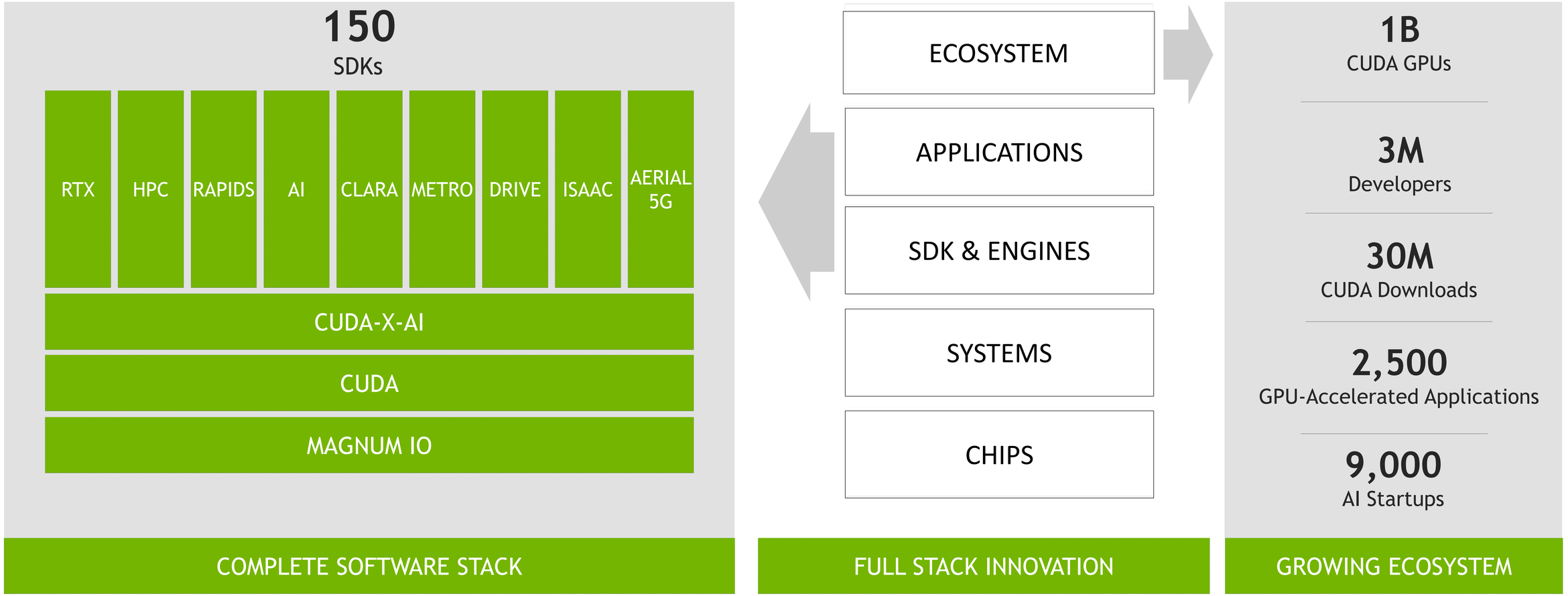
NVIDIA DATACENTER PLATFORM

SPEAKER NAME, TITLE

NVIDIA IS A FULL STACK COMPUTING PLATFORM

Amazing Innovation and Expansion of NVIDIA Ecosystem



65 updates from last GTC

NVIDIA DATA CENTER PLATFORM

HGX
AI SUPERCOMPUTING PLATFORM

EGX
MAINSTREAM ACCELERATED COMPUTING PLATFORM

WORKLOADS

HPC | DATA CENTER AI TRAINING & INFERENCE | DATA ANALYTICS & MACHINE LEARNING

DATA CENTER AI TRAINING & INFERENCE | EDGE AI INFERENCE | DATA ANALYTICS & MACHINE LEARNING | PROFESSIONAL VISUALIZATION | TRADITIONAL APPLICATIONS

SOFTWARE

Ecosystem of Accelerated Applications and Frameworks

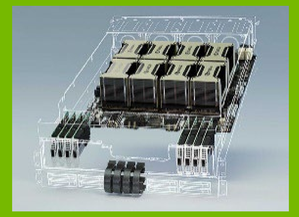
CUDA-X-AI

MAGNUM IO | DOCA | CUDA

Orchestration and Management Integrations

Bare Metal | Containers | Virtualization (NVIDIA vGPU)

HARDWARE



HGX Server

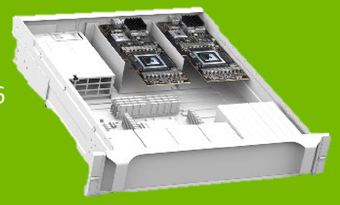
- HGX A100 (8- and 4-way baseboards)
- A100 PCIe GPUs in dense 4 GPU+ Configs
- ConnectX-6, ConnectX-6 Dx, BlueField-2

Partner Servers

NVIDIA CERTIFIED

EGX Platform

- A100 PCIe, A30, A40, A2, A16 GPUs
- ConnectX-6, ConnectX-6 Lx, BlueField-2



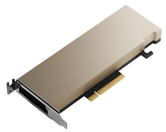
NVIDIA

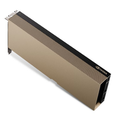
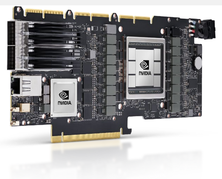

CURRENT NVIDIA DC PRODUCT OFFERING FOCUS

DATA CENTER PRODUCT COMPARISON

	Optimized for Compute			Optimized for Graphics		Converged Accelerators for Compute	
	A100	A30	A2	A40	A16	A100X	A30X
Design	Highest Perf Compute	Mainstream Compute	Entry-Level Small Footprint	High Perf Graphics	High Density Virtual Desktop	High Perf Converged Accelerator	Mainstream Converged Accelerator
Max Power	300W	165W	40-60W	300W	250W	300W	230W
Form Factor	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 1 NVLink Bridge	x16 PCIe Gen4 2 Slot FHFL	x16 PCIe Gen4 2 Slot FHFL 3 NVLink Bridge	x16 PCIe Gen4 2 Slot FHFL 1 NVLink Bridge
GPU Memory	80GB HBM2e	24GB HBM2	16GB GDDR6	48GB GDDR6	4x 16GB GDDR6	80GB HBM2e	24GB HBM2e
Multi-Instance GPU (MIG)	Up to 7	Up to 4	-	-	-	Up to 7	Up to 4
Media Acceleration	1 JPEG Decoder 5 Video Decoder	1 JPEG Decoder 4 Video Decoder	1 Video Encoder 2 Video Decoder (+AV decode)	1 Video Encoder 2 Video Decoder (+AV1 decode)	4 Video Encoder 8 Video Decoder (+AV1 decode)	1 JPEG Decoder 5 Video Decoder	1 JPEG Decoder 4 Video Decoder
Ray Tracing	-		Yes	Yes	Yes	-	
Fast FP64	Yes		-	-	-	Yes	
Graphics	For in-situ visualization (no NVIDIA vPC or RTX vWS)		Good	Best	Better	For in-situ visualization (no NVIDIA vPC or RTX vWS)	
vGPU	Yes		Yes	Yes	Yes	Yes	
Hardware Root of Trust	Yes		Yes	Yes	Yes	Yes	
Integrated DPU	-	-	-	-	-	BlueField-2	
Server Availability	In Production	In Production	Q1 ’22	In Production	In Production	Q1 ‘22	



