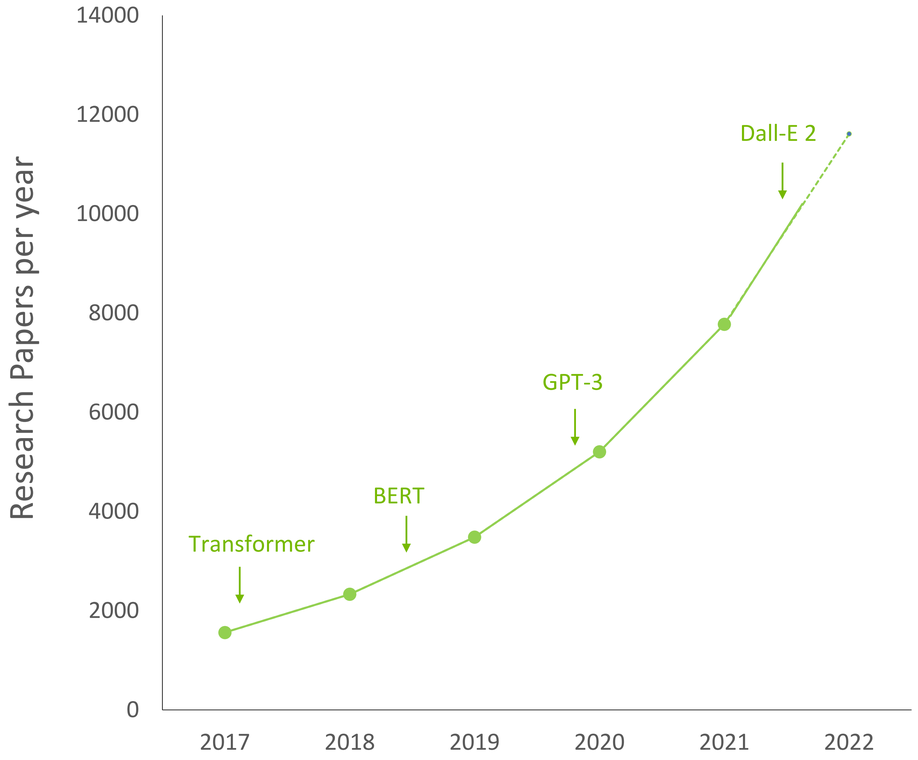
NVIDIA H100 Customer Deck

September 2022

Large Language Models Codifying Intelligence

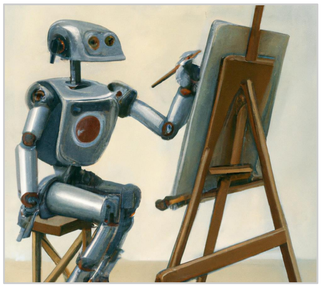
LLM research accelerating innovation and abilities

Transformer and LLM Research Papers Per Year

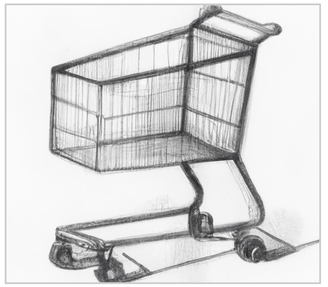


Explosion of Use Cases

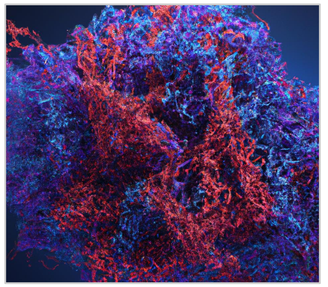
IMAGE GENERATION
Brand Creation
Gaming Characters



RECOMMENDATIONS
eCommerce
Personalized Content



LIFE SCIENCE RESEARCH
Molecular Representations
Drug Discovery



TRANSLATION
Translating Wikipedia
Real-Time Metaverse Translation

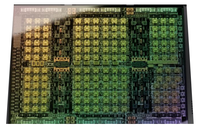
TEXT GENERATION
Summarization
Marketing Copy

CODING
Dynamic Code Comments
Function Generation

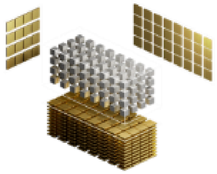
*Research paper published on Arxiv.org related to Transformers and LLMs in computer science subject area, with projected count for rest of 2022

NVIDIA Hopper

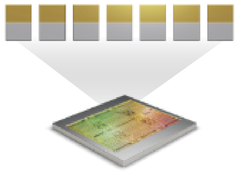
The new engine for the world's AI infrastructure



World's Most Advanced Chip



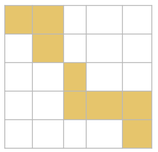
Transformer Engine



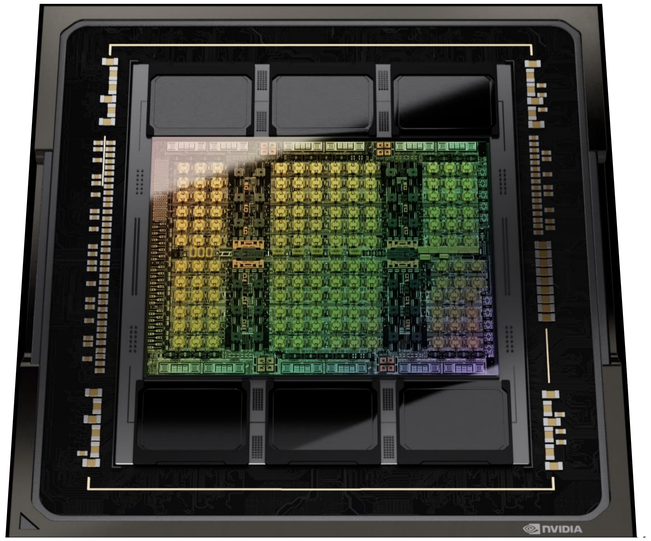
2nd Gen MIG

Confidential Computing

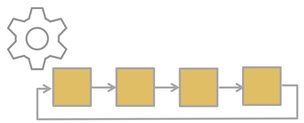
4th Gen NVLink



DPX Instructions



Custom 4N TSMC Process | 80 billion transistors



NVIDIA AI Enterprise Software Suite

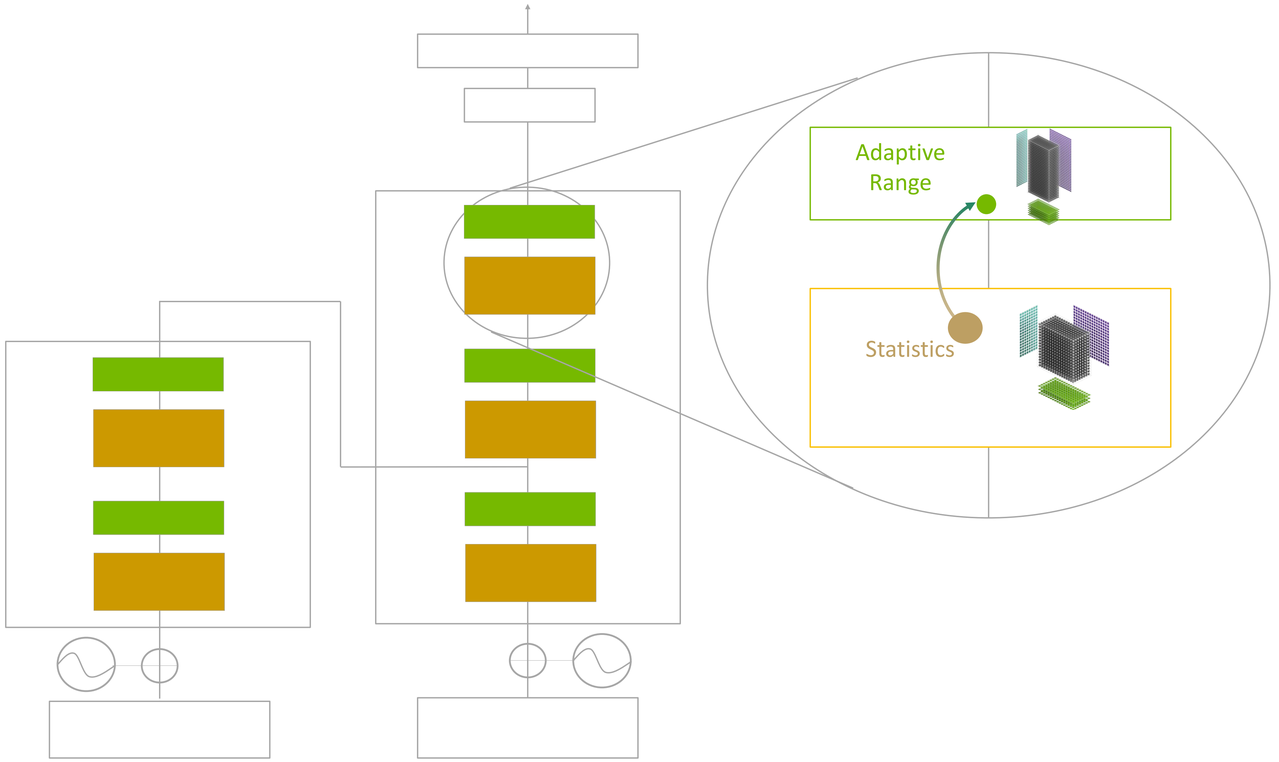
*Redeemable NVIDIA AI Enterprise 5 Year Subscription**

