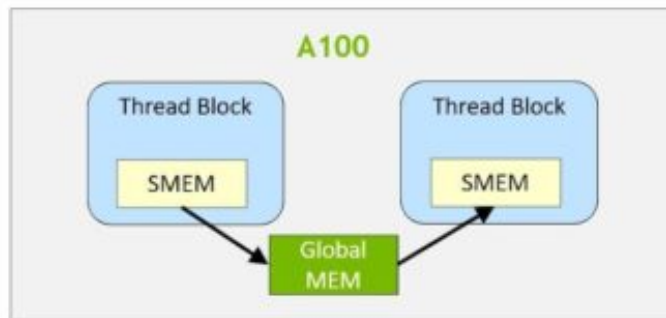
	Tesla V100 (Volta)	Titan RTX (Turing)	A100 (Ampere)	H100 * (Hopper)
GPU (chip)	GV100	TU102	GA100	GH100
fp32 cores	5120	4608	6912	16896
fp64 cores	2560	144	4096	8448
Frequency (base-boost)	1370-1455 MHz	1440-1770 MHz	1410 MHz	n/a
TFLOPS (fp16, fp32, fp64)	30, 15, 7.5	32.6, 16.3, 0.51	78, 19.5, 9.7	120, 60, 30
Memory interface	HBM2 4096 bits	GDDR6 384 bits	HBM2 5 stacks	HBM3 5 stacks
Memory bandwidth	900 GB/s.	672 GB/s.	1555 GB/s.	3000 GB/s.
Video memory	16 ó 32 GB	24 GB	48 GB	80 GB
L2 cache	6 MB	6 MB	40 MB	50 MB
Shared memory per multip.	Hasta 96 KB	Up to 64 KB	Up to 164 KB	Up to 228 KB.

(*) Preliminary specifications for H100 based on current expectations and are subject to change in the shipping product.

Matrix operations implemented in hardware



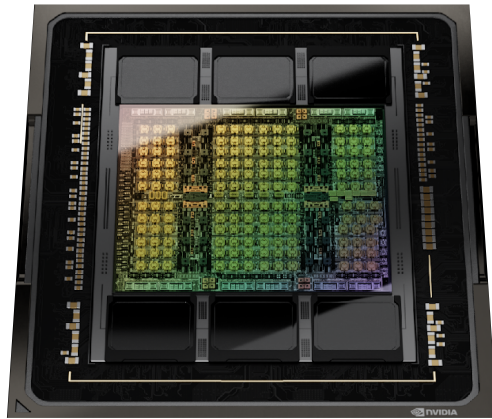
Thread block clusters: A new layer in the memory hierarchy



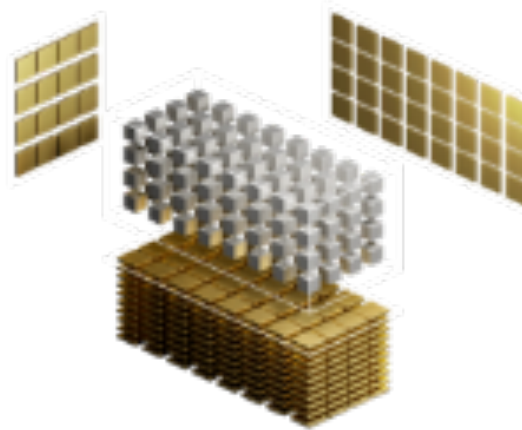


III. Major features

Hopper: The new engine for AI infrastructure. Performance, scalability & security



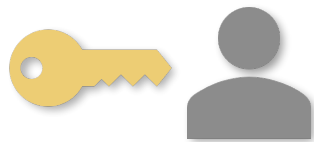
Custom 4N TSMC process
80 billion transistors