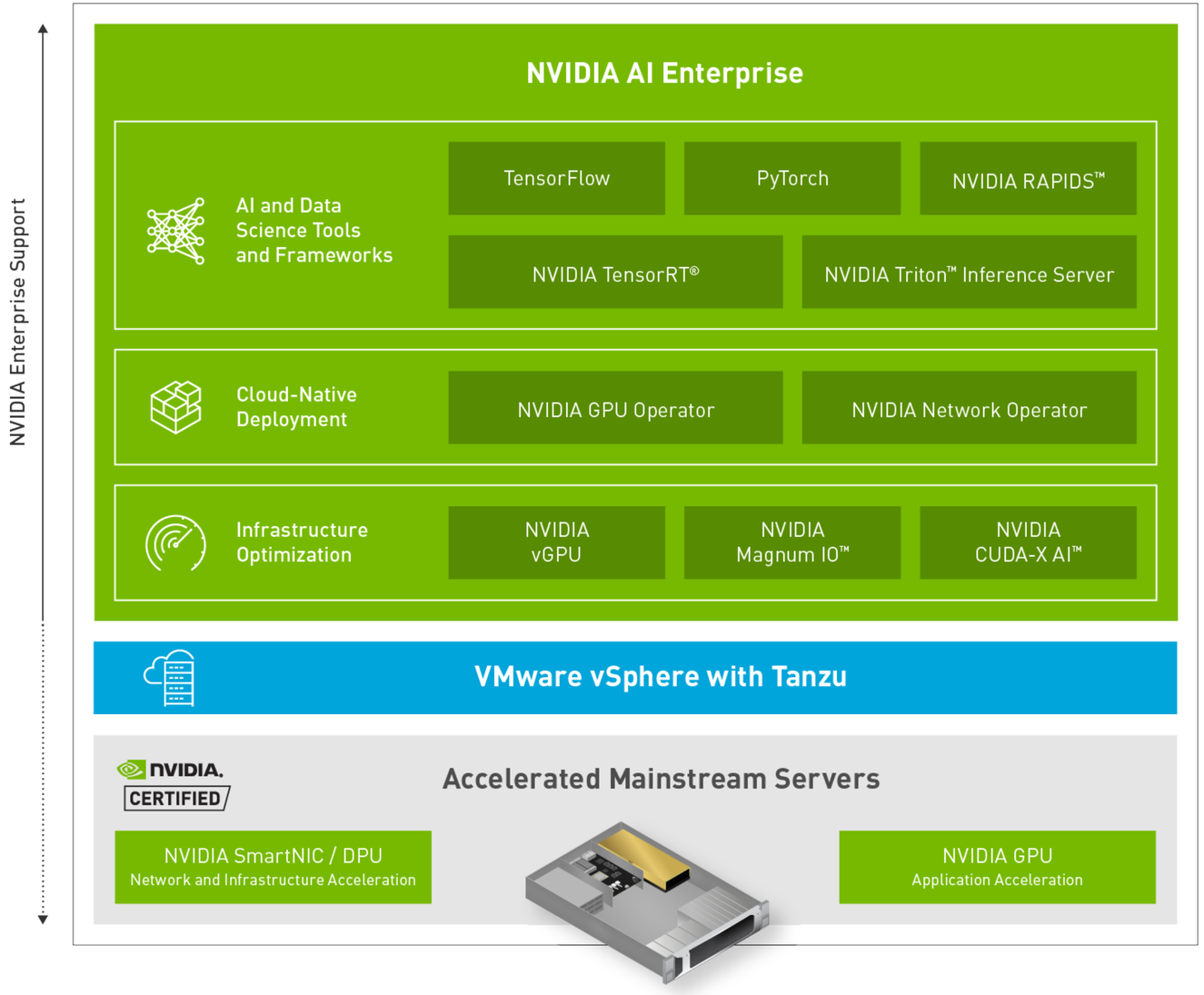


NVIDIA

NVIDIA AI ENTERPRISE SOFTWARE SUITE

Optimized, Certified, and Supported on VMware vSphere 7



NVIDIA DGX SYSTEMS

Setting the bar for enterprise AI

9 OF THE **TOP 10** GLOBAL UNIVERSITIES

7 OF THE **TOP 10** US HOSPITALS

6 OF THE **TOP 10** US BANKS

7 OF THE **TOP 10** GLOBAL CAR MANUFACTURERS

8 OF THE **TOP 10** GLOBAL TELCOS

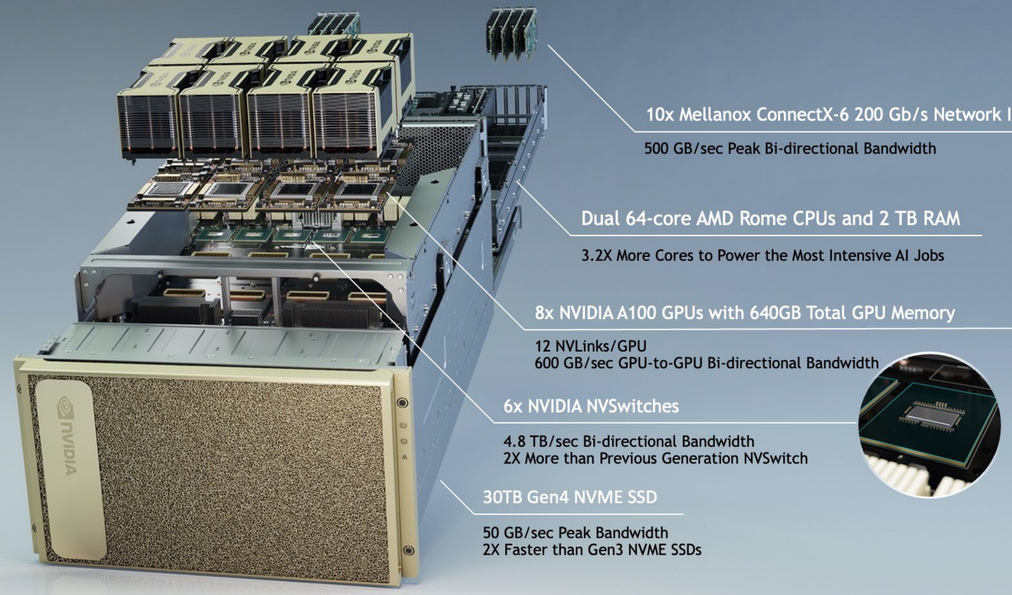
10 OF THE **TOP 10** US GOVERNMENT INSTITUTIONS