*Included for H100 PCIe in mainstream systems

Transformer Engine

Tensor core optimized for transformer models

- 6X Faster Training and Inference of Transformer Models
- NVIDIA Tuned Adaptive Range Optimization Across 16-bit and 8-bit Math
- Configurable Macro Blocks Deliver Performance Without Accuracy Loss

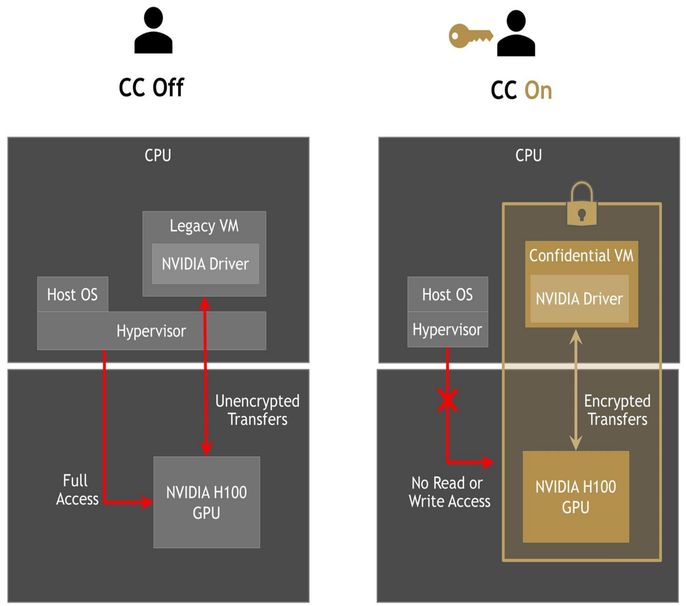


Statistics and Adaptive Range Tracking

Hopper Technological Breakthroughs

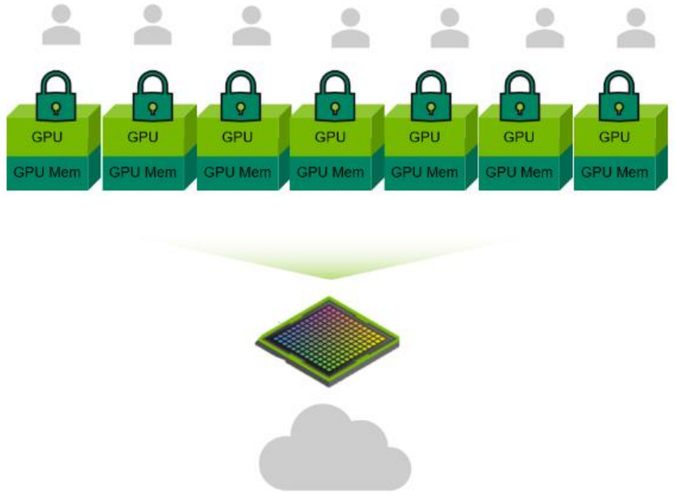
CONFIDENTIAL COMPUTING

Secure Data and AI Models In-Use



MULTI-INSTANCE GPU

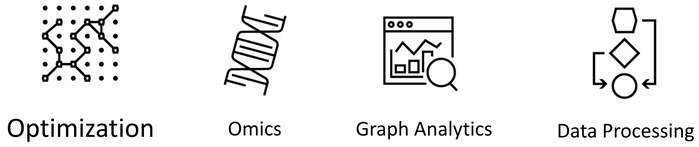
7 Secure Tenants on 1 GPU



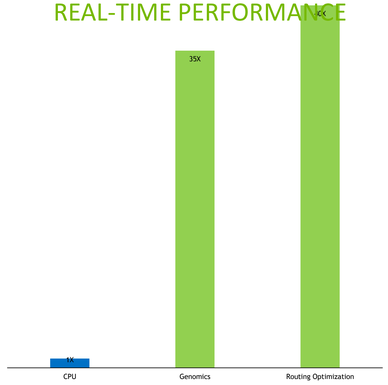
NEW DYNAMIC PROGRAMMING INSTRUCTIONS

Accelerate Dynamic Programming Algorithms

A BROAD RANGE OF USE CASES



REAL-TIME PERFORMANCE



Dynamic Programming Instructions HGX H100 4-GPU vs dual socket 32 core IceLake

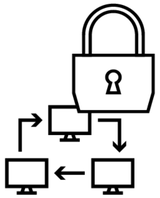
Confidential Computing Use Cases And Market Trends

CONFIDENTIAL COMPUTING

Trusted Execution Environment

Secure Data in Use with Confidential Computing

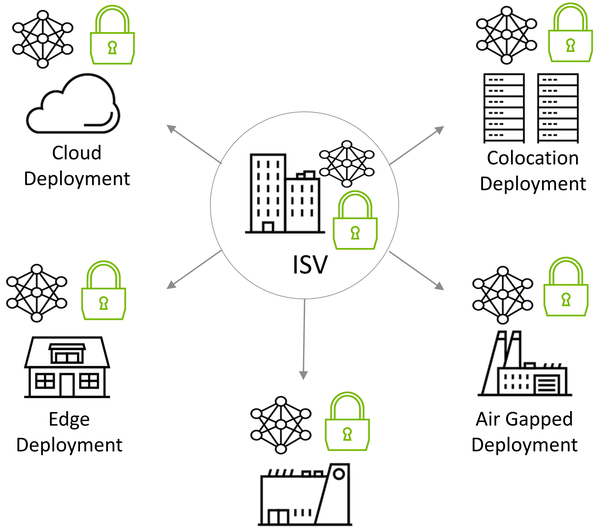
Secured Data At Rest



Secured Data in Transit

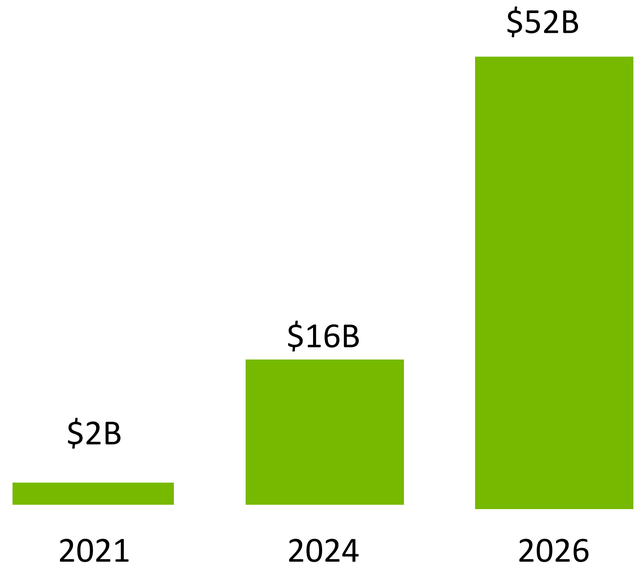
SECURE DATA AND AI MODELS IN USE

Protect Across Deployment Strategies



MARKET OPPORTUNITY

26X TAM Growth Over Next 5 Years



Confidential Computing TAM

Market opportunity reference
Everest Group: Confidential Computing The Next Frontier in Data Security report from confidential Compute Consortium

NVIDIA H100

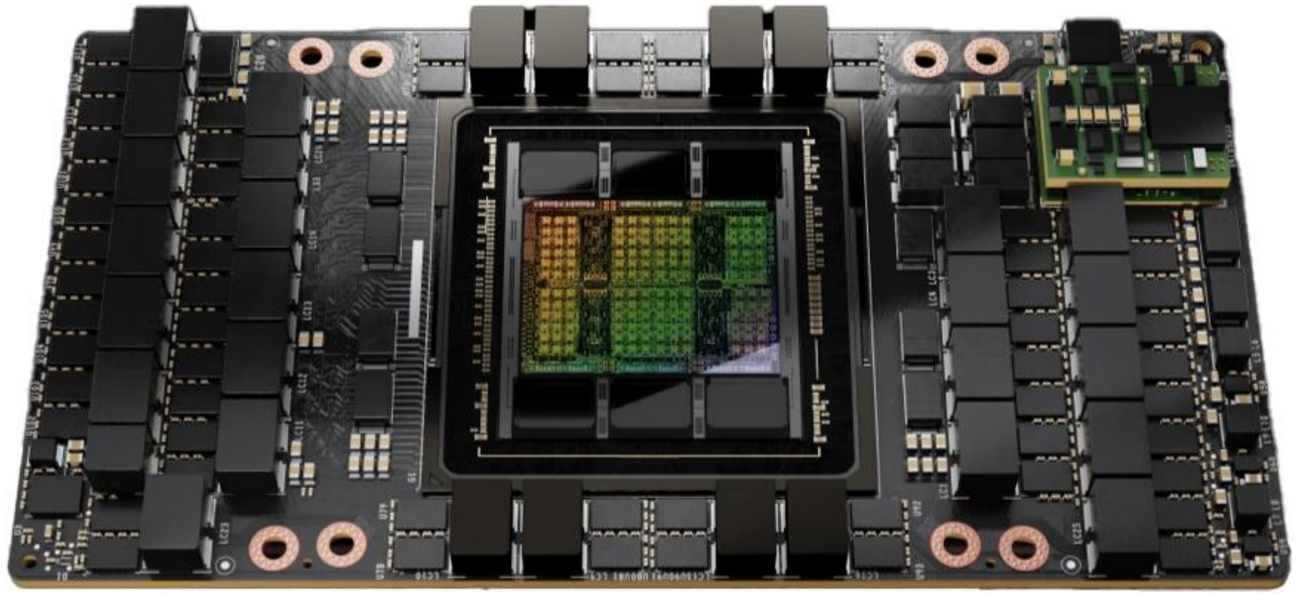
Unprecedented performance, scalability, and security for every data center