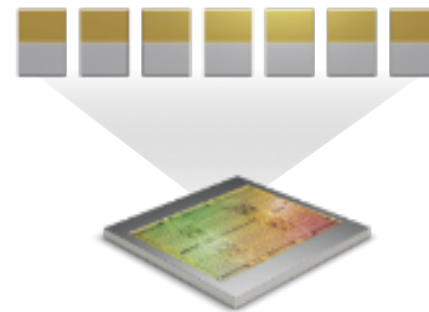
Transformer engine



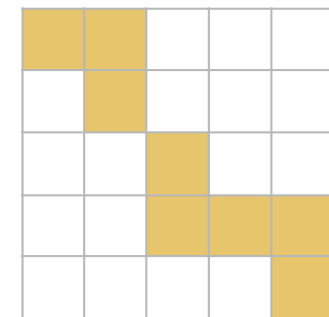
4th generation NVLink



Confidential computing



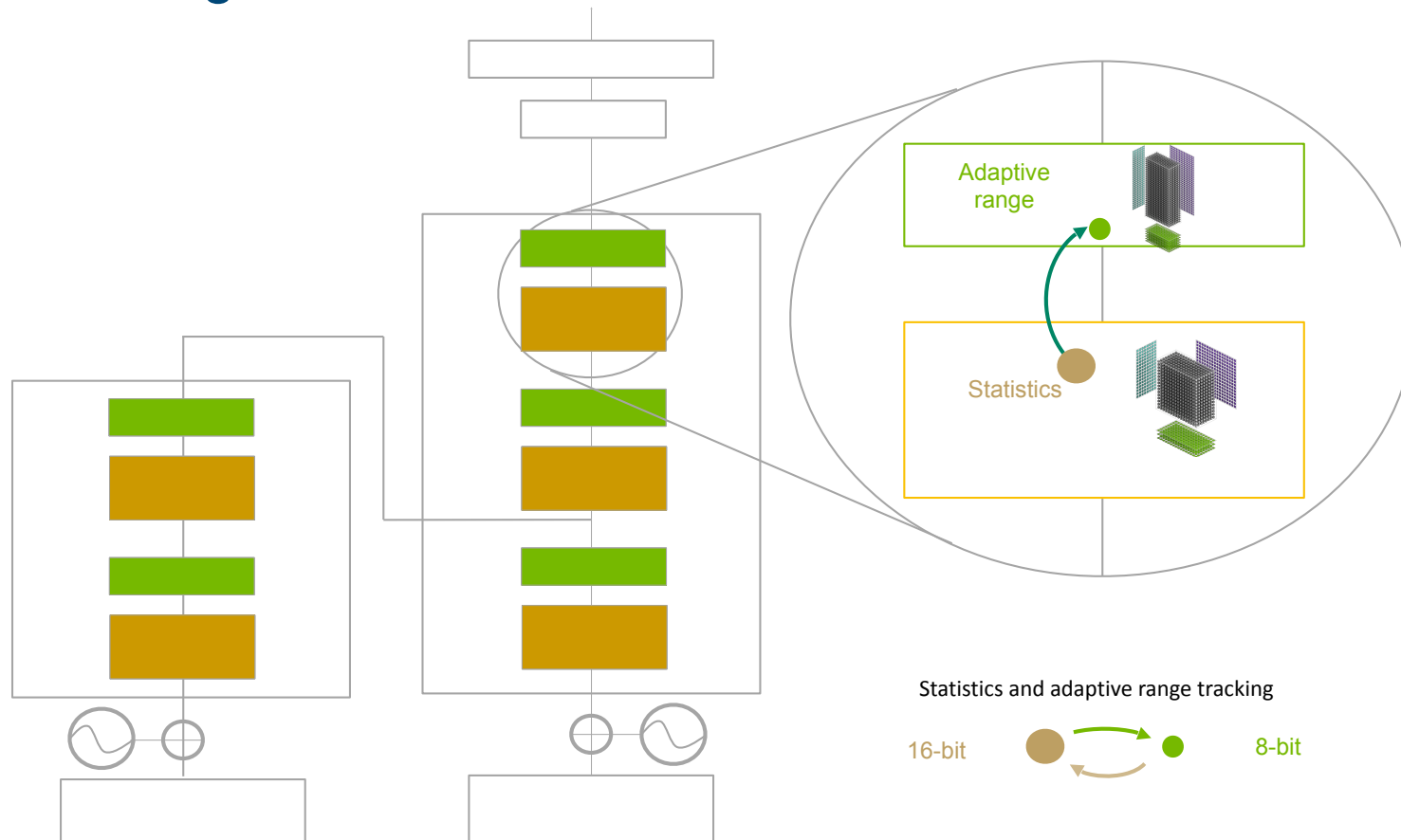
2nd generation MIG



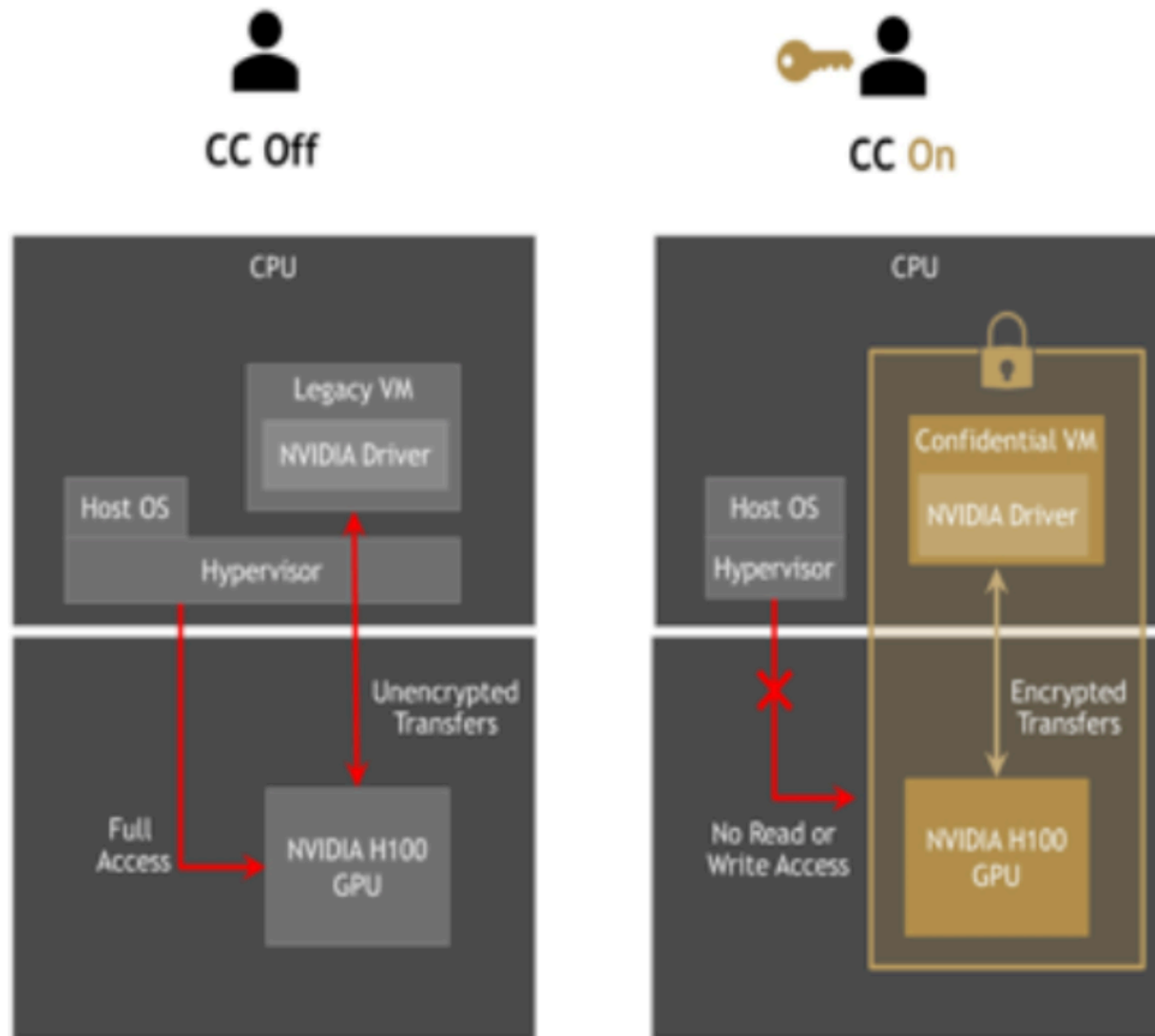
DPX instructions

Transformer engine: Tensor cores optimized for transformer models

- Nvidia tuned adaptive range optimization across 16-bit and 8-bit match.
- Configurable macro blocks deliver performance without accuracy loss.
- 6x faster training and inference of transformers models.



Confidential computing: Secure data and AI models in-use



Multi-GPU instance: 7 secure tenants within a single GPU

- 7 fully isolated and secured instances, QoS guaranteed.



DPX: New instructions for accelerating dynamic programming algorithms

A broad range of use cases



Optimization



Omics

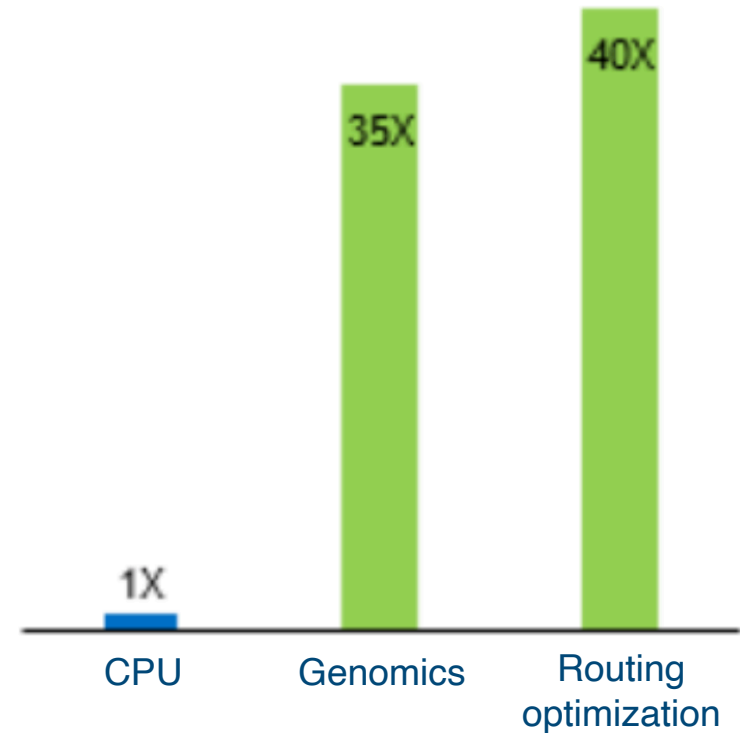


Graph analytics



Data processing

Real-time performance

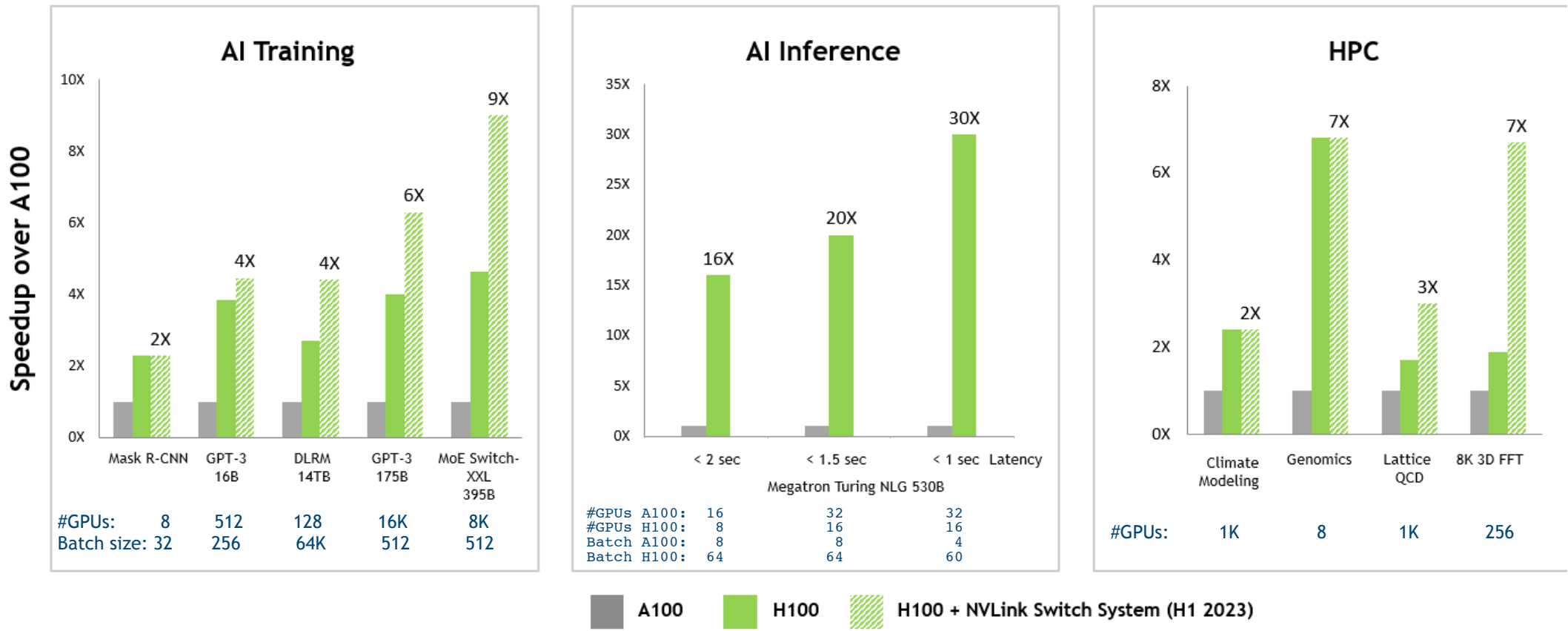




IV. Performance, scalability, connectivity

Substantial acceleration in all areas

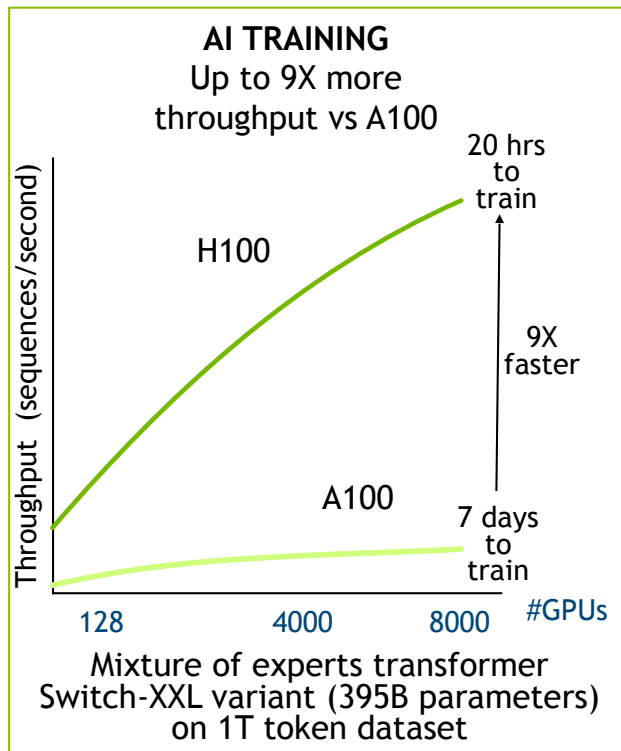
H100 speed-up vs. A100 on multiple GPUs:



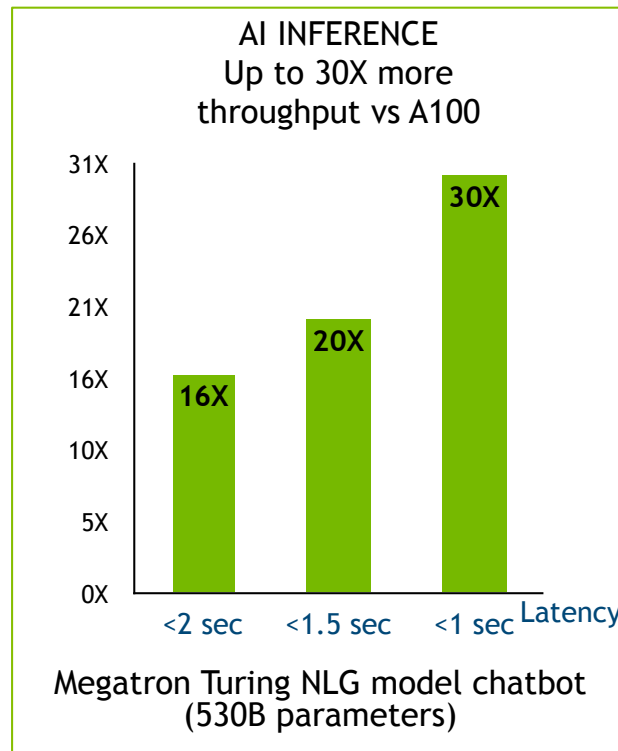
Projected performance subject to change

H100 brings order-of-magnitude leap in performance

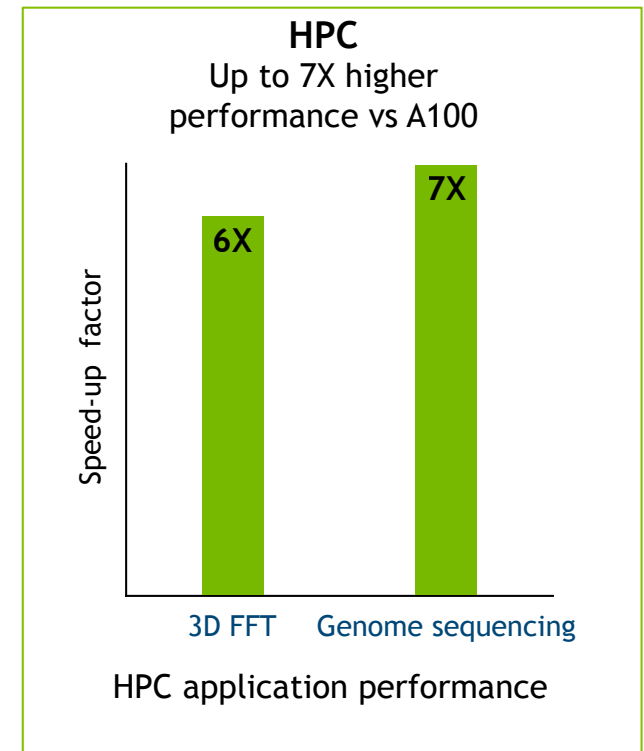
Performance and scalability for next-generation breakthroughs:



Projected performance subject to change

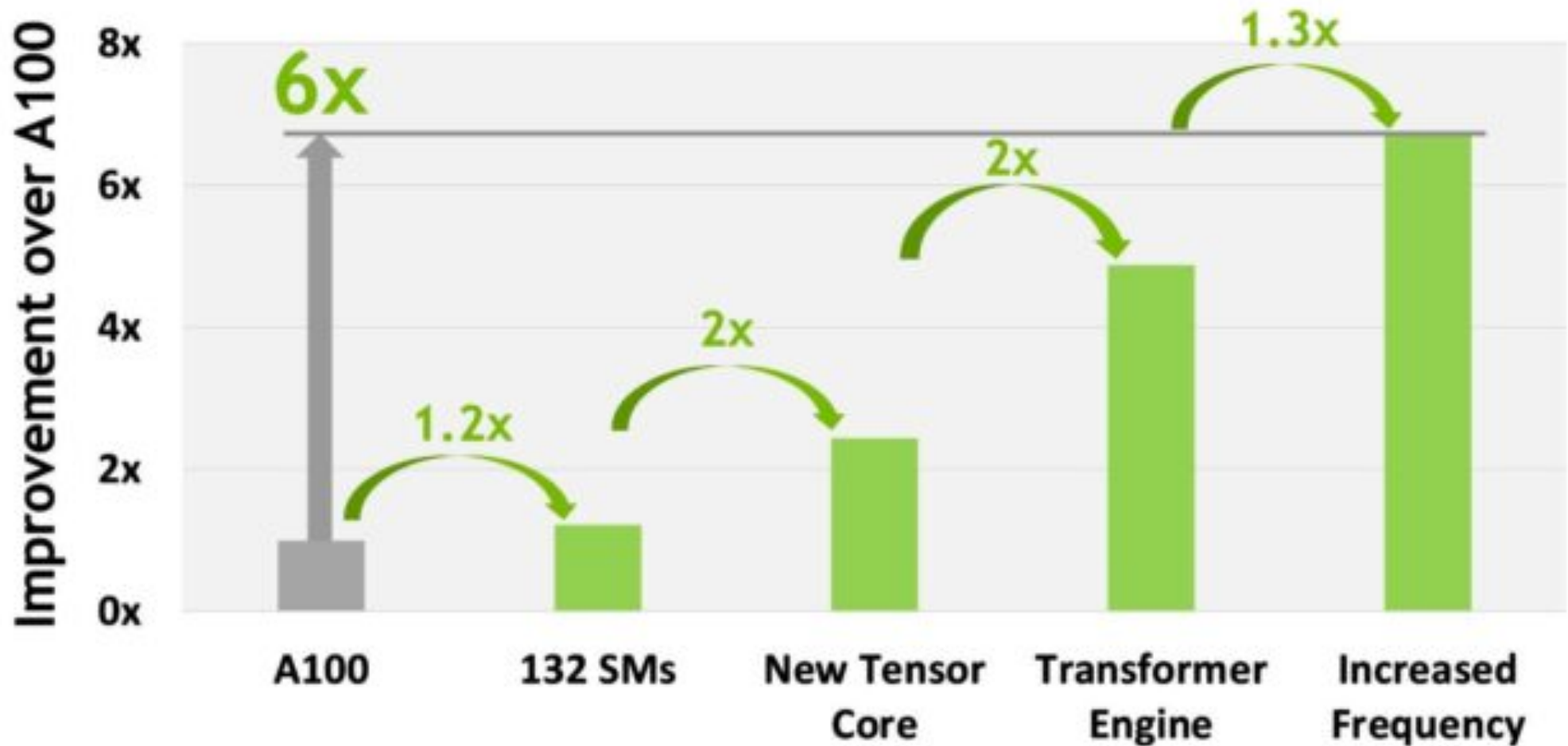


Input sequence length=128, output sequence length=20
 A100 cluster: HDR IB network
 H100 cluster: NDR IB network for 16 H100 configuration
 16 A100 vs 8 H100 for 2 sec
 32 A100 vs 16 H100 for 1 and 1.5 sec