7 OF THE **TOP 10** CONSUMER INTERNET COMPANIES

10 OF THE **TOP 10** AEROSPACE AND DEFENSE COMPANIES

NVIDIA DGX A100

The universal system for AI infrastructure



NVIDIA DGX STATION A100

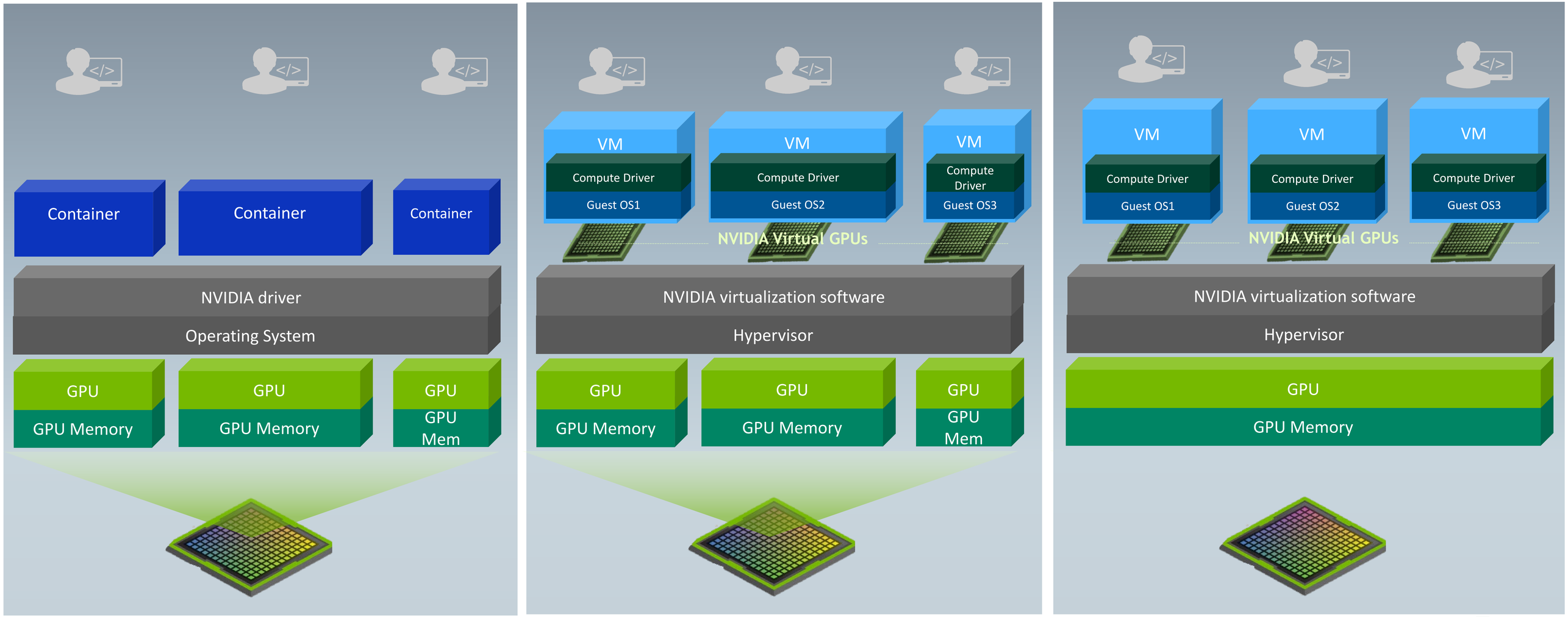
Workgroup appliance for the age of AI

- NVIDIA GPUs
  - 4x NVIDIA A100 GPUs with up to 320GB total GPU memory
  - 3rd generation NVLink
- CPU and Memory
  - 64-core AMD Epyc CPU, PCIe Gen4
  - 512GB system memory
- Internal Storage
  - NVME M.2 SSD for OS, NVME U.2 SSD for data cache
- Connectivity
  - 2x 10GbE (RJ45)
  - 4x Mini DisplayPort for display out
  - Remote management 1GbE LAN port (RJ45)

NVIDIA

Combining MIG and vGPU on NVIDIA GA100

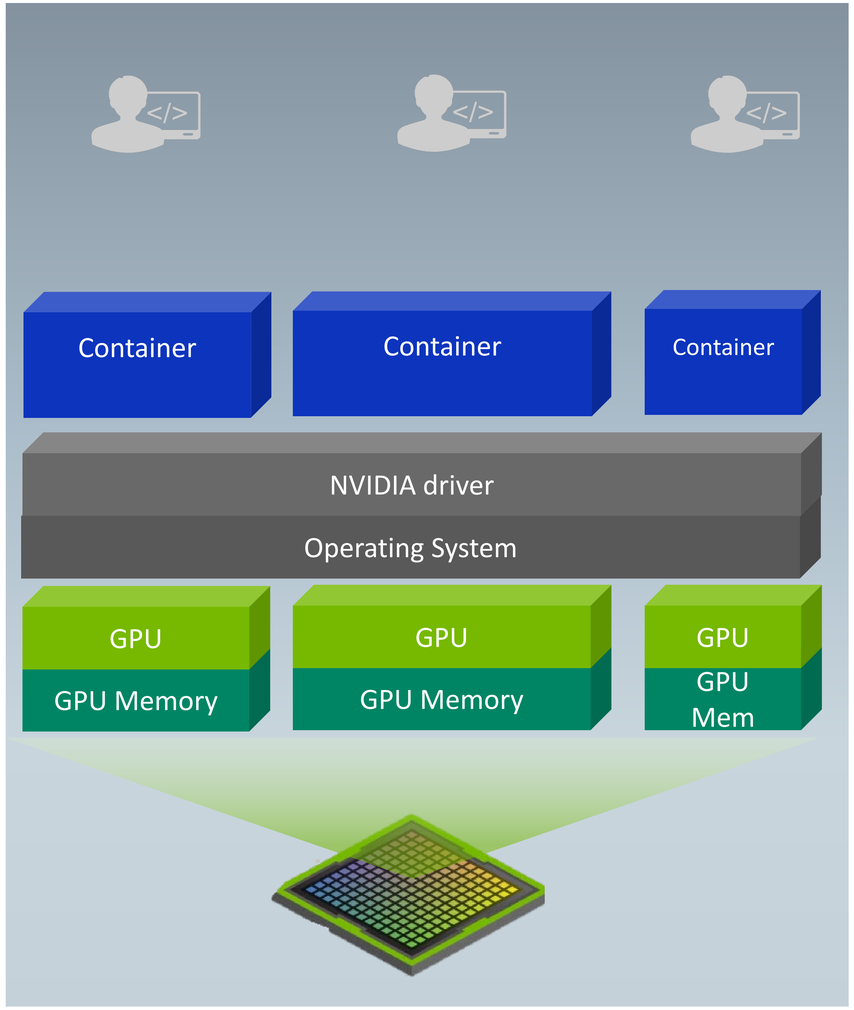
NVIDIA Virtual Compute Server Offers Flexibility with MIG Enabled or MIG Disabled



Combining MIG and vGPU on NVIDIA GA100

MIG only Mode - Spatial Partitioning

MIG
Spatial Partitioning



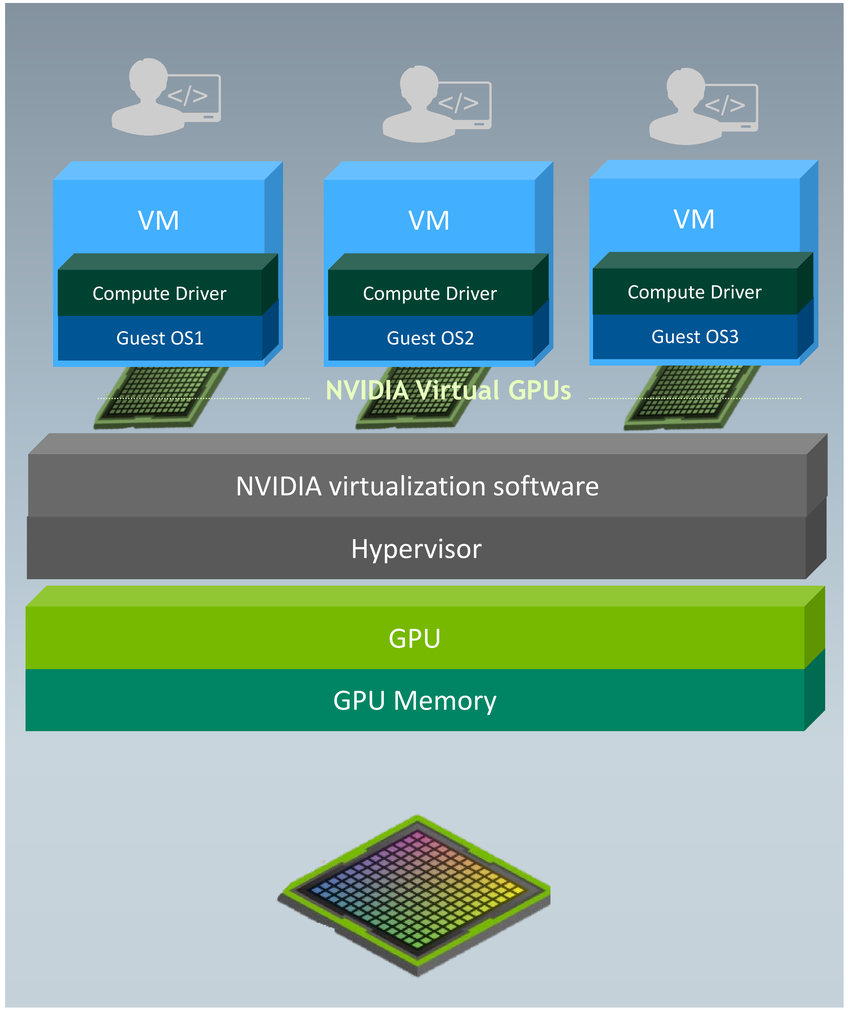
- GPU sharing with spatial partitioning
- All containers/apps share the same OS kernel
- 1 Container per MIG device
- Reboot/GPU reset required to toggle MIG Mode
- Max 7 MIG Devices per GPU
- 1:1 – 1 App/Container per MIG device
- Fully isolated and QoS instances at the hardware level with dedicated high-bandwidth memory, cache, and compute cores