HIGHEST AI AND HPC PERFORMANCE
4PF FP8 (6X)| 2PF FP16 (3X)| 1PF TF32 (3X)| 60TF FP64 (3X)
3TB/s (1.5X), 80GB HBM3 memory

TRANFORMER MODEL OPTIMIZATIONS
6X faster on largest transformer models

HIGHEST UTILIZATION EFFICIENCY AND SECURITY
7 Fully isolated & secured instances, guaranteed QoS
2nd Gen MIG | Confidential Computing

FASTEST, SCALABLE INTERCONNECT
900 GB/s GPU-2-GPU connectivity (1.5X)
up to 256 GPUs with NVLink Switch | 128GB/s PCI Gen5

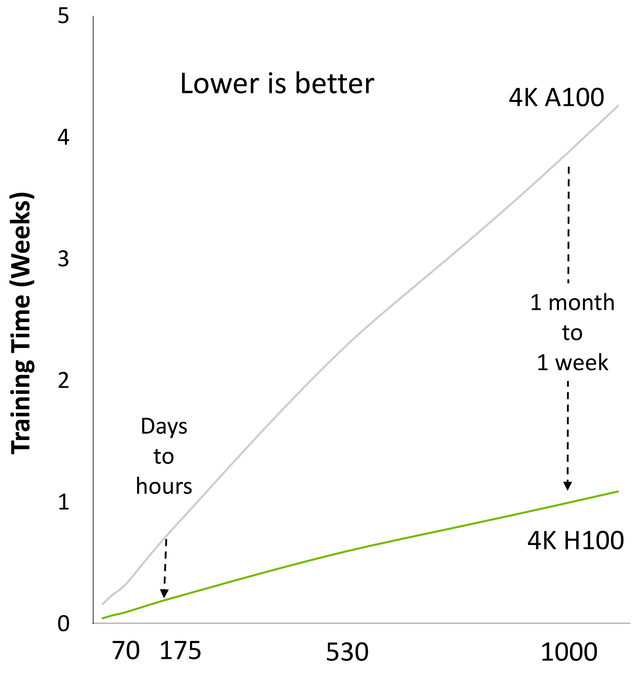


FP8, FP16, TF32 performance include sparsity. X-factor compared to A100

NVIDIA H100 Supercharges LLMs

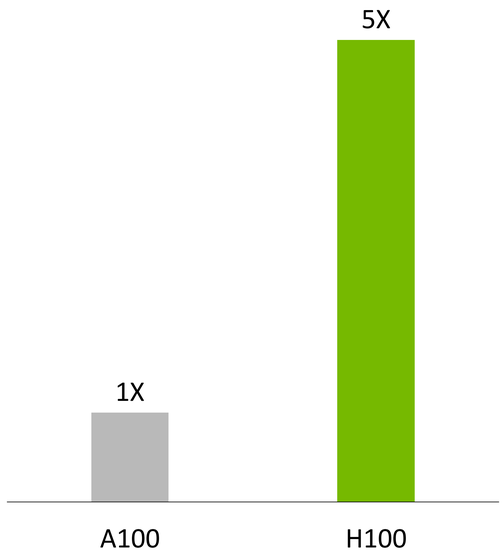
Hopper architecture addresses LLM needs at scale

Supercharged LLM Training



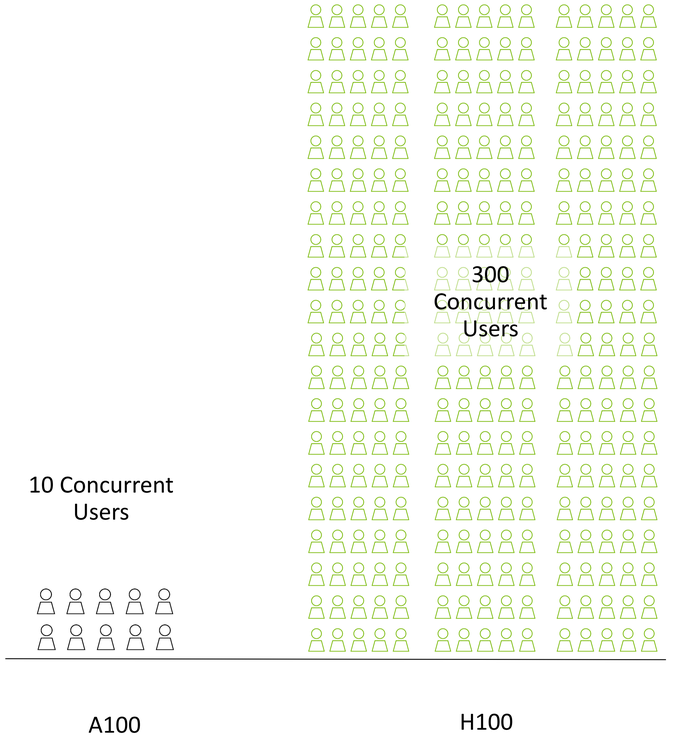
Time-to-Train by LLM Size
(Billion parameters)

High Performance Prompt Learning



530B P-Tuning Time-to-Train

30X Real-Time Inference Throughput



530B Inference on 10 DGX systems

LLM Training | 4096 GPUs | H100 NDR IB | A100 HDR IB | 300 Billion tokens.
P-Tuning | DGX H100 | DGX A100 | 530B Q&A tuning using SQuAD dataset
Inference | chatbot| 10 DGX H100 NDR IB | 10 DGX A100 HDR IB | <1 sec latency | 1 inference/second/user.
H100 data center projected workload performance, subject to change

NVIDIA HGX H100

The world’s most advanced enterprise AI infrastructure

HIGHEST PERFORMANCE FOR AI AND HPC

4-way / 8-way H100 GPUs with 32 PetaFLOPs FP8

3.6 TFLOPs FP16 in-network SHARP Compute

NVIDIA Certified High-Performance Offering from All Makers

FASTEST, SCALABLE INTERCONNECT

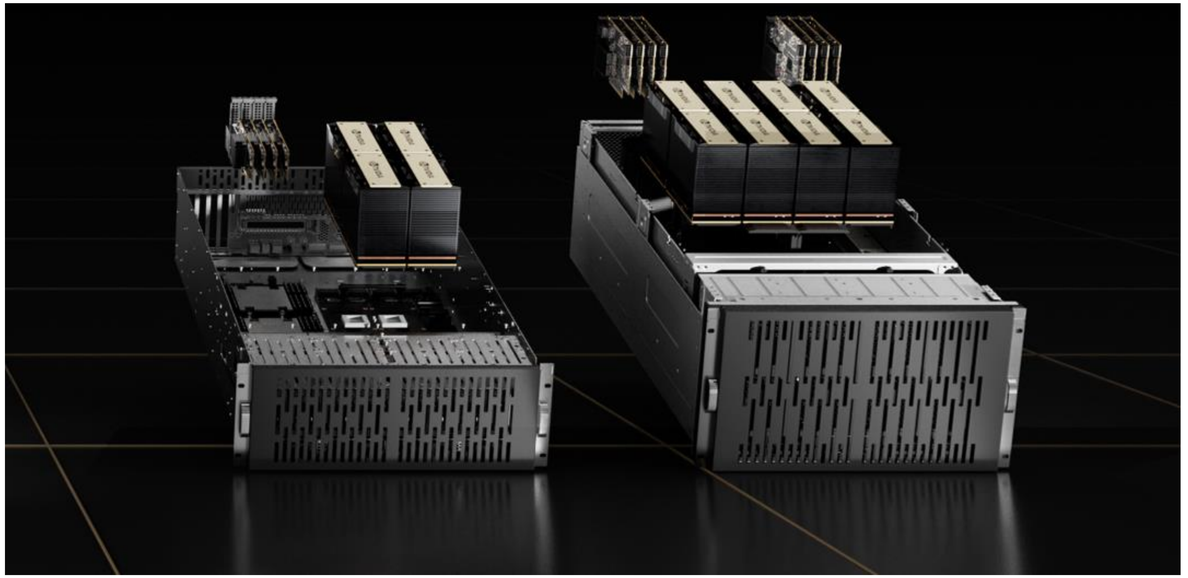
4th Gen NVLINK with 3X faster All-Reduce communications