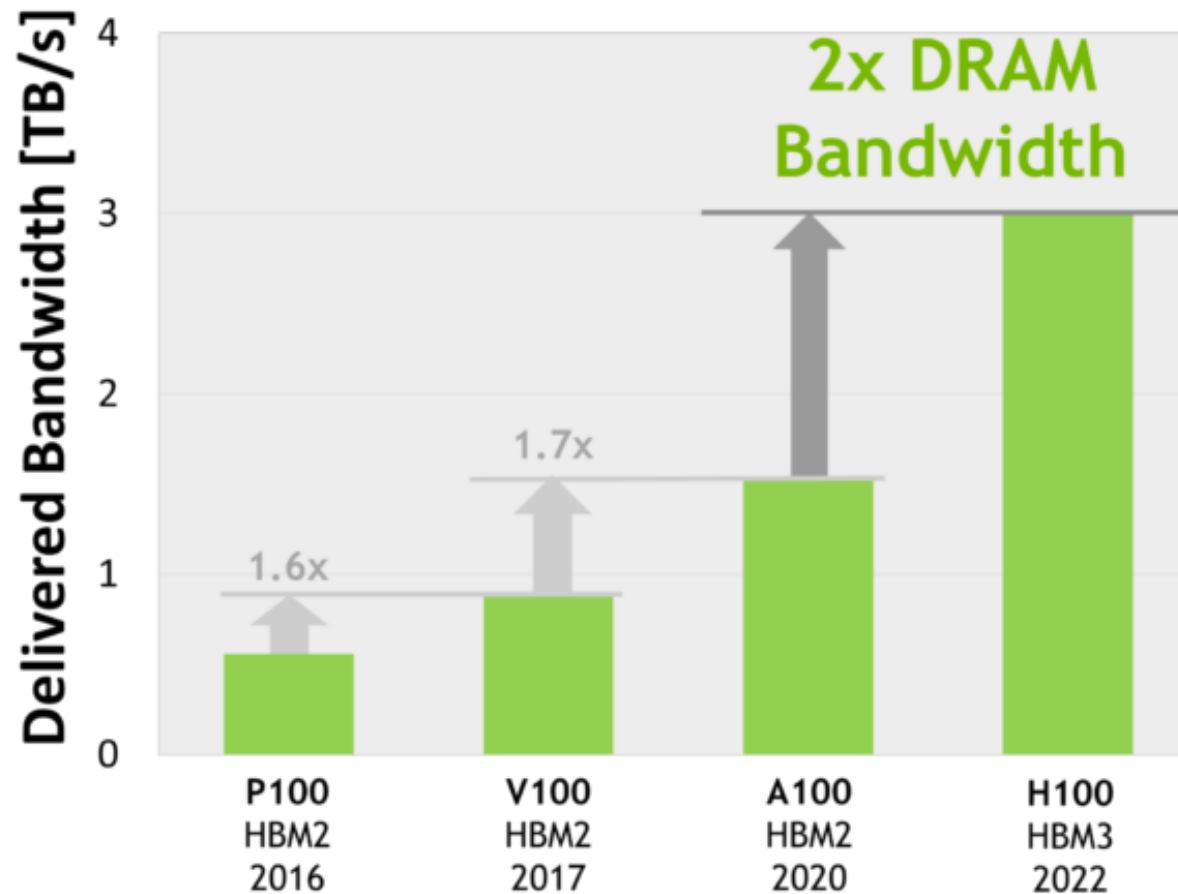


3D FFT (4K^3):
 - A100 cluster: HDR IB network
 - H100 cluster: NVLink Switch System, NDR IB
 Genome sequencing (Smith-Waterman):
 - 1 A100
 - 1 H100

Speed-up breakout vs. A100



Memory bandwidth improvement since the adoption of High Bandwidth Memory



Unprecedented AI and HPC performance, scalability and connectivity

Peak performance:

Data type	FLOPS (NVLink version)	FLOPS (PCI-e version)	Speed-up vs. A100 (including sparsity)	
			NVLink	PCI-e
fp8	4 Peta-	3.2 Peta-	6x	5x
fp16 (half)	2 Peta-	1.6 Peta-	3x	2.5x
fp32 (float)	1 Peta-	0.8 Peta-	3x	2.5x
fp64 (double)	60 Tera-	48 Tera-	3x	2.5x

	HBM3 memory (NVLink)	HBM2e mem. (PCI-e)
Size	80 Gbytes	80 Gbytes
Bandwidth	3 TB/s. (1.5x vs. A100)	2 TB/s.

Scalability:

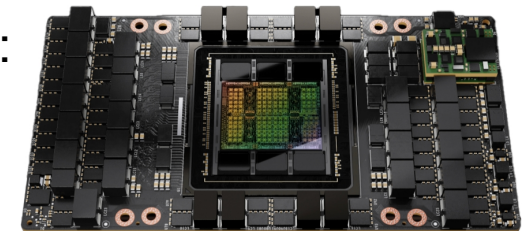
- NVLink Switch: Up to 256 GPUs (from NVSwitch@DGX).
- NVLink Bridge: Up to 2 GPUs (for PCI-e).

Connectivity:

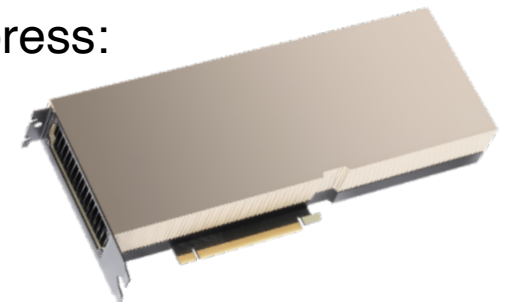
- GPU to GPU:
 - 900 GB/s (4th gen. NVLink).
 - 600 GB/s (5th gen. PCI-e).
- GPU to CPU:
 - 128 GB/s (5th gen. PCI-e).

Form factors:

- NVLink:



- PCI-express:



NVLink switch system

- High perf. 4th gener. NVLink network for up to 256 GPUs.



4th GEN NVLink

- 900 GB/s from 18 bi-directional ports @ 25 GB/s. each.
- GPU-2-GPU connectivity across nodes.

3rd GEN NVSwitch (from DGXs)

- All-to-all NVLink switching for 8-256 GPUs.
- Accelerate collectives - multicast and SHARP.

NVLink Switch

- 128 port cross-connect based on NVSwitch.

Representative hardware: H100 cluster (1 scalable unit)

- Servers: 32.
- NVLink switches: 18.
- NVLink optical cables: 1152.
- All-to-all bandwidth: 57.6 TB/s.



V. Products, market segments, roadmap

HGX-H100

HIGHEST PERFORMANCE FOR AI AND HPC

4-way / 8-way H100 GPUs in-network
 SHARP compute with sparsity:
 32 PetaFLOPS (FP8)
 3.6 TFLOPS (FP16)

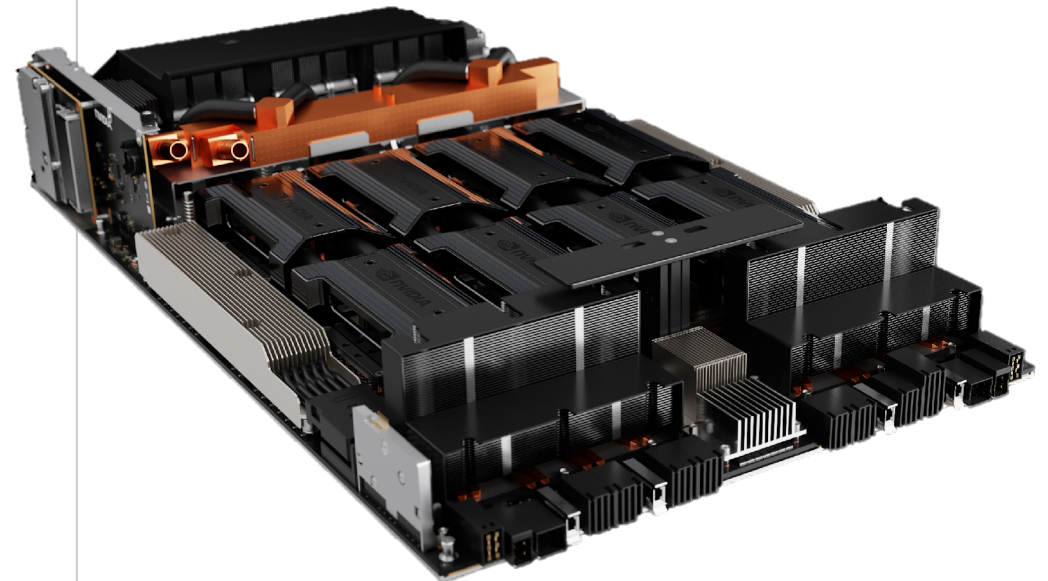
NVIDIA Certified HPC Offering from All Makers

FASTEST, SCALABLE INTERCONNECT

4th Gen NVLINK with 3X faster All-Reduce
 communications vs. previous generation.