Combining MIG and vGPU on NVIDIA GA100

vGPU only Mode - Temporal Partitioning

- Fractional GPU with temporal partitioning
- VM reboot required to configure vGPUs
- Up to 20 VMs on a single GPU*
- Fully isolated including OS kernel
- Homogenous profile sizes ranging from 4 GB (4C) to 80 GB (80C*)
- Max multi-tenancy 20:1*
- Live Migration
- Configurable scheduler

vGPU
Temporal Partitioning



* Available with NVIDIA A100 80GB GPU

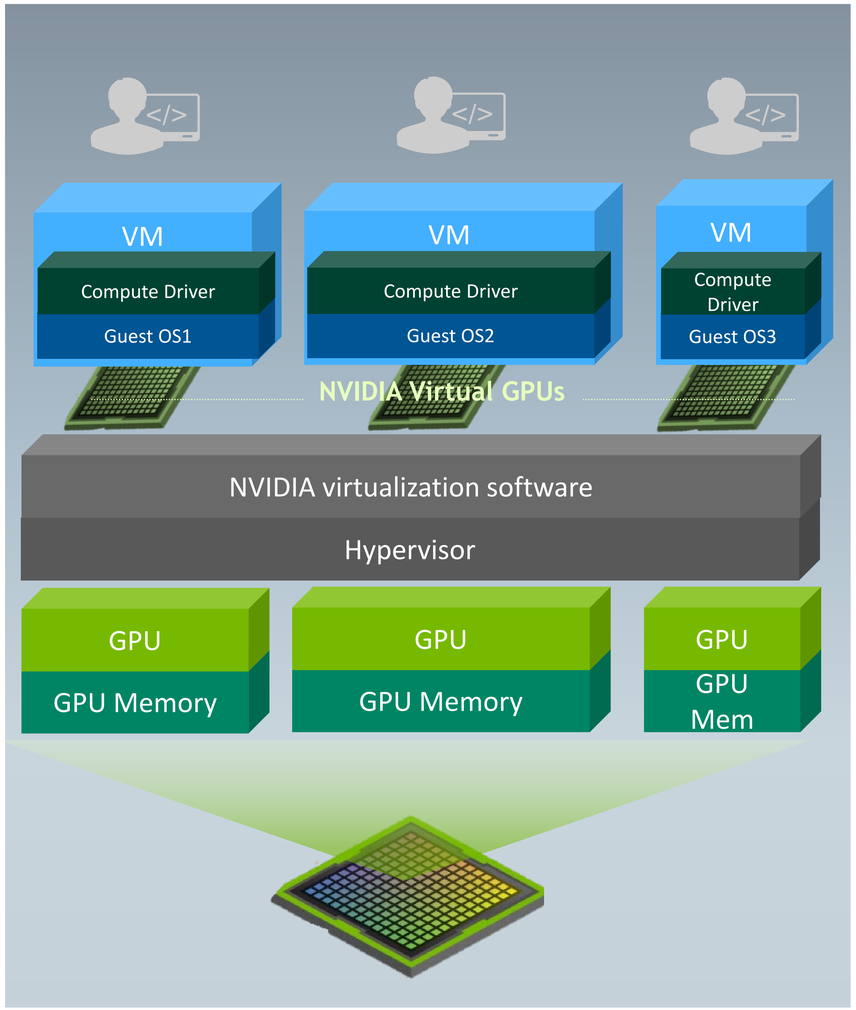
Combining MIG and vGPU on NVIDIA GA100

vGPU with MIG - Spatial Partitioning

MIG

- GPU sharing with spatial partitioning
- ~~All containers/apps share the same OS kernel~~
- ~~1 Container per MIG instance~~
- Reboot/GPU reset required to toggle MIG Mode
- Max 7 MIG Devices per GPU
- ~~1:1 – 1 App/Container per MIG device~~
- **1:1 – 1 VM per MIG Device**
- Fully isolated and QoS instances at the hardware level with dedicated high-bandwidth memory, cache, and compute cores

vGPU + MIG, Spatial (& Temporal) Partitioning



vGPU

- ~~Fractional GPU with temporal partitioning~~
- ~~No reboot/GPU reset required to configure partitions~~
- ~~Up to 20 VMs on a single GPU*~~
- ~~Homogenous profile sizes ranging from 4 GB (4C) to 80 GB (80C*)~~
- Fully isolated including OS kernel
- Heterogeneous profile sizes based on MIG instance configuration
- ~~Max multi-tenancy 20:1*~~
- **Max multi-tenancy 7:1**
- Live Migration
- ~~Configurable scheduler~~

NVIDIA

* Available with NVIDIA A100 80GB GPU

NVIDIA Cloud Native Technology

GPU Operator for Kubernetes-based Orchestration

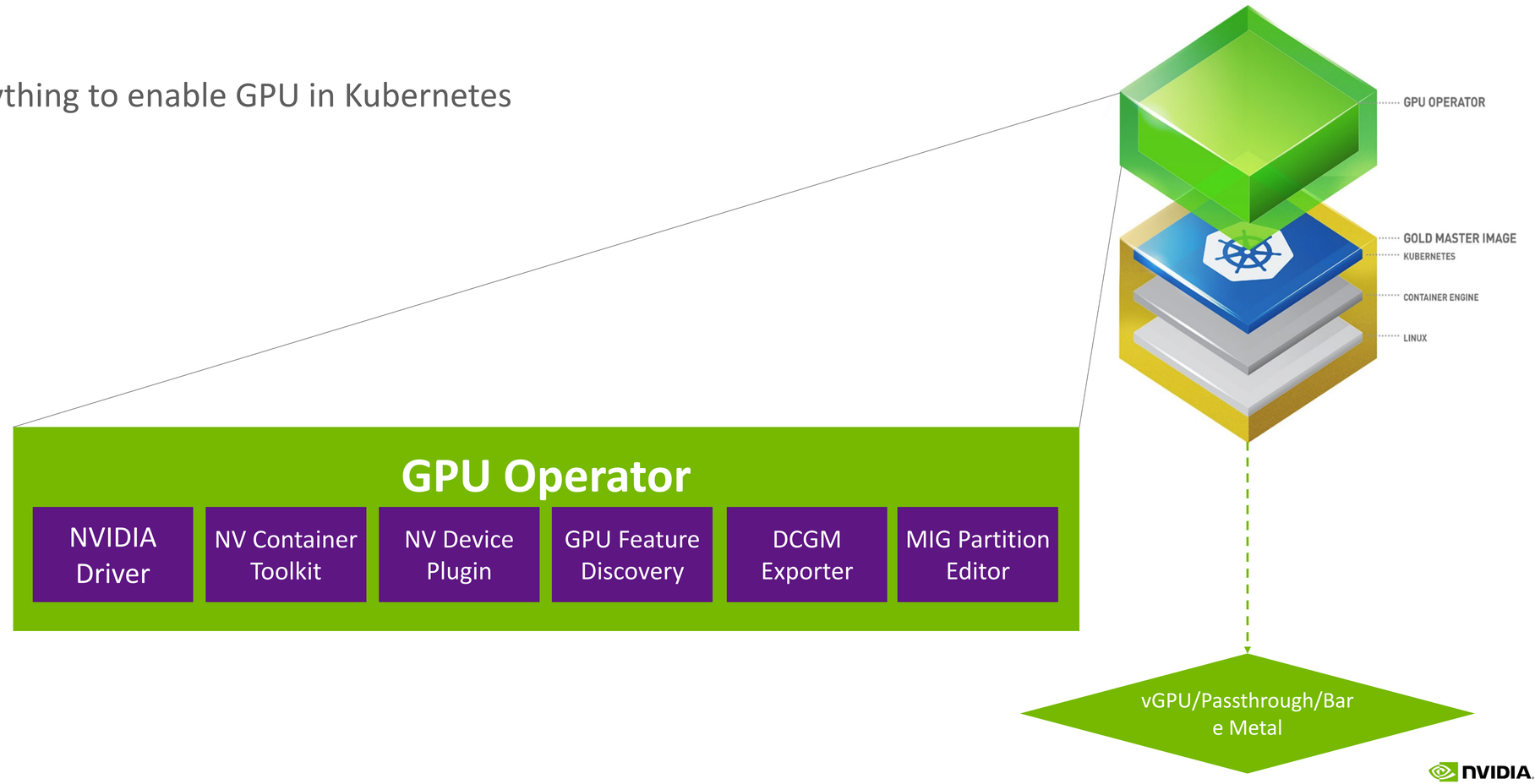
Single package that includes everything to enable GPU in Kubernetes