3.6 TB/s bisection bandwidth

NVLINK Switch System Option Scales Up to 256 GPUs

SECURE COMPUTING

First HGX System with Confidential Computing



Atos CISCO DELL Technologies FUJITSU

GIGABYTE Hewlett Packard Enterprise Lenovo SUPERMICRO

Tensor Core FLOPs shown with sparsity | Speedups compared to prior generation

NVIDIA H100 PCIe

Unprecedented performance, scalability, and security for mainstream servers

HIGHEST AI AND HPC MAINSTREAM PERFORMANCE

3.2PF FP8 (5X) | 1.6PF FP16 (2.5X) | 800TF TF32 (2.5X) | 48TF FP64 (2.5X)
6X faster Dynamic Programming with DPX Instructions
2TB/s , 80GB HBM2e memory

HIGHEST COMPUTE ENERGY EFFICIENCY

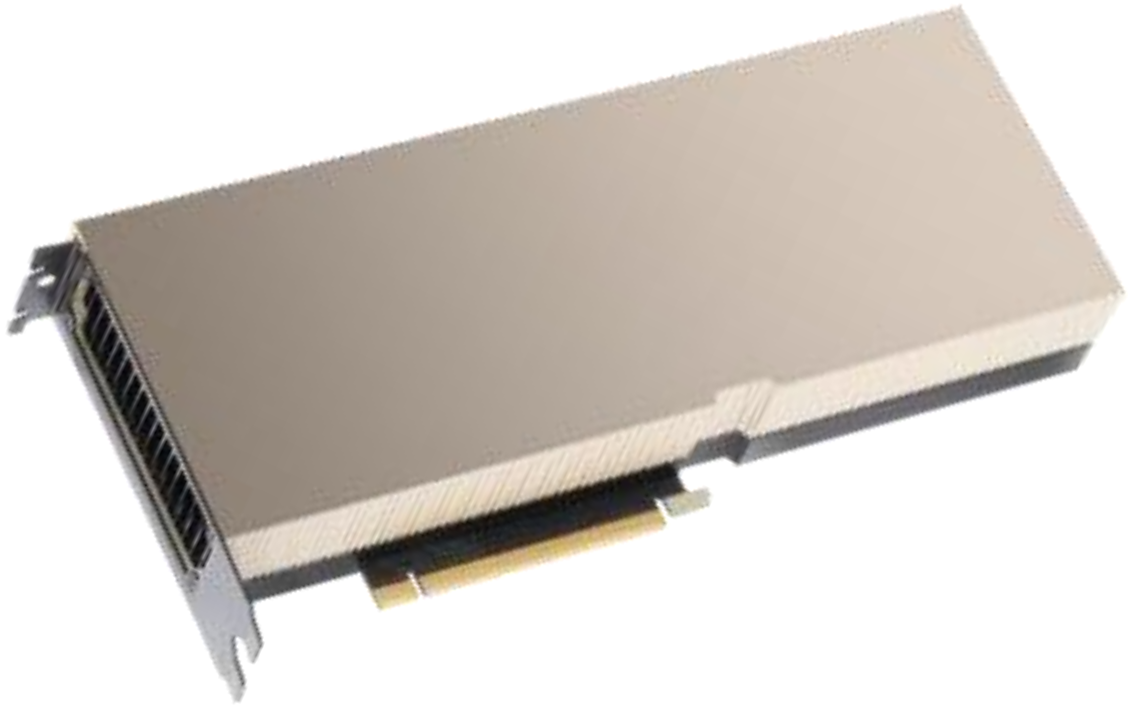
Configurable TDP - 150W to 350W
2 Slot FHFL mainstream form factor

HIGHEST UTILIZATION EFFICIENCY AND SECURITY

7 Fully isolated & secured instances, guaranteed QoS
2nd Gen MIG | Confidential Computing

HIGHEST PERFORMING SERVER CONNECTIVITY

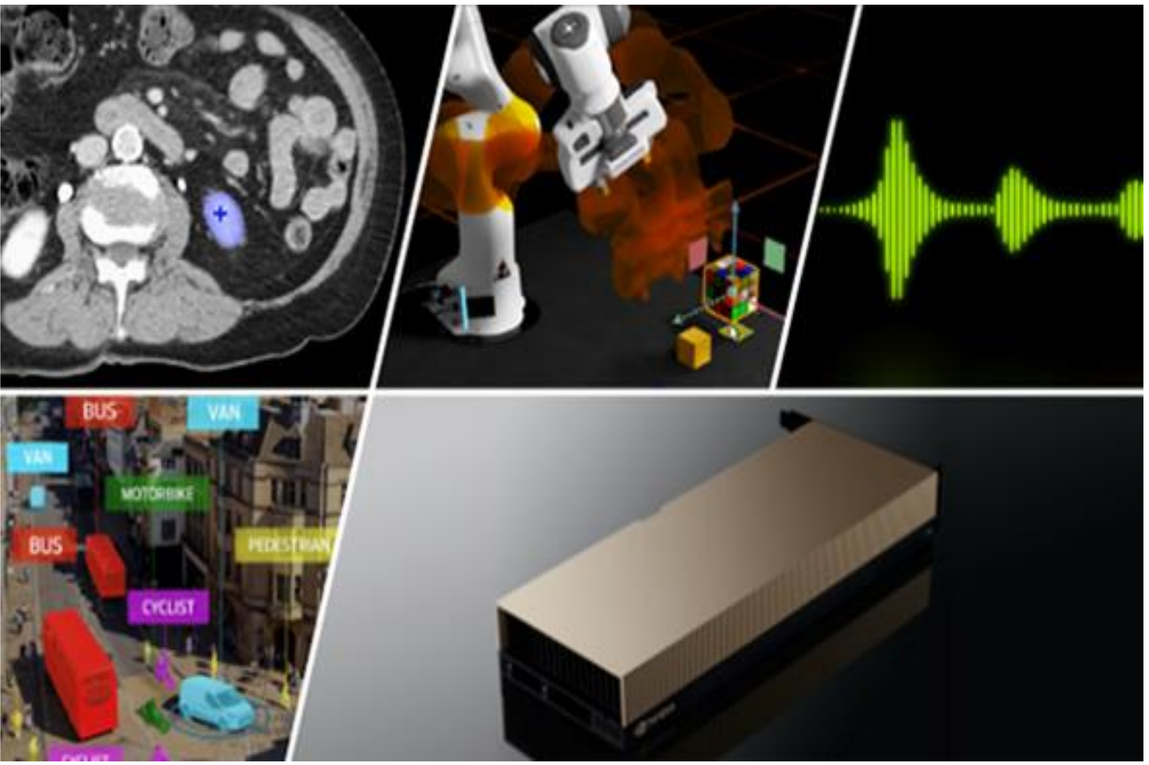
128GB/s PCI Gen5
600 GB/s GPU-2-GPU connectivity (5X PCIe Gen5)
up to 2 GPUs with NVLink Bridge



FP8, FP16, TF32 performance include sparsity. X-factor compared to A100

Production AI with NVIDIA H100 and NVIDIA AI Enterprise

Develop and deploy enterprise AI with unmatched performance, security, and scalability



5-YEAR SUBSCRIPTION OF NVIDIA AI ENTERPRISE

A cloud native software suite for development and deployment of AI

NVIDIA ENTERPRISE SUPPORT

Including access to NVIDIA AI experts, priority notifications of the latest security fixes and maintenance releases

ENTERPRISE TRAINING SERVICES

Developers, data scientists, and IT professionals learn how to get the most out of the NVIDIA AI platform

Software activation: www.nvidia.com/activate-h100

NVIDIA H100 GPU + NVIDIA AI Enterprise Value Proposition

AI-READY PLATFORM

- Turnkey solutions with NVIDIA H100 integrated NVIDIA-Certified Systems
- Quick time to value
- Large ecosystem of AI-accelerated applications
- Rapid PoC
- Services Partner network

END-TO-END AI SOFTWARE STACK

- Best-in-class AI tools and frameworks
- AI Domain frameworks
- Pre-trained models[1]
- Cloud Native and agile
- Kubernetes-based applications

PERFORMANT & EFFICIENT

- Fastest training with Transformer Engine
- Highest efficiency and security with MIG and confidential computing
- Streamlined market-leading inference performance with NVIDIA Triton Inference Server

AGILE AND SCALABLE

- Accelerate a wide range of use cases
- Flexibility to service entire AI pipeline on one infrastructure
- Effective resource sharing[2]
- Scales to multi-GPU and multi-node configurations