3.6 TB/s bisection bandwidth

NVLINK Switch System Option Scales Up to 256 GPUs



SECURE COMPUTING

First HGX System with Confidential Computing



H100 PCI-express

HIGHEST AI AND HPC MAINSTREAM PERFORMANCE

3.2PF fp8 (5x)

1.6PF fp16 (2.5x)

800TF TF32 (2.5x)

48TF fp64 (2.5x)

(x-factors vs. A100 and including sparsity)

2TB/s , 80GB HBM2e memory

HIGHEST COMPUTE ENERGY EFFICIENCY

Configurable TDP - 150W to 350W

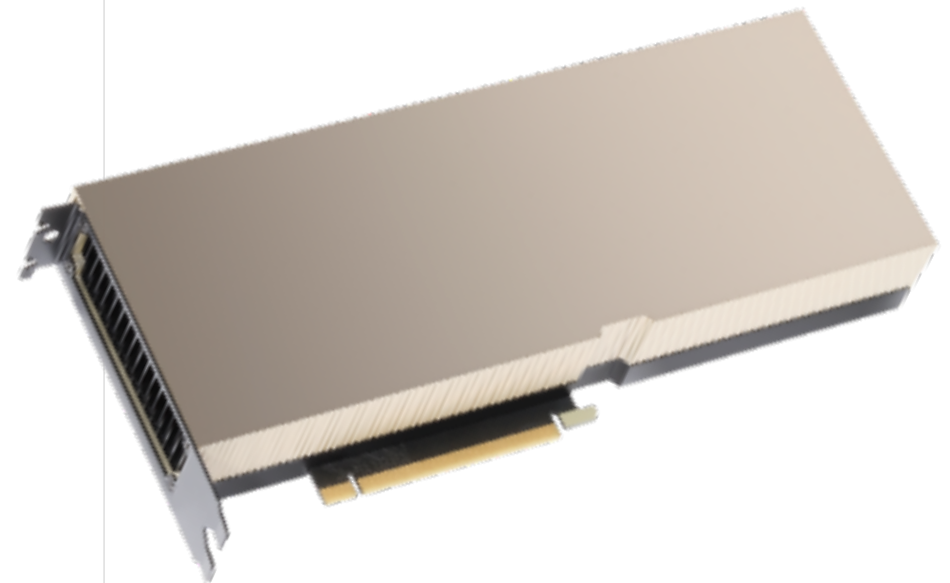
2 Slot FHFL mainstream form factor

HIGHEST PERFORMING SERVER CONNECTIVITY

128GB/s PCI Gen5

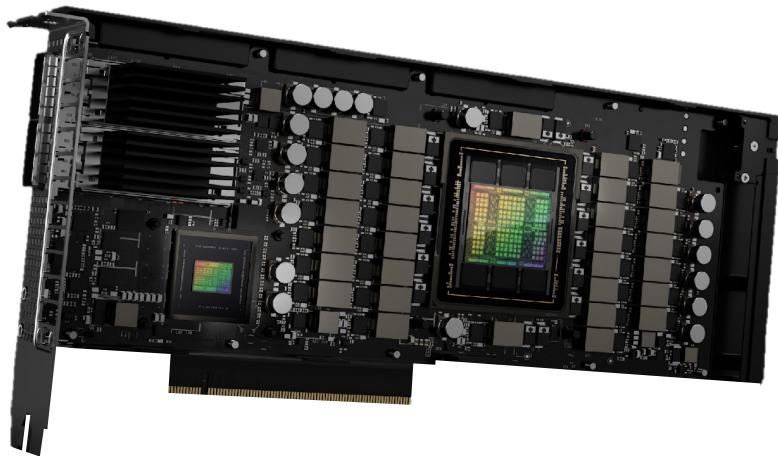
600 GB/s GPU-2-GPU connectivity (5X PCIe Gen5)

up to 2 GPUs with NVLink Bridge



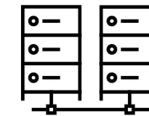
H100 CNX converged accelerator

- Delivering high-speed GPU-network I/O to mainstream servers



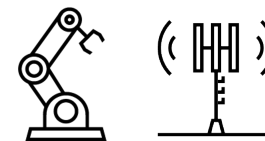
350W | 80GB | 400 Gb/s Ethernet or InfiniBand
PCIe Gen 5 | 2-Slot FHFL | NVLink

MULTI-NODE TRAINING



High performance and scalability

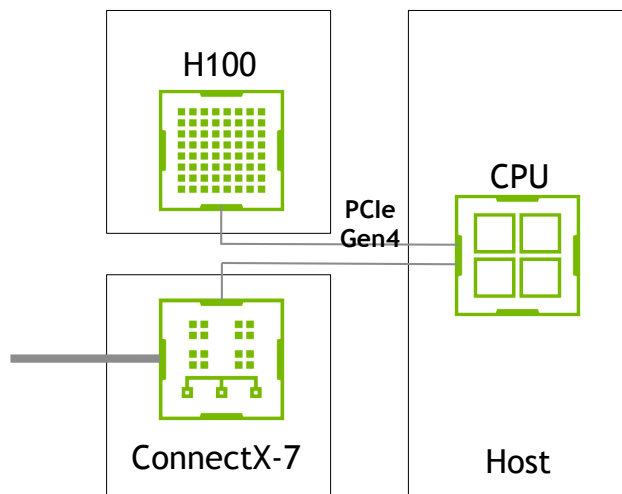
5G AI / PROCESSING



5G processing and AI services on a single commodity server

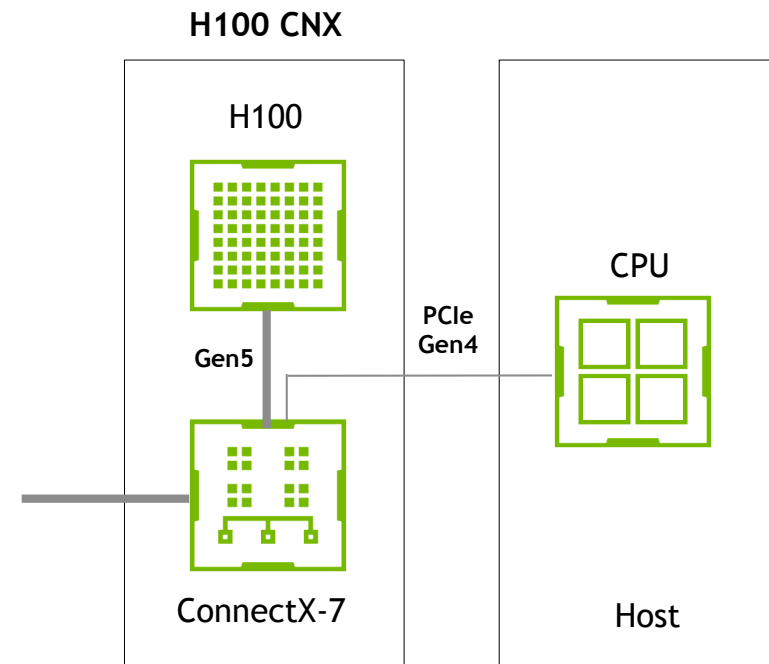
H100 CNX

● The mainstream choice for Multi-GPU



PCIe Gen4 Mainstream Server

- Throughput limited by Gen4 and CPU processing bottlenecks
- Reduced CPU performance from managing data transfers



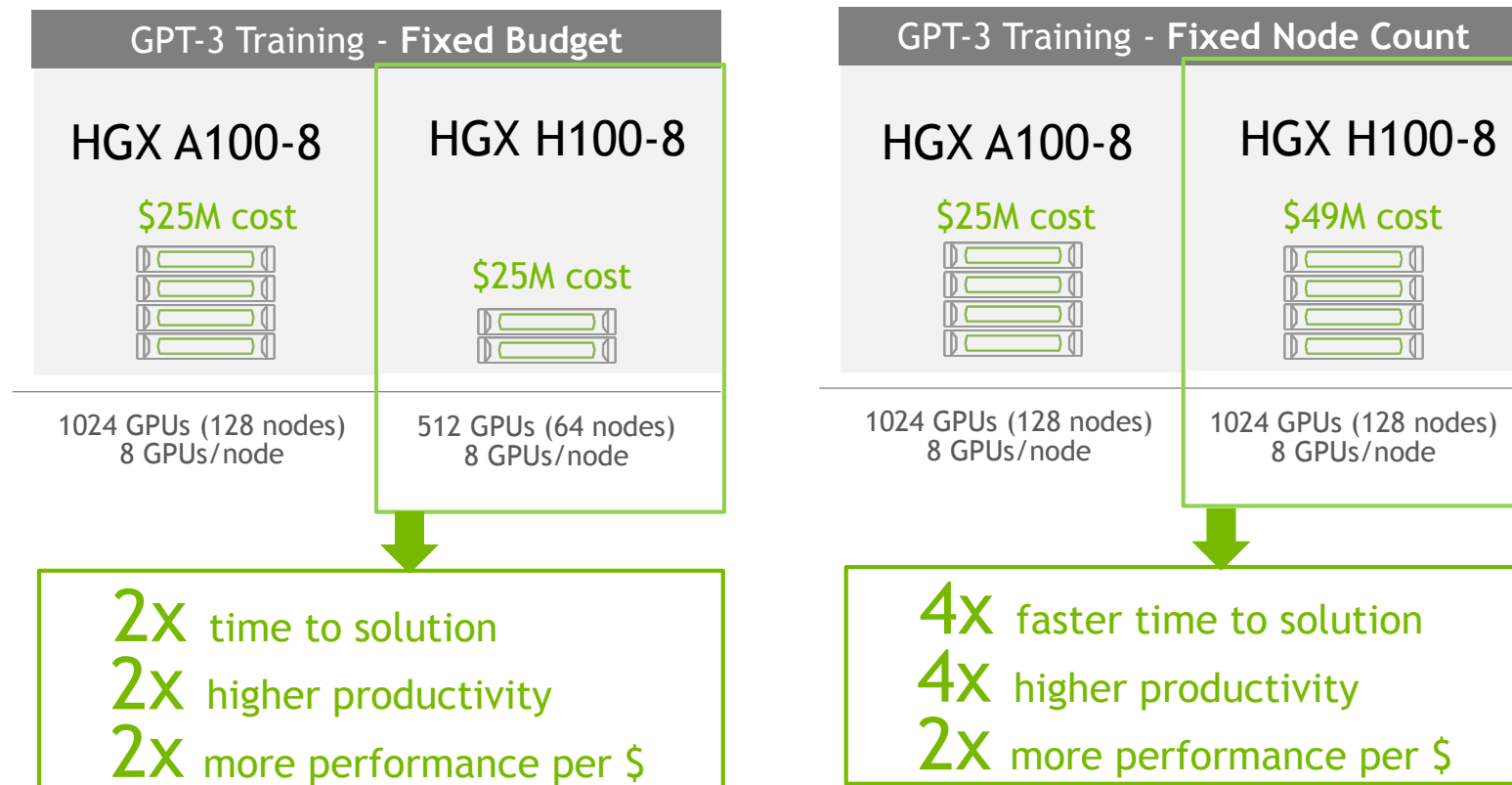
PCIe Gen4 Mainstream Server with H100 CNX