Includes

- NVIDIA Driver
- Container Toolkit
- Device Plugin
- Feature Discovery
- DCGM exporter
- Mig-parted

Supports

- NVIDIA GPUs
- MIG-enabled GPUs
- vGPUs

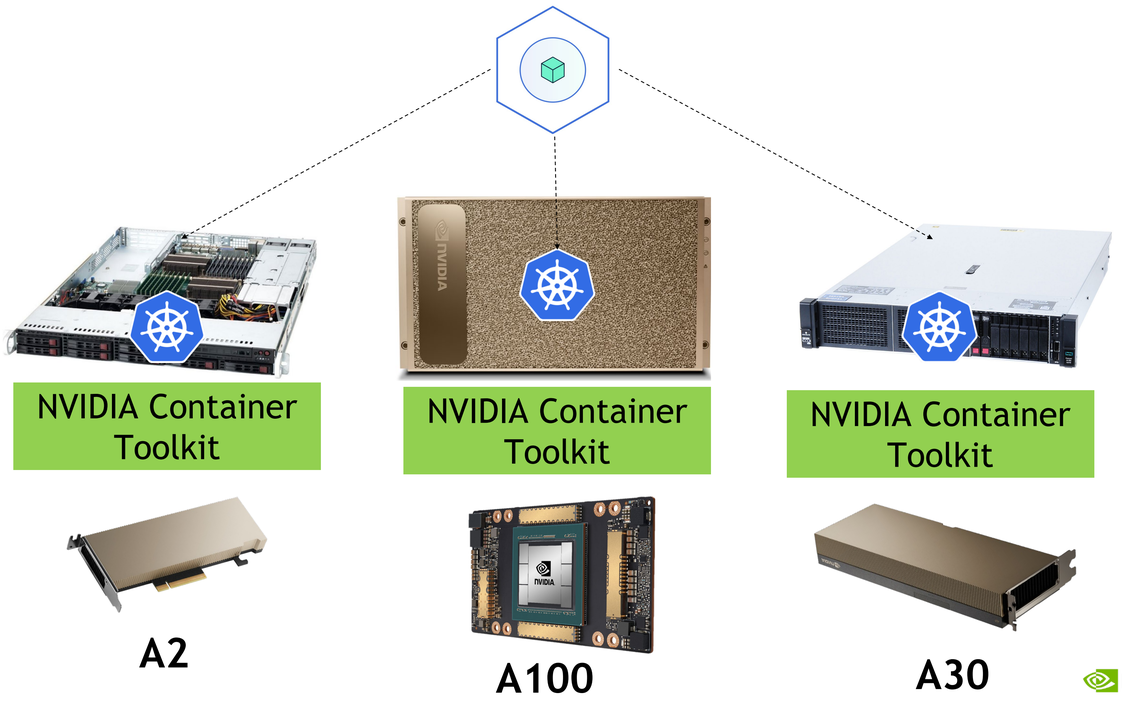


NVIDIA Cloud Native Technology

Accessing GPU resources from Kubernetes-based Orchestration

```
apiVersion: v1
kind: Pod
metadata:
  name: gpu-example
spec:
  containers:
    - name: gpu-example
      image: nvidia/cuda
      resources:
        limits:
          nvidia.com/gpu: 4
  nodeSelector:
    nvidia.com/gpu.product: A100-SXM4-80GB
    nvidia.com/cuda.runtime: 11.5
    nvidia.com/cuda.driver: 495.29.05
```

- **k8s-device-plugin**
- **gpu-feature-discovery**

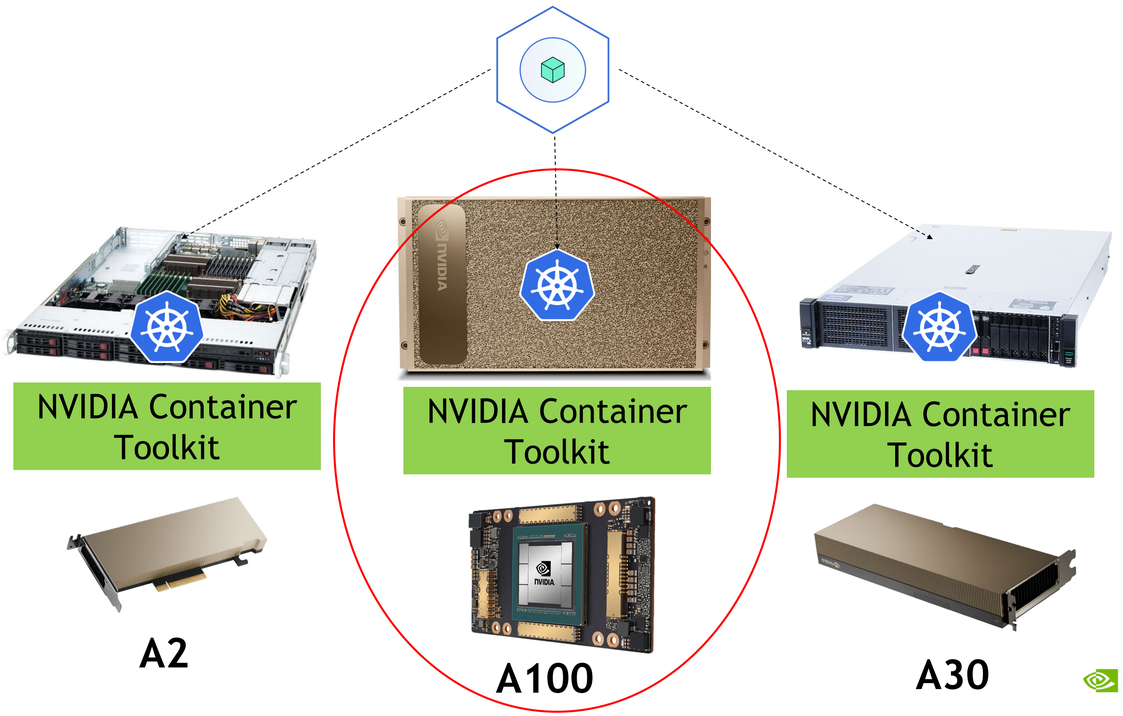


NVIDIA Cloud Native Technology

Specifying GPU resources from Kubernetes-based Orchestration

```
apiVersion: v1
kind: Pod
metadata:
  name: gpu-example
spec:
  containers:
    - name: gpu-example
      image: nvidia/cuda
      resources:
        limits:
          nvidia.com/gpu: 4
  nodeSelector:
    nvidia.com/gpu.product: A100-SXM4-80GB
    nvidia.com/cuda.runtime: 11.5
    nvidia.com/cuda.driver: 495.29.05
```

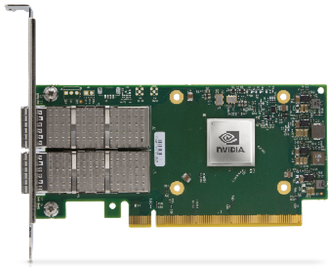
- **k8s-device-plugin**
- **gpu-feature-discovery**



NVIDIA NETWORKING

NVIDIA ConnectX

World's Leading Ethernet NICs



- High-Performance, Multi-Purpose SmartNIC
- All Speeds from 10Gb/s to 200Gb/s Ethernet Connectivity
- Software-Defined, Hardware-Accelerated Networking