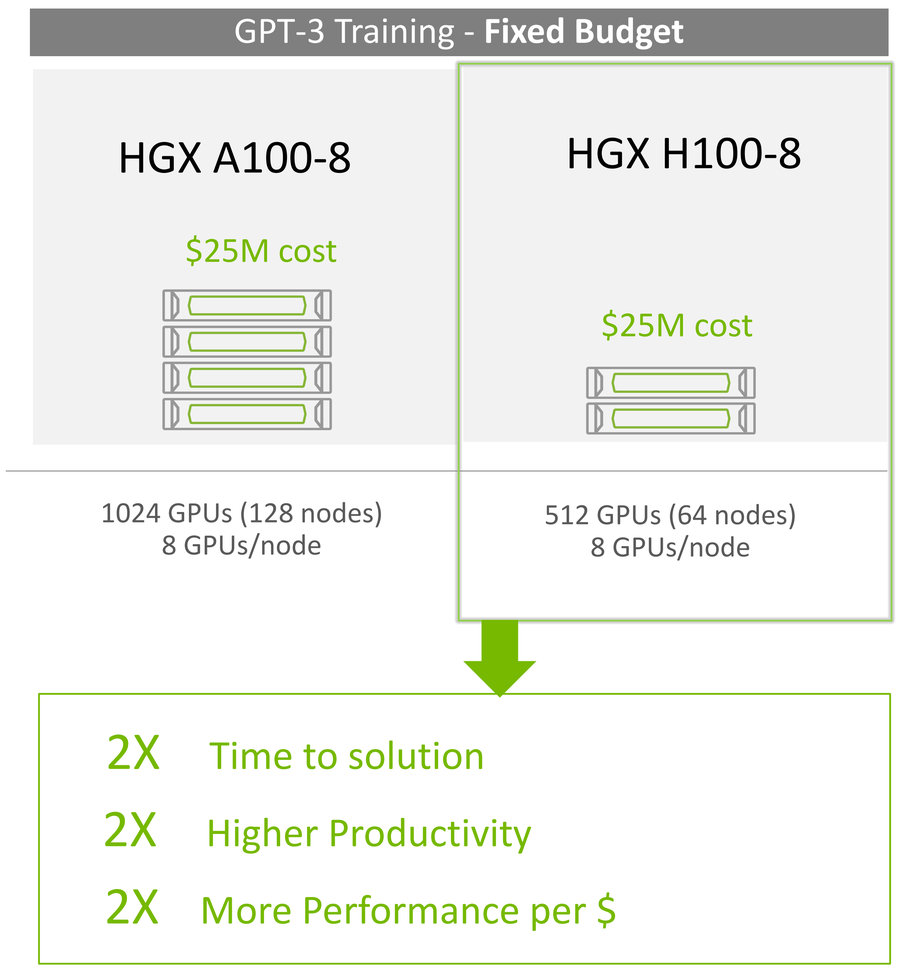
OPEN AND ENTERPRISE READY

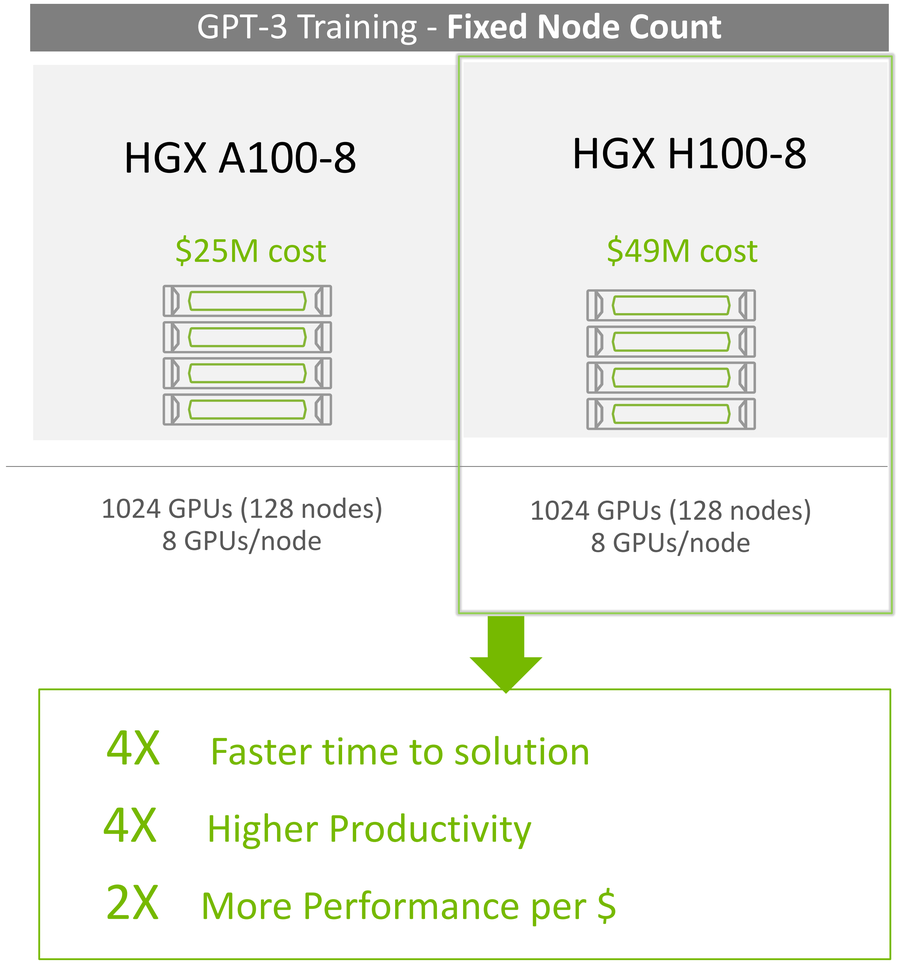
- Streamlined enterprise AI, built from proven open-source
- Certified to run on broadly adopted enterprise platforms in the data center and cloud
- Assurance of NVIDIA Support worldwide
- Guaranteed response times
- Priority security notifications

[1] Available in Q4 2022
[2] vGPU support for NVIDIA H100 available in Q4 2022

Highest Performance Training with H100

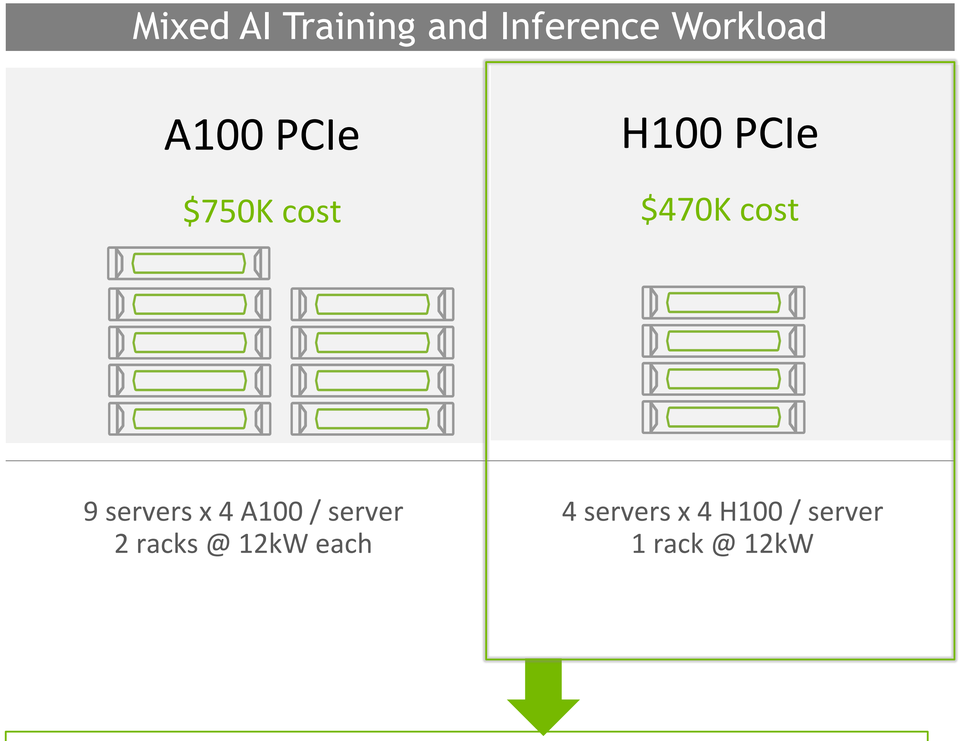
4x higher performance over A100





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

H100 PCIE – Delivers Dramatic TCO for AI



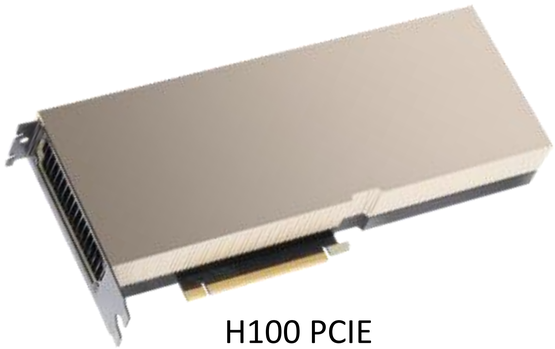
55% Less Server Nodes

1.6X More Performance per $

35% Cost savings

2.2X More Energy Efficient

NVIDIA AI Enterprise
5 Year Subscription



H100 PCIE

Server cost for representation purpose. Please contact your OEM/ODM for actual costs
Workload mix of 33% BERT-Large Training, 33% BERT-Large Inference, 33% Recommender Inference
System configuration: 2U, 2S 64C CPU, 1024GB RAM, 2TB SSD, ConnectX-7 Dx NIC. 3 Year Hosting Cost: $150/kW/m