- Gen5 GPUDirect between network and H100 delivers 2X higher throughput
- CPU performance increases
- Scalable multi-node GPU processing

System configuration: 2U, 2S 64C CPU, 1024GB RAM, 2TB SSD, ConnectX-7 Dx NIC on H100 PCIe config

Highest performance training with H100

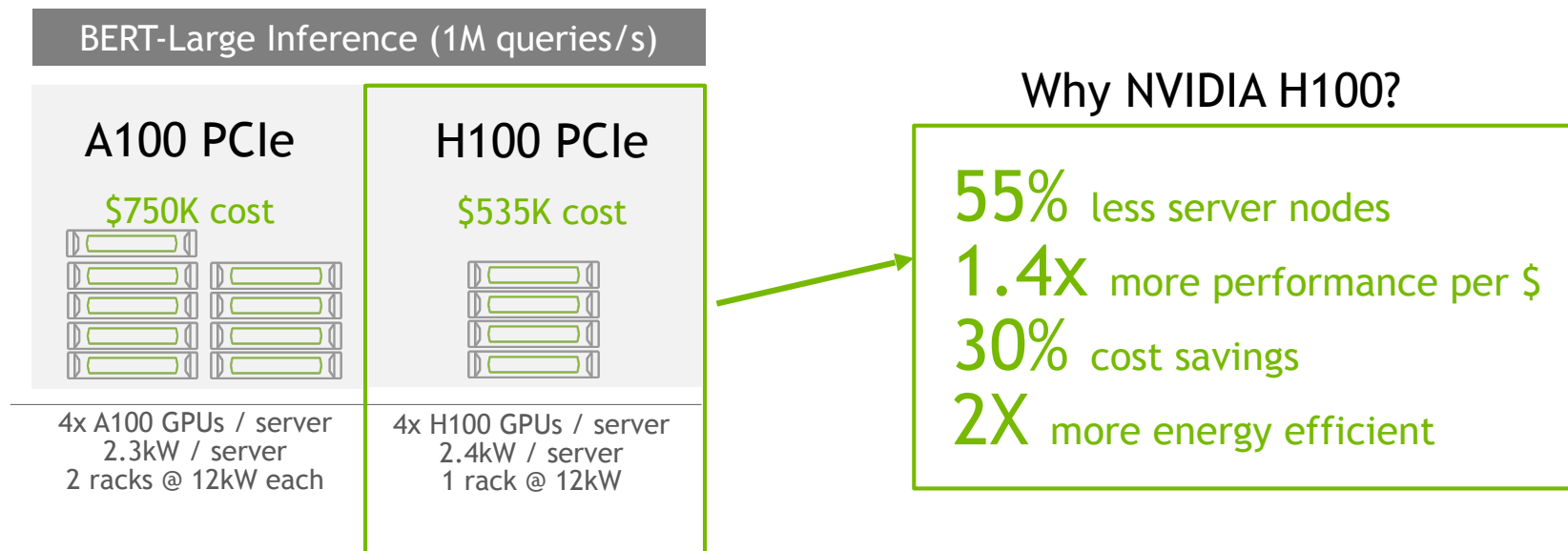
4x higher performance over A100



Server cost for representation purpose. Please contact your OEM/ODM for actual costs
 System configuration (Training): HGX A100 8-way | HGX H100 8-way excludes NVlink Switch System

Reduce cost for inference workloads with H100

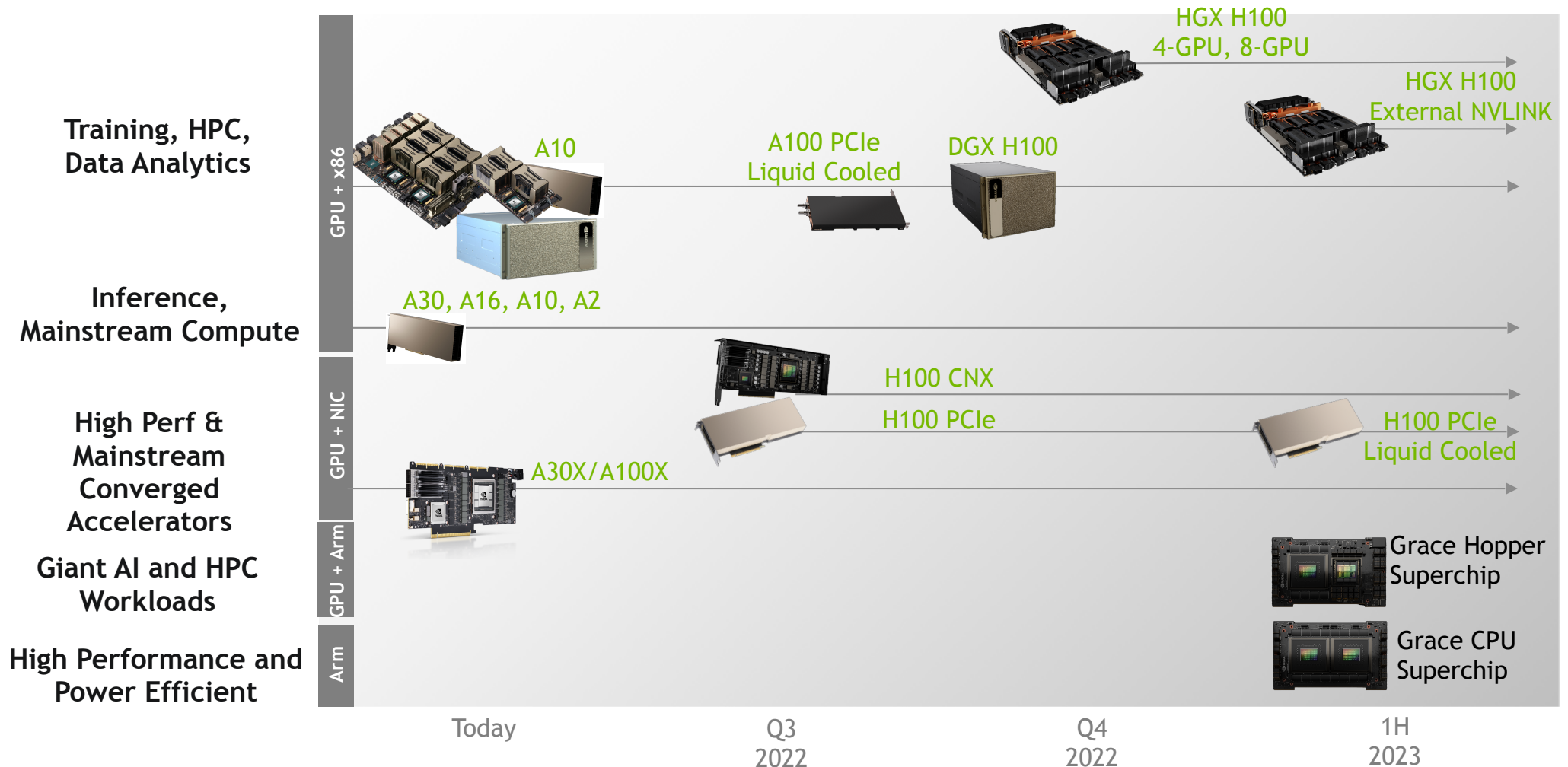
- Optimal compute for large inference deployments



Server cost for representation purpose. Please contact your OEM/ODM for actual costs
System configuration: 2U, 2S 64C CPU, 1024GB RAM, 2TB SSD, ConnectX-7 Dx NIC. 3 Year Hosting Cost: \$150/kW/m

Portfolio availability

Hopper coming soon



Choose the right H100 GPU

GPU	Availability	Training, Inference, HPC, Data Analytics		
		Highest Performance	Mainstream Servers	
			Multi-Node Jobs	Single-Node Jobs
H100 SXM	Q3 '22	DGX HGX 4-Way/8-Way ●		
H100 CNX	Q4 '22		HGX CNX ●	HGX CNX ●
H100 PCIe	Q3 '22			HGX PCIe ●