NVIDIA BlueField

World's Most Advanced Data Center Infrastructure on-a-Chip

- Fully programmable DPU for Accelerated Networking, Storage and Security
- Powerful Arm, Advanced Hardware Accelerations
- 200Gb/s Ethernet and InfiniBand

NVIDIA SPECTRUM

World's Leading Open Ethernet Switches

- Built for Scale
- Easiest AI Config
- Best in Class Telemetry
- Highest Operational Efficiency
- Highest Performance Fair/predictable QOS

NVIDIA QUANTUM

World's Highest Performance HPC & AI InfiniBand Networking

- ConnectX adapters and Quantum switches
- HDR 200G & NDR 400G
- Full transport offload
- In-Network Computing
- RDMA, GPU Direct, GDS
- Adaptive routing, congestion control and quality of service

NVIDIA LINKX

World's Most Reliable Optical Transceivers, AOCs & Copper Cables

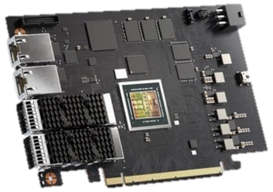
- Unmatched Quality
- Copper Direct Attach Cables (DAC)
- DAC splitter cables & adapters
- Active Optical Cables Multi-mode and single-mode transceivers

NVIDIA-CERTIFIED SYSTEMS

Simplifies Deployment of Accelerated Computing at Scale

SYSTEM DESIGN OPTIONS

NVIDIA SERVER GPUs



NVIDIA SMARTNICs AND DPUs

LEADING PARTNER SERVERS

NVIDIA WORKSTATION GPUs

LEADING PARTNER LAPTOPS AND DESKTOPS

NVIDIA CERTIFIED

Validates the Best Baseline Configuration for

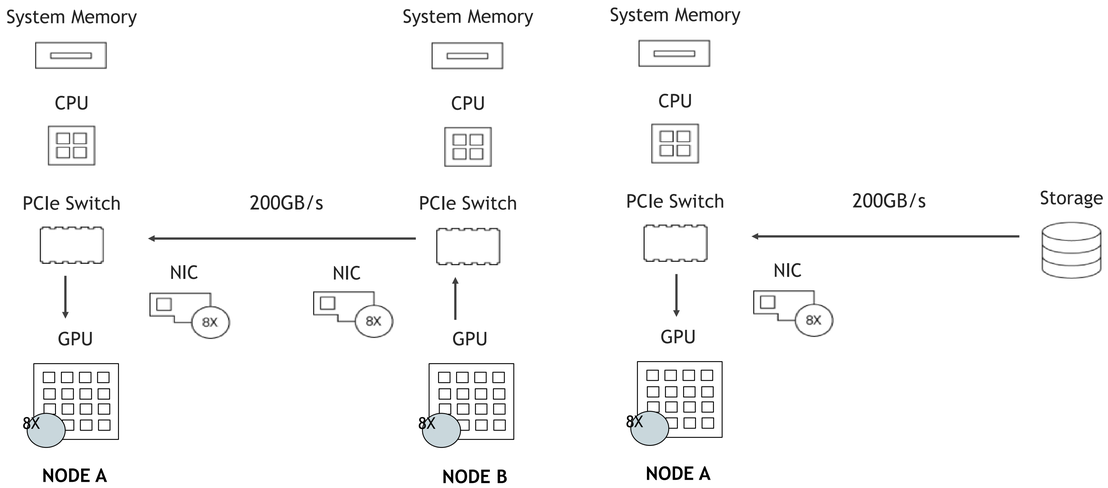
PERFORMANCE

MANAGEABILITY

SECURITY

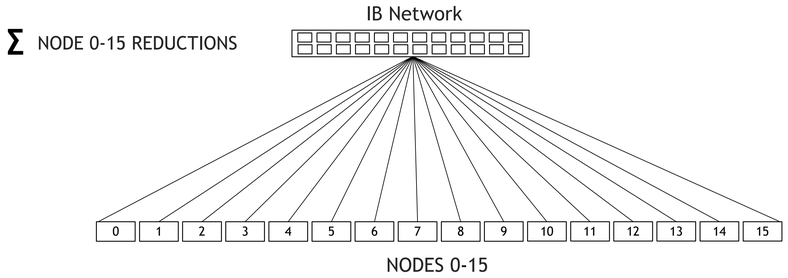
SCALABILITY

AI SUPERCOMPUTING NEEDS EXTREME IO

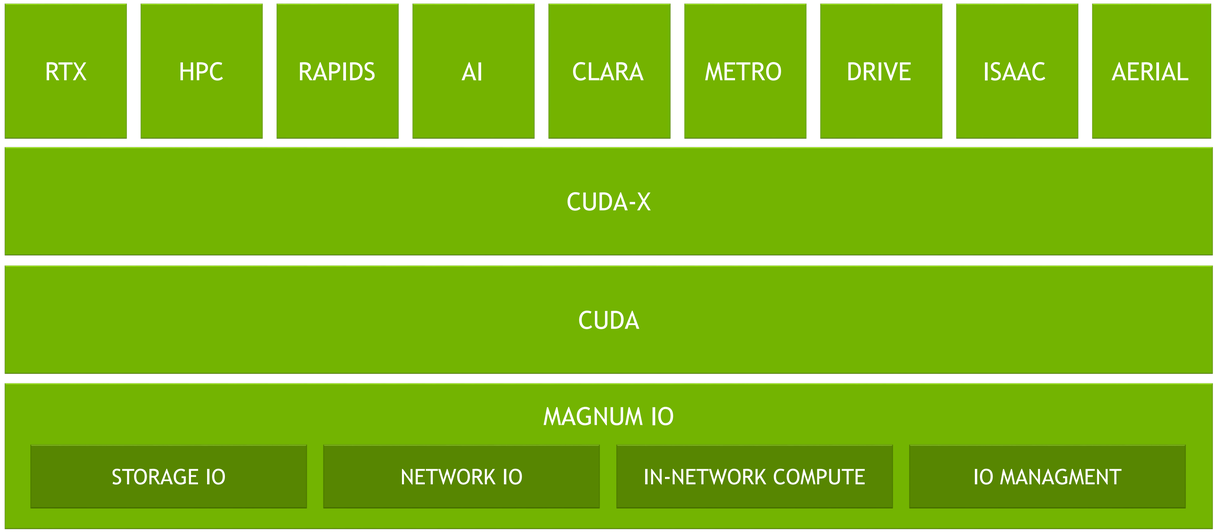


GPUDIRECT RDMA

GPUDIRECT STORAGE



SHARP IN-NETWORK COMPUTING



2X DL Inference Performance

10X IO Performance | 6X Lower CPU Utilization

ddn · DELL EMC · Hewlett Packard Enterprise · HITACHI Inspire the Next · IBM · KIOXIA · LIQID · Micron · PAVILION · SAMSUNG · ScaleFlux · VAST · NetApp · WEKA · Western Digital