Delivering the AI Center of Excellence for Enterprise

Best-of-breed infrastructure for AI development built on NVIDIA DGX

NVIDIA DGX H100

The World's Proven Choice for Enterprise AI

8x NVIDIA H100 GPUs | 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

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

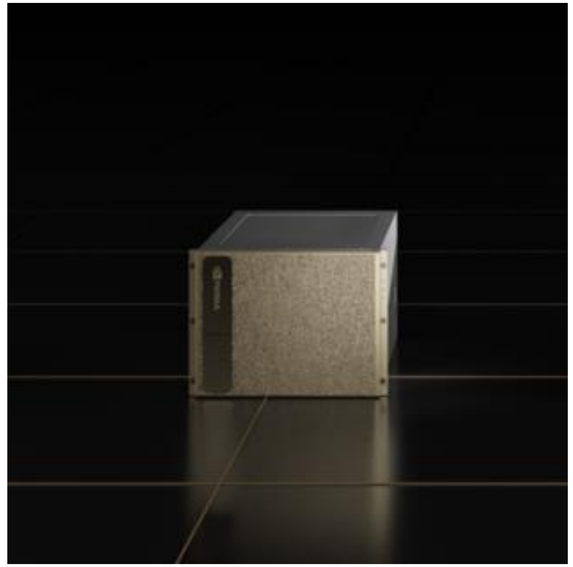
Blueprint for OEM and Cloud Partner Offerings

Cloud Native | Performance Isolation | Multi-Tenant

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

NVIDIA H100 on LaunchPad
Available Now

NVIDIA DGX H100
Order for Q1’23 Delivery



NVIDIA H100 Partner Systems
Shipping Starting Oct’ 22

NVIDIA H100 Cloud
Services Starting Early 2023

aws
Google Cloud
Microsoft Azure
ORACLE CLOUD

H100 Available at Every Scale

Get started with H100 today

NVIDIA Data Center GPU Portfolio

	GPU	Networking Solutions	DL Training & DA	DL Inference	HPC / AI	Omniverse / Render Farms	Virtual Workstation	Virtual Desktop (VDI)	Mainstream Acceleration	Far Edge Acceleration	AI-on-5G
Compute	H100	QTM2 SPTM4	SXM PCIE	SXM PCIE	SXM PCIE				PCIE		
	A100	QTM1 SPTM3	SXM PCIE A100X	SXM PCIE	SXM PCIE A100X				PCIE A100X		A100X
	A30	SPTM3		PCIE	PCIE				PCIE		A30X
Graphics / Compute	L40	SPTM4									
	A40	SPTM3									
	A10	SPTM3									
	A16	SPTM3									
Small Form Factor Compute/Graphics	A2	SPTM3									
	T4	SPTM3									

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

QTM1 Qauntum-1 IB switch plus BlueField2 DPUs or ConnectX-6/6 DX SmartNICs

QTM2 Qauntum-2 IB switch plus BlueField3 DPUs or ConnectX-7 SmartNICs

SPTM3 Spectrum-3 ethernet switch plus Bluefield2 DPUs or ConnectX-6 /6 Dx SmartNICs

SPTM4 Spectrum-4 ethernet switch plus Bluefield3 DPUs or ConnectX-7 SmartNICs

NVIDIA

Distinguishing Between Form Factors

Offering the right H100 configuration

GPU	LLM Training, Inference, HPC	
	Highest Performance	Mainstream Compute
H100 SXM	DGX, HGX 4-GPU, HGX 8-GPU	
H100 PCIe *(Includes NVIDIA AI Enterprise)*		

Good Better Best

Price-performance comparison for entire table

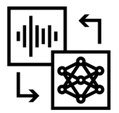
H100 and A100 Tensor Core GPUs

Both available to meet your data center needs

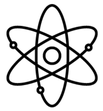
H100

PERFORMANCE

The Highest Performance



LLM
& Transformers

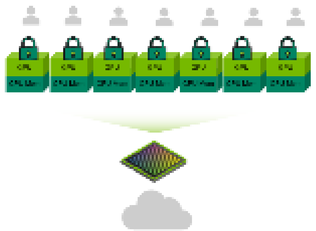


HPC

SECURITY

The Most Secure

Confidential
Computing



Secured MIG Instances

SCALABILITY

The Most Scalable

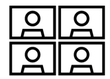
4th Gen NVLink
Quantum-2 IB

A100 & Ampere

FOR ALL OTHER NEEDS

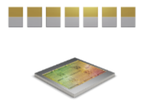
- Best for CNN models (RESNET-50, R-CNN…)

Non-Transformer AI



Mainstream
Acceleration

3rd Gen NVLink
Quantum-1 IB



Multi-instance
GPU

Includes NVIDIA AI Enterprise Software Suite and Support

5-Year Subscription and Support included for H100 PCIe

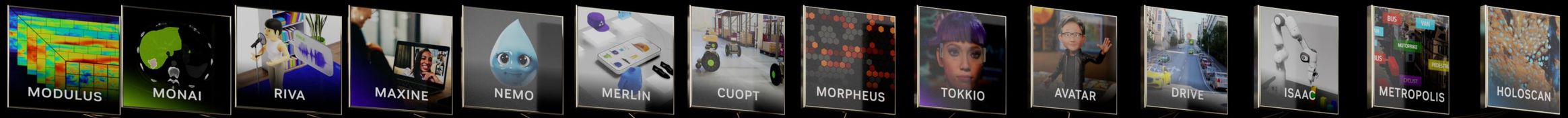
NVIDIA AI Enterprise
Sold Separately

DATA CENTER GPU COMPARISON - SEPT ‘22

	H100		A100		A30	A2	L40	A40	A10	A16
Design	Highest Perf AI, Big NLP, HPC, DA		High Perf Compute		Mainstream Compute	Entry-Level Small Footprint	Powerful Universal Graphics + AI	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 Gen4 2 Slot FHFL	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	300W	300W	150W	250W
FP64 TC | FP32 TFLOPS[2]	60 | 60	48 | 48	19.5 | 19.5		10 |10	NA | 4.5	NA | TBD[3]	NA | 37	NA | 31	NA | 4x4.5
TF32 TC | FP16 TC TFLOPS[2]	1000 | 2000	800 | 1600	312 | 624		165 | 330	18 | 36	TBD[3] | TBD[3]	150 | 300	125 | 250	4x18 | 4x36
FP8 TC | INT8 TC TFLOPS/TOPS[2]	4000 | 4000	4000 | 4000	NA | 1248		NA | 661	NA | 72	TBD[3] | TBD[3]	NA | 600	NA | 500	NA | 4x72
GPU Memory / Speed	80GB HBM3	80GB HBM2e	80GB HBM2e		24GB HBM2	16GB GDDR6	48GB 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)	3 Video Encoder 3 Video Decoder 4 JPEG Decoder	1 Video Encoder 2 Video Decoder (+AV1 decode)		4 Video Encoder 8 Video Decoder (+AV1 decode)
Ray Tracing	-		-			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	Top-of-Line	Best	Better	Good
vGPU	Yes						Yes*	Yes		
Hardware Root of Trust	Internal and External		Internal with Option for External				Internal	Internal with Option for External		
Confidential Computing	Yes		*(1)*		-	-	-	-	-	-
NVIDIA AI Enterprise	Add-on	Included	Add-on				Add-on			

1. *Supported on Azure NVIDIA A100 with reduced performance compared to A100 without Confidential Computing or H100 with Confidential Computing.*
2. *All Tensor Core numbers with sparsity. Without sparsity is ½ the value.*
3. *Precision TFLOP performance will be added in future update*