Price-performance comparison relative within each column

● Good
 ● Better
 ● Best

Data center GPU comparison

1. Coming soon
2. Supported on [Azure NVIDIA A100](#) with reduced performance compared to
3. A100 without Confidential Computing or H100 with Confidential Computing.
4. All Tensor Core numbers with sparsity



	H100		A100		A30	A2	T4	A40	A10	A16
Design	Highest Perf AI, Big NLP, HPC, DA		High Perf Compute		Mainstream Compute	Entry-Level Small Footprint	Small Footprint Datacenter Inference	High Perf Graphics	Mainstream Graphics & Video with AI	High Density Virtual Desktop
Form Factor	SXM5	x16 PCIe Gen5 2 Slot FHFL 3 NVLINK Bridge	SXM4	x16 PCIe Gen4 2 Slot FHFL 3 NVLink Bridge	x16 PCIe Gen4 2 Slot FHFL 1 NVLink Bridge	x8 PCIe Gen4 1 Slot LP	x16 PCIe Gen3 1 Slot LP	x16 PCIe Gen4 2 Slot FHFL 1 NVLink Bridge	x16 PCIe Gen4 1 slot LP	x16 PCIe Gen4 2 Slot FHFL
Max Power	700W	350W	500W	300W	165W	40-60W	70W	300W	150W	250W
FP64 TC FP32 TFLOPS ³	60 60	48 48	19.5 19.5		10 10	NA 4.5	NA 8	NA 37	NA 31	NA 4x4.5
TF32 TC FP16 TC TFLOPS ³	1000 2000	800 1600	312 624		165 330	18 36	NA 65	150 300	125 250	4x18 4x36
FP8 TC INT8 TC TFLOPS/ TOPS ³	4000 4000	4000 4000	NA 1248		NA 661	NA 72	NA 130	NA 600	NA 500	NA 4x72
GPU Memory / Speed	80GB HBM3	80GB HBM2e	80GB HBM2e		24GB HBM2	16GB GDDR6	16GB GDDR6	48GB GDDR6	24GB GDDR6	4x 16GB GDDR6
Multi-Instance GPU (MIG)	Up to 7		Up to 7		Up to 4	-	-	-	-	-
NVLink Connectivity	Up to 256	2	Up to 8	2	2	-	-	2	-	-
Media Acceleration	7 JPEG Decoder 7 Video Decoder		1 JPEG Decoder 5 Video Decoder		1 JPEG Decoder 4 Video Decoder	1 Video Encoder 2 Video Decoder (+AV1 decode)	1 Video Encoder 2 Video Decoder	1 Video Encoder 2 Video Decoder (+AV1 decode)		4 Video Encoder 8 Video Decoder (+AV1 decode)
Ray Tracing	-		-		-	Yes	Yes	Yes	Yes	Yes
Transformer Engine	Yes		-		-	-	-	-	-	-
DPX Instructions	Yes		-		-	-	-	-	-	-
Graphics	For in-situ visualization (no NVIDIA vPC or RTX vWS)		For in-situ visualization (no NVIDIA vPC or RTX vWS)		-	Good	Good	Best	Better	Good
vGPU	Yes ¹		Yes		-	Yes	Yes	Yes	Yes	Yes
Hardware Root of Trust	Yes		Optional		-	Optional	-	Optional	Optional	Optional
Confidential Computing	Yes		(2)		-	-	-	-	-	-
Server Availability	Q3'22	Q3'22	In Production		In Production	In Production	In Production	In Production	In Production	In Production

Choose the right data center GPU

	GPU	Availability	DL Training & DA	DL Inference	HPC / AI	Render Farms	Virtual Workstation	Virtual Desktop (VDI)	Mainstream Acceleration	Far Edge Acceleration	AI-on-5G
Compute	H100	Q3 '22	●	●	●				●		●
	A100	Now	●	●	●				●		●
	A30	Now		●	●				●		●
Graphics / Compute	A40	Now				●	●		●		
	A10	Now		●		●	●	●	●	●	
	A16	Now					●	●			
Small Form Factor Compute/Graphics	A2	Now		●			●	●	●	●	
	T4	Now		●			●	●	●	●	

Good
 Better
 Best

Price-performance comparison within each product group (Compute, Graphics & Compute, SFF Compute & Graphics) and workload column

SXM form factor
 H100 + ConnectX7 Converged PCIe card
 PCIe form factor
 A100X / A30X
 A100 or A30 + BlueField2 Converged PCIe Card

Delivering the AI center of excellence for enterprise

- Best of breed infrastructure for AI development built on DGX

NVIDIA DGX H100



The World's First AI System with NVIDIA H100

8x NVIDIA H100 | 32 PFLOPS FP8 (6X) | 0.5 PFLOPS FP64 (3X)
640 GB HBM3 | 3.6 TB/s (1.5X) BISECTION B/W

4th Generation of the World's Most Successful Platform Purpose-Built for Enterprise AI

COMING LATE 2022

DGX SuperPOD with DGX H100



32 DGX H100 | 1 EFLOPS AI
NVLINK SWITCH SYSTEM | QUANTUM-2 IB |
20TB HBM3 | 70 TB/s BISECTION B/W (11X)

1 ExaFLOPS of AI Performance in 32 Nodes
Scale as large as needed in 32 node increments

X-Factors compare performance over DGX SuperPOD with DGX A100 supercomputer configuration with same number of nodes

Announcing Nvidia EOS Supercomputer

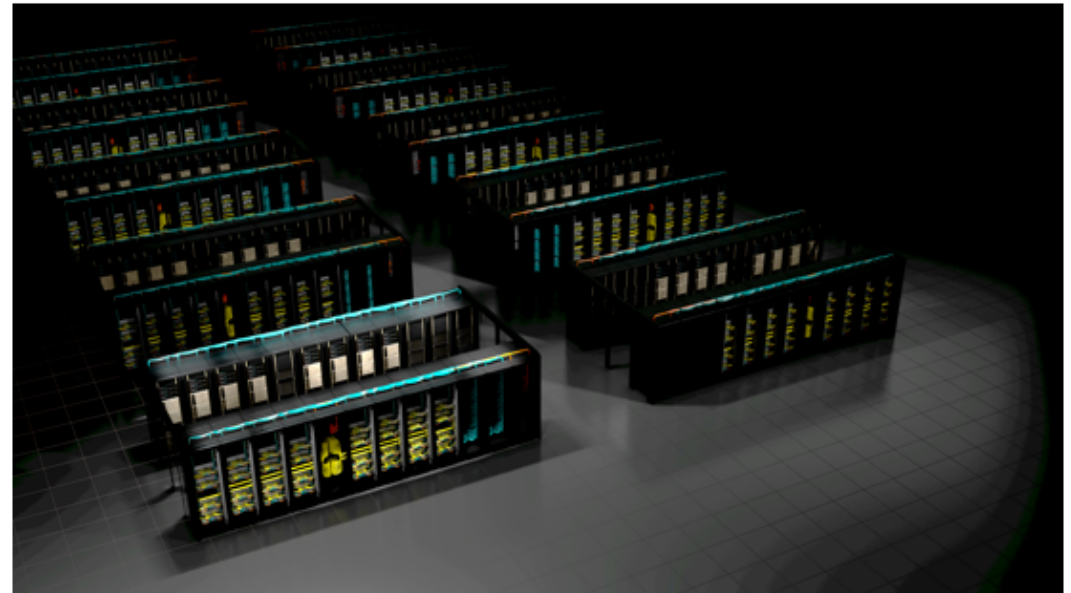
● The world's most advanced AI infrastructure

NVIDIA Eos

DGX SuperPOD Powered by 576 DGX H100 Systems |
500 Quantum-2 IB Switches | 360 NVLink Switches

FP8	18 EFLOPS	6X
FP16	9 EFLOPS	3X
FP64	275 PFLOPS	3X
In-Network Compute	3.7 PFLOPS	36X
Bisection Bandwidth	230 TB/s	2X
NVLINK Domain	256 GPUs	32X

Blueprint for OEM and Cloud Partner Offerings



Cloud Native | Performance Isolation | Multi-Tenant

X-Factors compare performance over DGX A100 SuperPOD based supercomputer configuration with same number of Nodes

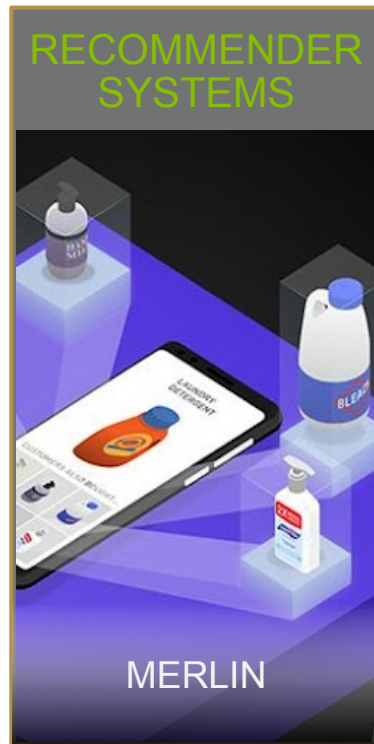


VI. Nvidia AI Platform

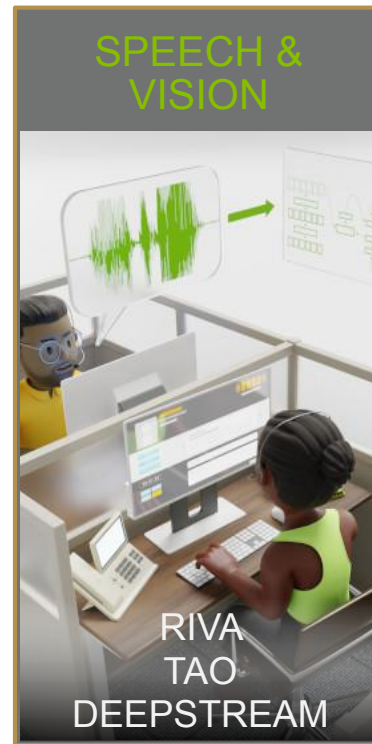
Accelerating the Next Wave of AI: AI Platform Updates



Analytics, ML,
Visualization



Personalization
Engine, Simplified



Conversational AI,
Video Analytics















Fast, Scalable
Predictions



Large Language
Model

TAO

- Framework for creating custom, production-ready models to power speech and vision AI applications.