NVIDIA


NVIDIA AI AND DATA SCIENCE SOFTWARE OFFERINGS

NVIDIA AI PLATFORM LEADERSHIP

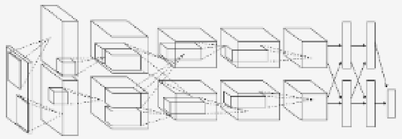
#1

MLPerf

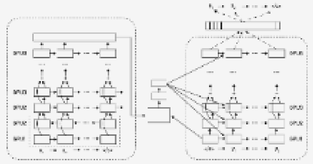
AI Training & Inference

Performance
#1 MLPerf

CNNs



RNNs



GANs

•••

Versatility
Runs Everything

HugeCTR · mxnet · ONNX

PaddlePaddle · PYTORCH

RAPIDS · Spark · TensorFlow

CUDA-X-AI

CUDA

MAGNUM IO

Productivity
Mature Software Stack

Alibaba Cloud · aws

Google Cloud · Microsoft Azure

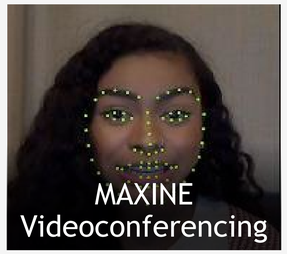
•••

DELL · Hewlett Packard Enterprise

inspur · Lenovo

•••

Available Everywhere
Every Cloud and OEM



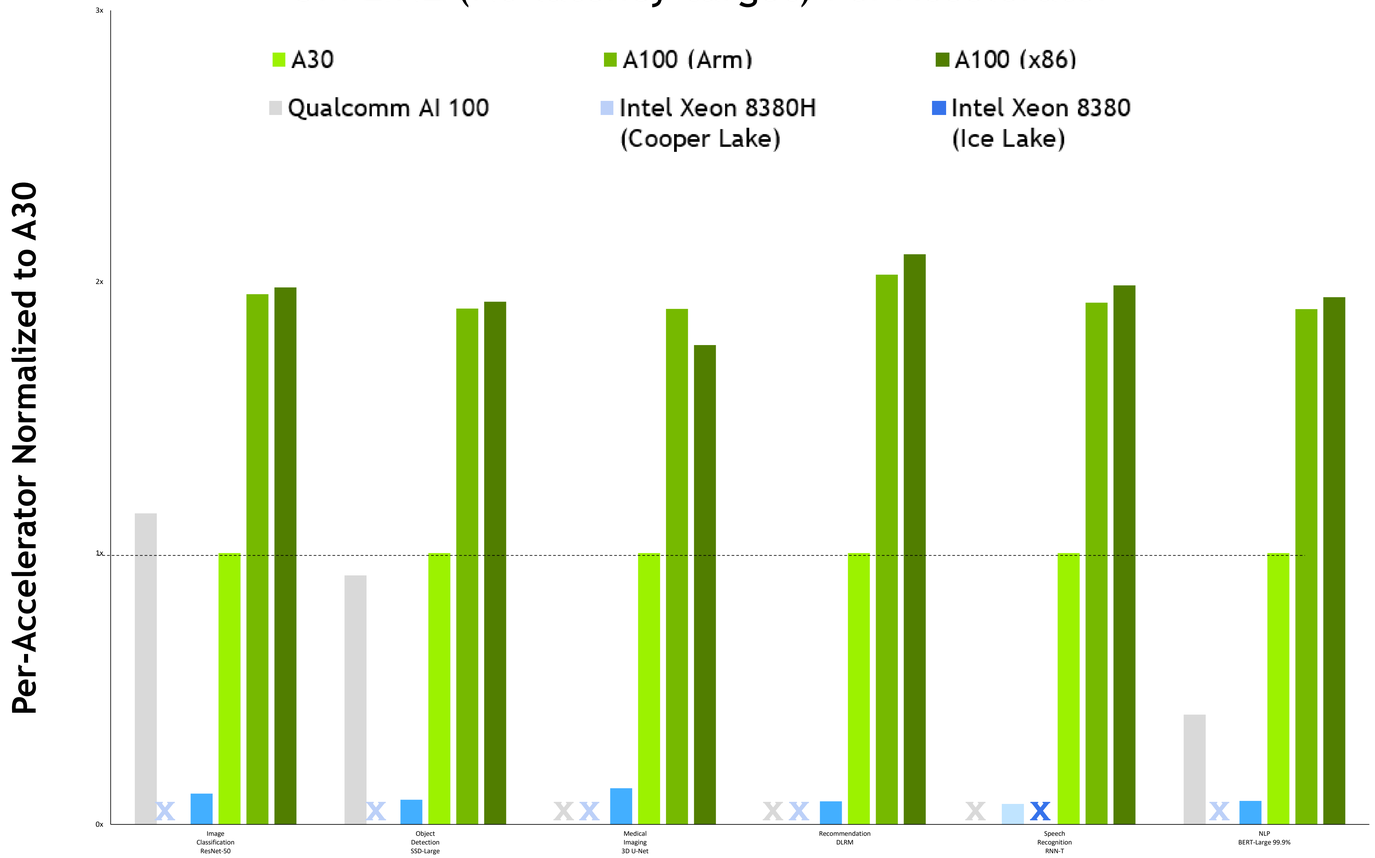
•••

Fastest to Production
Ready Application Frameworks

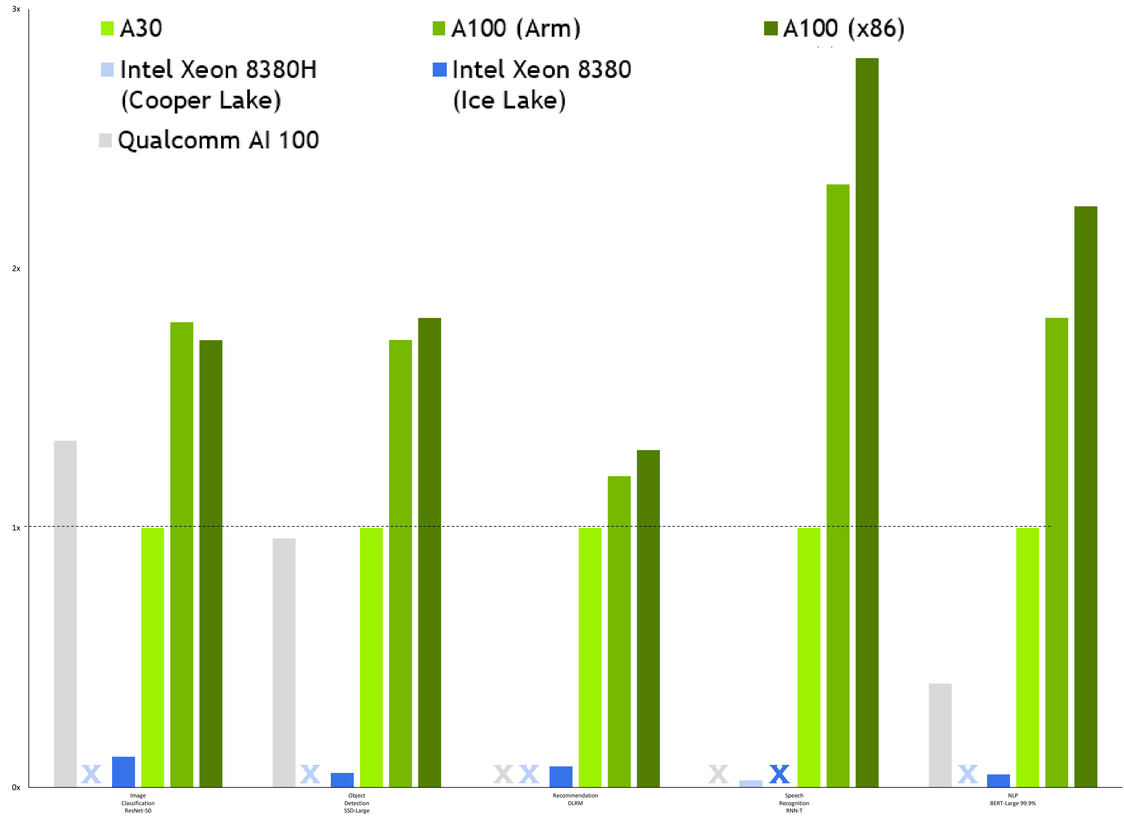
NVIDIA TOPS MLPERF DATA CENTER BENCHMARKS