NVIDIA AI Platform



AI APPLICATION FRAMEWORK

PLATFORMS

NVIDIA HPC

NVIDIA AI

NVIDIA Omniverse

ACCELERATION LIBRARIES

cuNumeric | CV-CUDA | cuQuantum | Parabricks | Sionna | JetPack

RAPIDS | Spark | cuDNN | cuGraph | TensorRT | Triton | DeepStream | Flare

DOCA | Mag IO | Aerial

CLOUD-TO-EDGE DATACENTER-TO-ROBOTIC SYSTEMS

RTX | DGX | HGX | EGX | OVX | Super POD | AGX

3 CHIPS

GPU | CPU | DPU

NVIDIA AI

End-to-end open platform for production AI

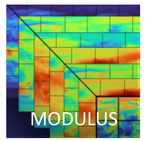
Application Workflows

Medical Imaging

Speech AI

Customer Service

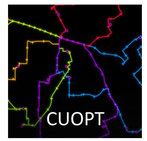
Recommenders



Physics ML

Video

Video Analytics



Logistics

Conversational AI

Robotics

Autonomous Vehicles

Cybersecurity

NVIDIA AI Enterprise

AI and Data Science Development and Deployment Tools

Cloud Native Management and Orchestration

Infrastructure Optimization

Accelerated Infrastructure

Cloud

Data Center

Edge

Embedded

NVIDIA LaunchPad

Hands-on Labs

Accelerating the Next Wave of AI

AI Platform Updates

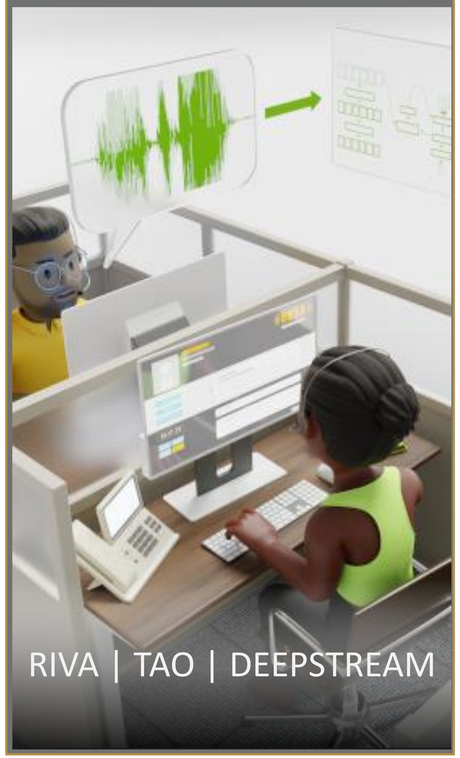
DATA SCIENCE
Analytics, ML, Visualization

RAPIDS | SPARK | cuOPT

RECOMMENDER
Personalization Engine, Simplified

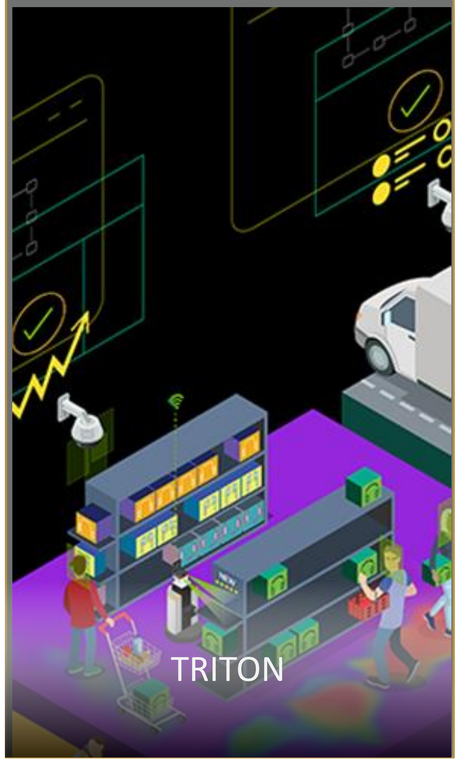
MERLIN

SPEECH & VISION
Conv AI, Video Analytics



RIVA | TAO | DEEPSTREAM

INFERENCE
Fast, Scalable Predictions



TRITON

NLP
Large Language Model

NEMO MEGATRON

Getting Started with NVIDIA AI

NVIDIA AI Enterprise trial programs

NVIDIA LaunchPad

- AI development and deployment trial program
- Deep dive, hands-on labs for AI practitioners and IT staff
- Requires ~8 hrs / Access for 2 wks

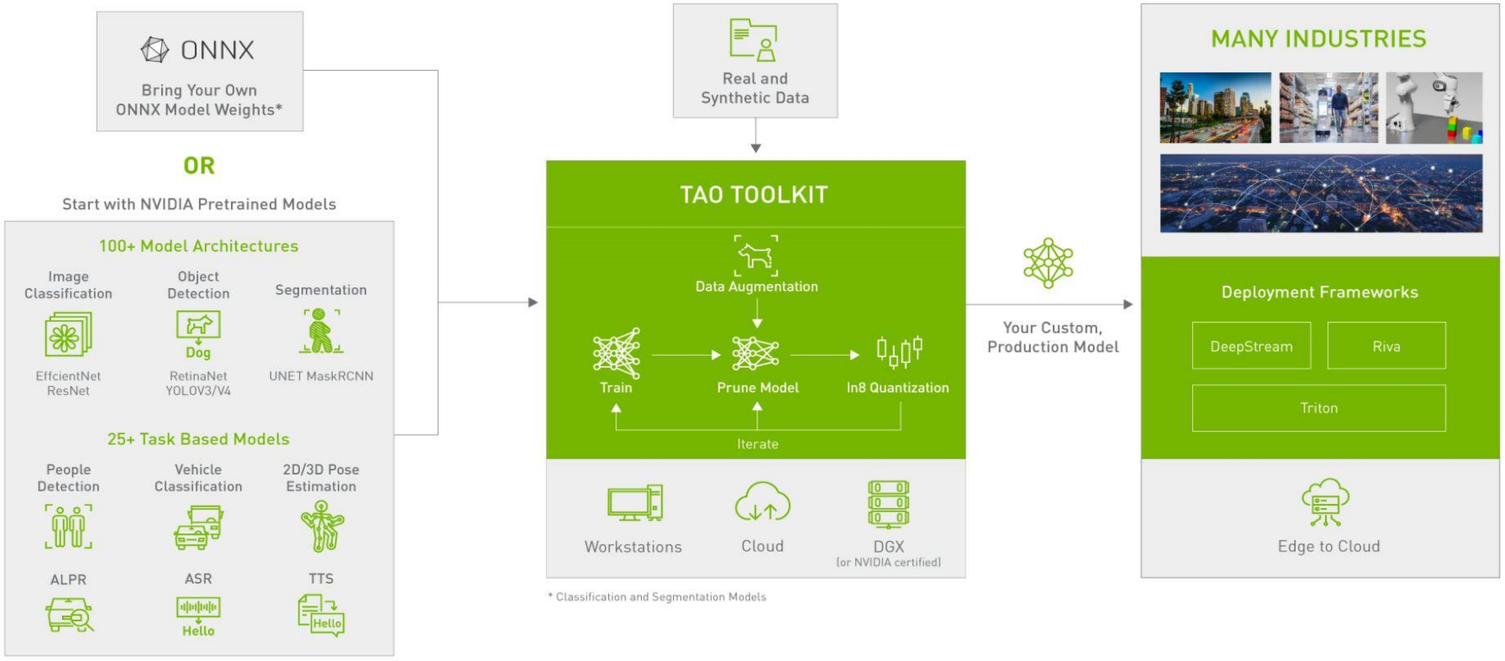
Evaluation Software

- Requirements: NVIDIA-Certified System,
- Free evaluation licenses for on premises POC
- 90 days to test and experience

NVIDIA TAO Toolkit

Create custom, production-ready AI models in hours rather than months

1. Bring your own model weights or choose from NVIDIA's library of model architectures or task-based models
2. Quickly train, adapt, and optimize models with your real or synthetic data
3. Integrate your customized models into your application and deploy



TRAIN EASILY

Fine tune NVIDIA pretrained models with fraction of the data

CUSTOMIZE FASTER

Built on TensorFlow and PyTorch that abstracts away the AI framework complexity

OPTIMIZE FOR DEPLOYMENT

Optimize for inference and integrate with Riva or DeepStream

SUPPORTED BY EXPERTS*

Supported by NVIDIA experts to help resolve issues from development to deployment

* Requires NVIDIA AI ENTERPRISE SUBSCRIPTION. Learn more here: https://www.nvidia.com/en-us/data-center/products/ai-enterprise/

Triton Inference Server

Open-source inference serving software for fast, scalable, simplified inference serving

Any Framework

Multiple DL/ML Frameworks e.g., TensorFlow, PyTorch, TensorRT, XGBoost, ONNX, Python & More

Multi-GPU Multi-Node Inference for Large Language Models

Any Query Type

Optimized for Real Time and Batch Requests

Audio & Video Streaming

Model Ensembles

Any Platform

X86 CPU | Arm CPU | NVIDIA GPUs | MIG

Linux | Windows | Virtualization

Public Cloud, Data Center and Edge/Embedded (Jetson)