<p>Train, Adapt and Optimize in hours, rather than months</p> <ul style="list-style-type: none">  Removes the need for AI expertise  Build custom models faster  Optimized for inference  Simplified model deployment with Riva, DeepStream and Triton 	<p>New Version of TAO Toolkit Low-Code Version of TAO</p> <ul style="list-style-type: none">  Import model weights from publicly available models  Deploy as a service with REST APIs  Visualize with TensorBoard  New vision models 	<p>Accelerating AI Across Industries</p> <ul style="list-style-type: none">    
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

<https://developer.nvidia.com/tao>

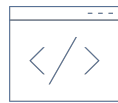
Triton

- Open-source inference serving software for fast, easy inference deployment.

Fast And Scalable AI In Every App



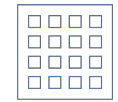
Any Model



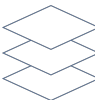
Any Framework



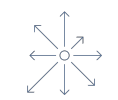
Any Query Type



Any Processor



Any Deployment Platform



Any Deployment Location

Key Feature Updates

Shapley Values in FIL Backend
Explanation of model prediction

Triton Management Service
Efficient scaling in Kubernetes

Model Navigator
Accelerated time to production

1000's Of Users | 1.3M+ Downloads/Clones



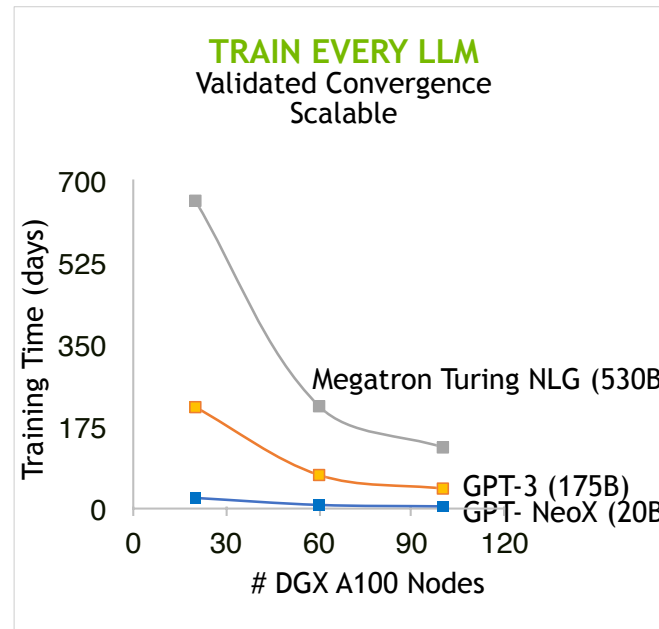
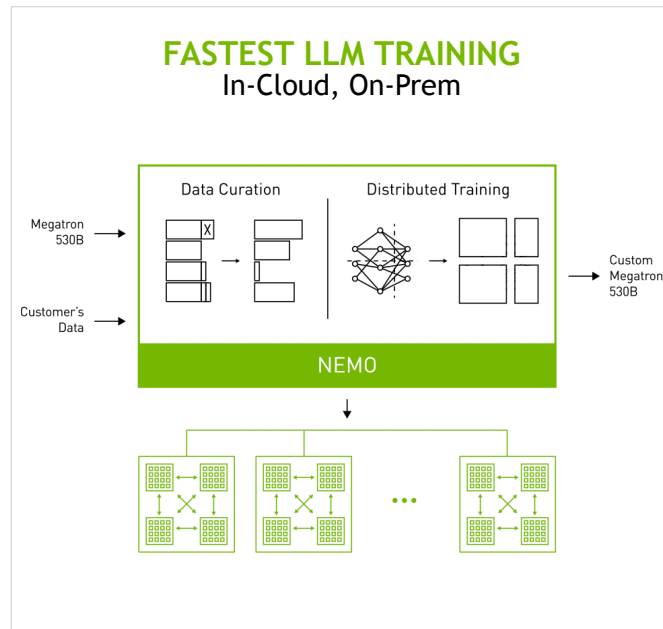
Real time Spell Check in Product Search



Document Translation

Nemo Megatron

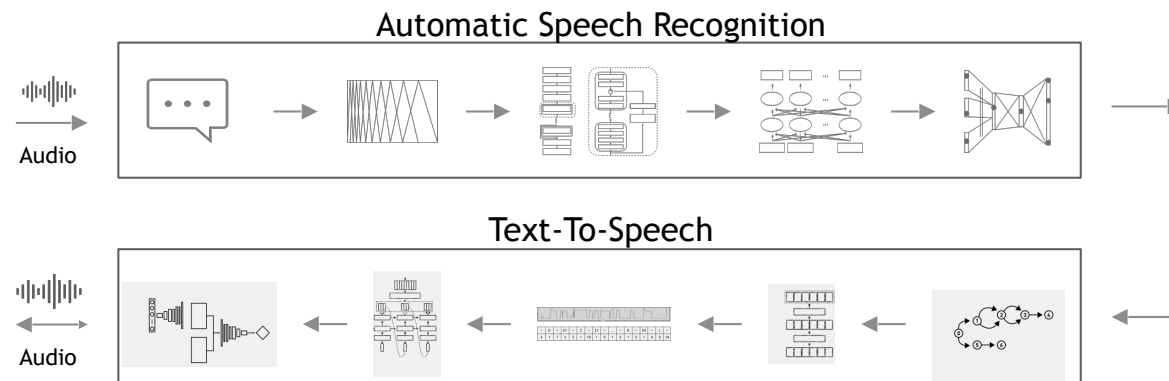
- Accelerated framework for training large language models.



MULTIPLE LANGUAGES & INDUSTRIES

Riva 2.0

- World-class speech AI.
- Fully customizable.
- Supported with Riva Enterprise.



UNLIMITED USAGE

Scale to any cloud & on-premises

ACCESS TO NVIDIA AI EXPERTS

8-5 local business hours
Guidance on configurations, performance

CONTROL MAINTENANCE & UPGRADE SCHEDULE

Long-term support for up to 3 years

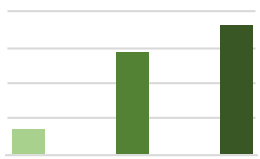
TICKET PRIORITIZATION

Latest security fixes and maintenance releases


NGC Catalog

The hub of GPU-optimized software

PERFORMANCE OPTIMIZED
Tested across GPU-accelerated Platforms



Monthly software container updates

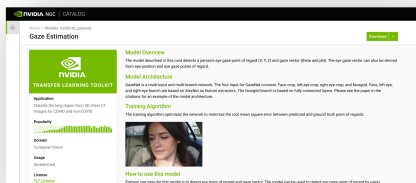


SOTA models

FULLY TRANSPARENT
Quickly identify and deploy the right software


vulnerabilities	OS package	Medium	(CVE-2021-3995) libmount1
vulnerabilities	OS package	Medium	(CVE-2021-3995) fdisk
vulnerabilities	OS package	Medium	(CVE-2019-9157) hdf5-helpers
vulnerabilities	OS package	Medium	(CVE-2018-17233) hdf5-helpers

Detailed security scan reports

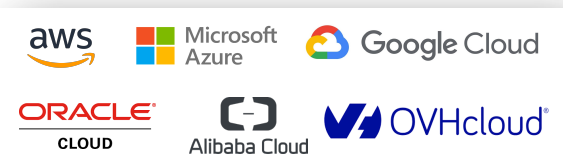


Model resumes

ACCELERATES DEVELOPMENT
Focus on building, not setup



One click deploy from NGC



Develop once.
deploy anywhere with NVIDIA VMI

1.5M+ users millions of downloads

*8x NVIDIA A100 40GB. NVIDIA DGX. ResNet-50. Mixed Precision. 256 batch size.

Thanks for your attention!

- You can always reach me in Spain at the Computer Architecture Department of the University of Malaga:
 - e-mail: ujaldon@uma.es
 - Web page: <http://manuel.ujaldon.es> (english/spanish versions available).