DevOps/MLOps Ready

Microservice in Kubernetes, KServe

Available Across All Major Cloud AI Platforms

Integration with major MLOPS solutions

High Performance

Optimized for High GPU/CPU Utilization, Throughput & Low Latency

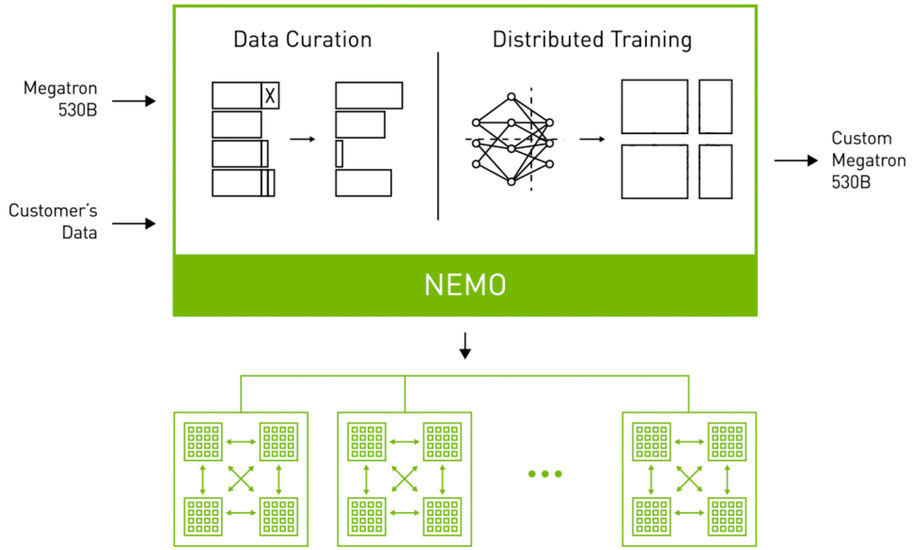
https://developer.nvidia.com/nvidia-triton-inference-server

NeMo Megatron

Accelerated Framework For Training Large Scale NLP Models

- Scale To Models with Trillions of Parameters
- Automated Data Curation for Training
- Pipeline, Tensor & Data Parallelism
- 20B Parameter Model in 1 month on DGX SuperPOD
- Optimized for DGX SuperPOD

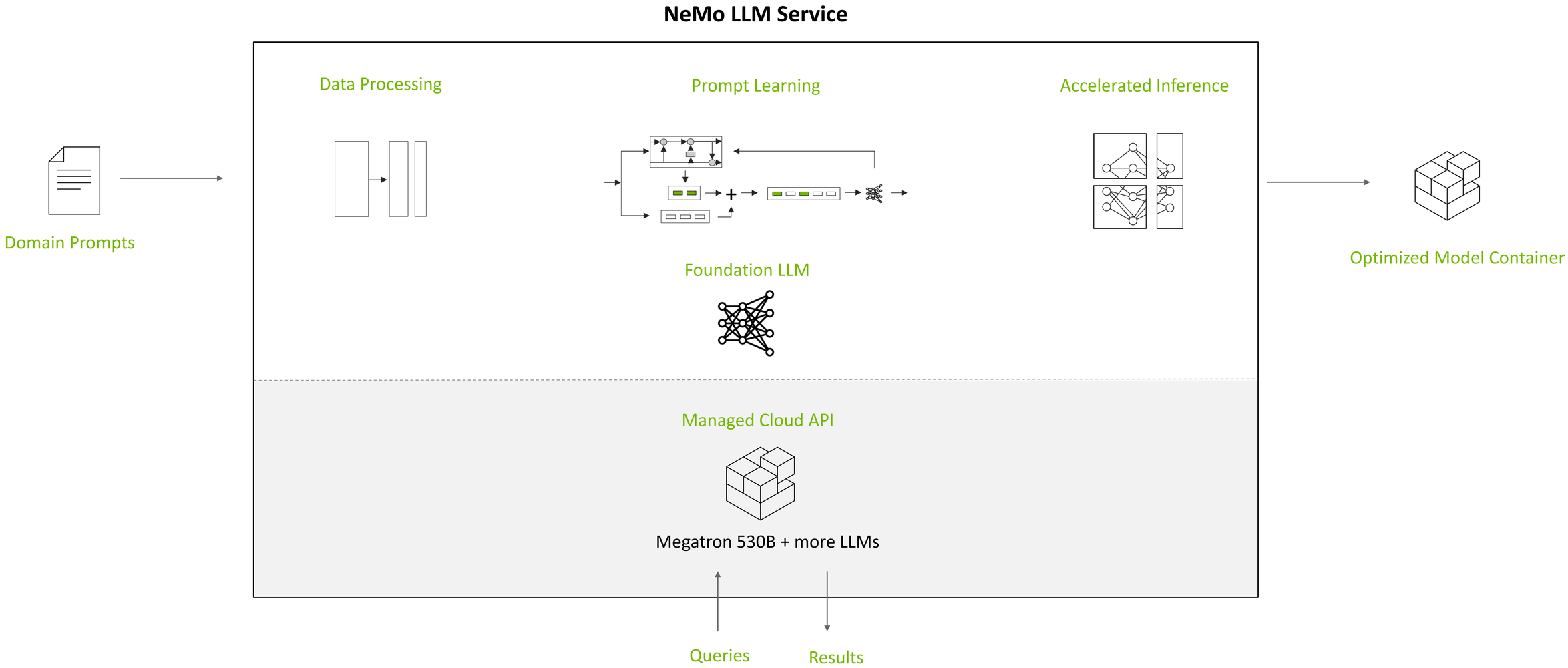
Sign up for Early Access



https://docs.nvidia.com/deeplearning/sdk/mixed-precision-training/index.html

Announcing NVIDIA NeMo LLM Service

Unlocking The Promise Of LLMs



Signup for EA now. EA starting in Oct ‘22

Develop, Customize and Deploy Large Language Models

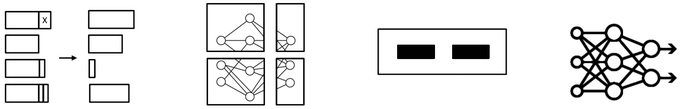
NeMo Megatron

Data Curation

Distributed Training

Prompt Learning

Accelerated Inference



Build, customize & deploy LLMs

NeMo LLM Service

Prompt Learning

NeMo Megatron 530B API

Customize and deploy LLMs or use API

BioNeMo Service

Biology Language Model Cloud API

API to predict biomolecular properties

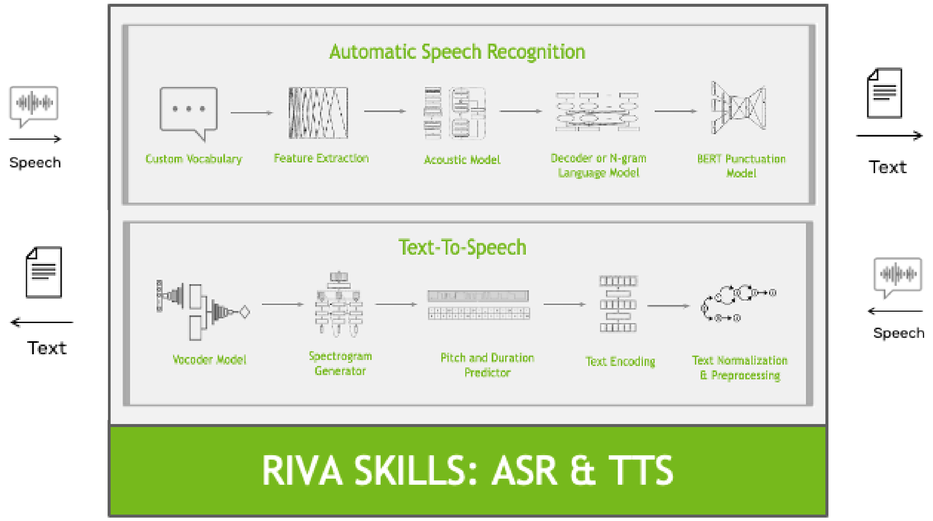
Framework - On-Prem or Cloud

Cloud Services

NVIDIA Riva

Empowering speech AI applications development and deployment

- **State-of-the-art pre-trained models** trained for 1M+hours on 70K+ hours of speech
- **Fully customizable** for achieving the best possible accuracy and voice expressivity
- **Real-time performance** far below 300 ms for interactive engaging conversations
- **Highly scalable** to hundreds of thousands of concurrent users
- **Runs everywhere**: on-premises, all clouds, edge, embedded



Riva enables speech AI workflows:

- Contact center agent assists, virtual assistants, and digital avatars
- Transcriptions and live captions
- Custom brand voices

Get started in GitHub or NVIDIA GPU Cloud, or try it in LaunchPad
https://www.nvidia.com/en-us/ai-data-science/products/riva

NGC – Portal To Enterprise Services, Software, Support

Artificial Intelligence | Metaverse | HPC

NGC
ngc.nvidia.com

Services

NeMo LLM service
LLM Customization | Managed API | Playground

Riva Studio
Custom voice | No-code GUI

BioNeMo service
Predict 3D protein structures, properties

Base Command
Training tools | Workflow mgmt | Monitoring, reporting

Fleet Command
Managed Edge AI | Layered Security | Monitoring, reporting

Enterprise Catalog
NVIDIA AI SW | Support | Proprietary SW

Software

Containers
Performance optimized

AI Models
Speech, CV, data science, more

Jupyter Notebooks
End-to-end example workflows

Helm Charts
Automate deployments on K8s

Support

NVIDIA AI Enterprise Support
Business Standard | Business critical |Dedicated TAM

ON-PREM | CLOUD | EDGE


NVIDIA