A100 up to 104x Faster Than CPU

OFFLINE (No latency target) Per Accelerator



SERVER (w/ latency target) Per Accelerator



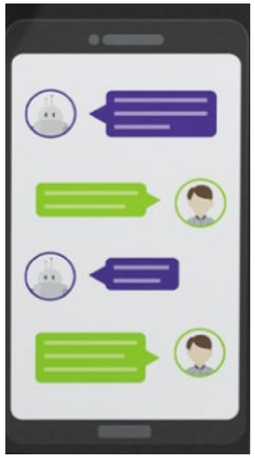
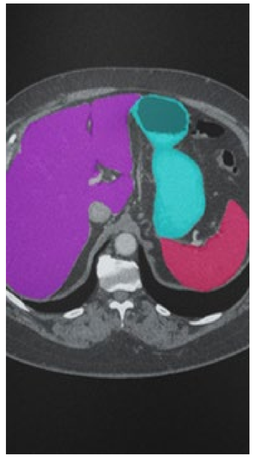
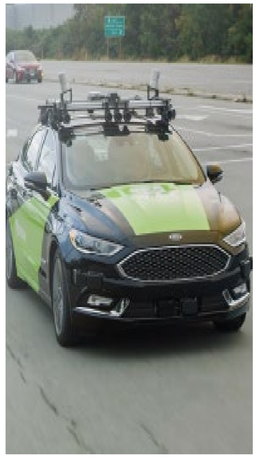
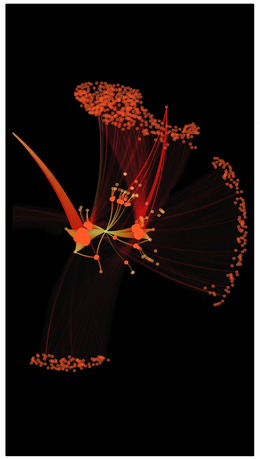
MLPerf v1.1 Inference Closed; Per-accelerator performance derived from the best MLPerf results for respective submissions using reported accelerator count in Data Center Offline and Server. Qualcomm AI 100: 1.1-057 and 1.1-058, Intel Xeon 8380: 1.1-023 and 1.1-024, Intel Xeon 8380H 1.1-026, NVIDIA A30: 1.1-43, NVIDIA A100 (Arm): 1.1-033, NVIDIA A100 (X86): 1.1-047. MLPerf name and logo are trademarks. See www.mlcommons.org for more information.

X = No result submitted

NVIDIA.

ENABLING ENTERPRISE TRANSFORMATION WITH AI

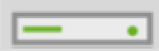
End to End Application Frameworks

Speech AI	Big NLP	Recommender Systems	Smart Cities	Healthcare	Robotics	Autonomous Vehicles	Telecom	Cybersecurity
								
Riva	NeMo	Merlin	Metropolis	Clara	Isaac	Drive	Aerial	Morpheus

Desktop Development

Data Center Solutions

Accelerated Edge



Supercomputers

GPU-Accelerated Cloud

NVIDIA TAO

AI-Model-Adaptation Framework

Build models easily with no AI expertise

Create custom, production-ready models in hours, rather than months with fraction of data, as opposed to training from scratch

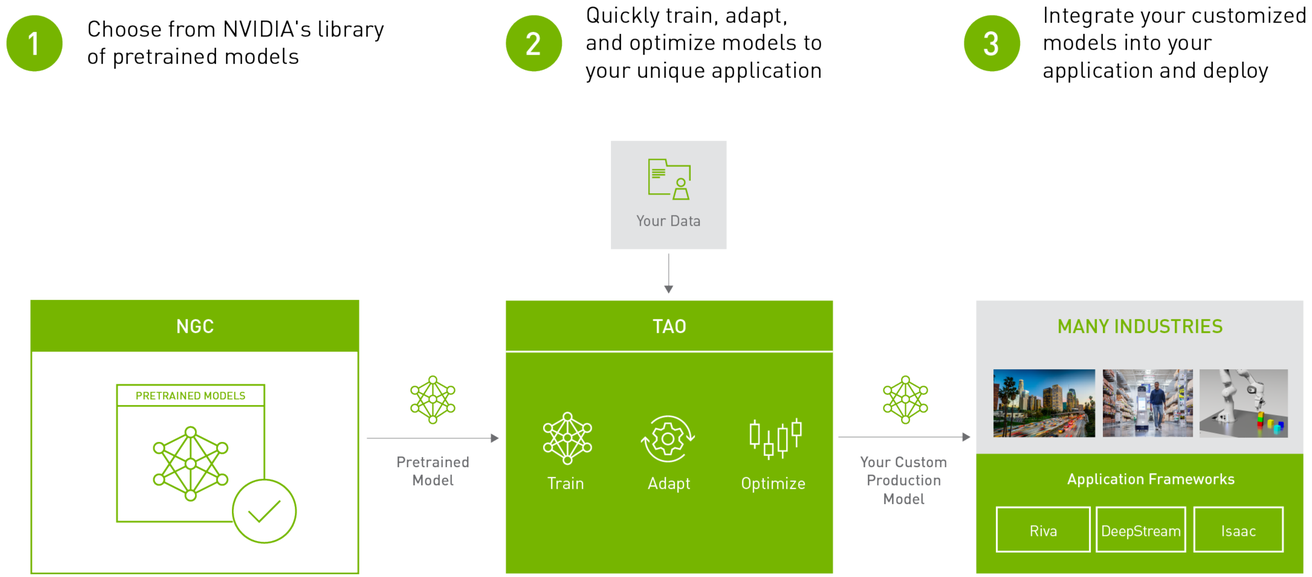
Optimize models for throughput and latency

Easily integrate models into DeepStream and Riva

Product Updates:

TAO Toolkit (CLI Version) - New Version Available Now
- New 2D/3D action recognition
- New text-to-speech models

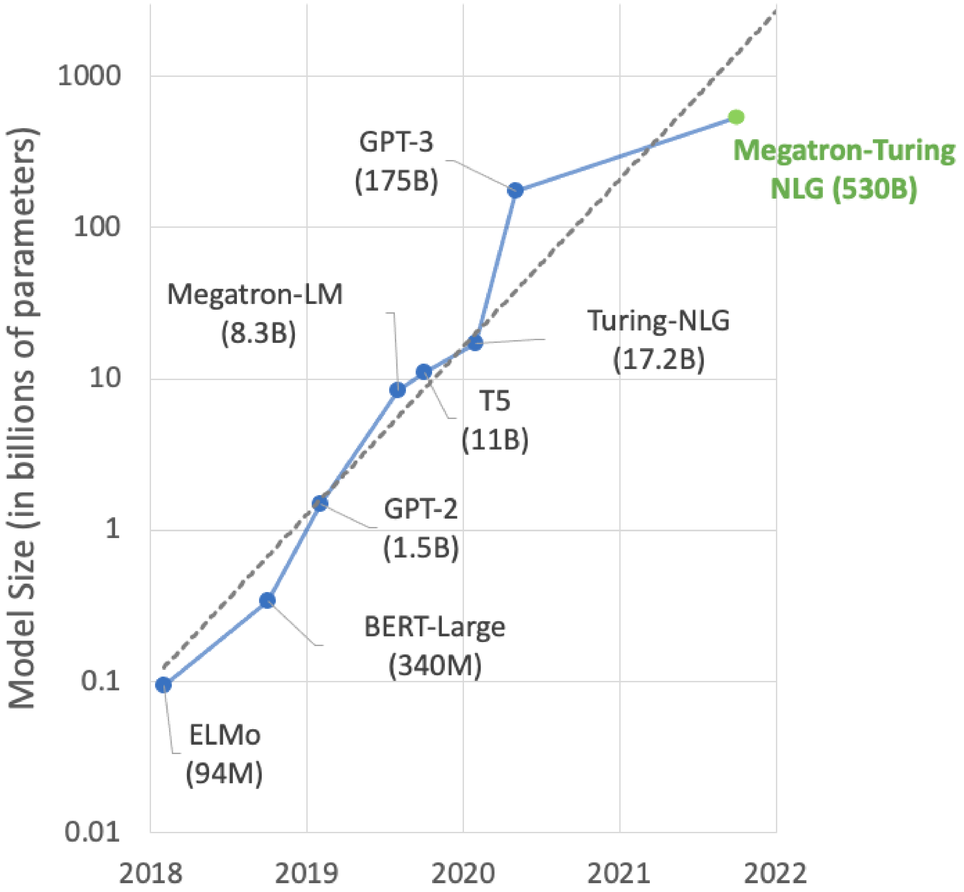
TAO GUI Version – EA Early 2022
- Zero-Code model development
- Train, Adapt and Optimize models with just a few clicks

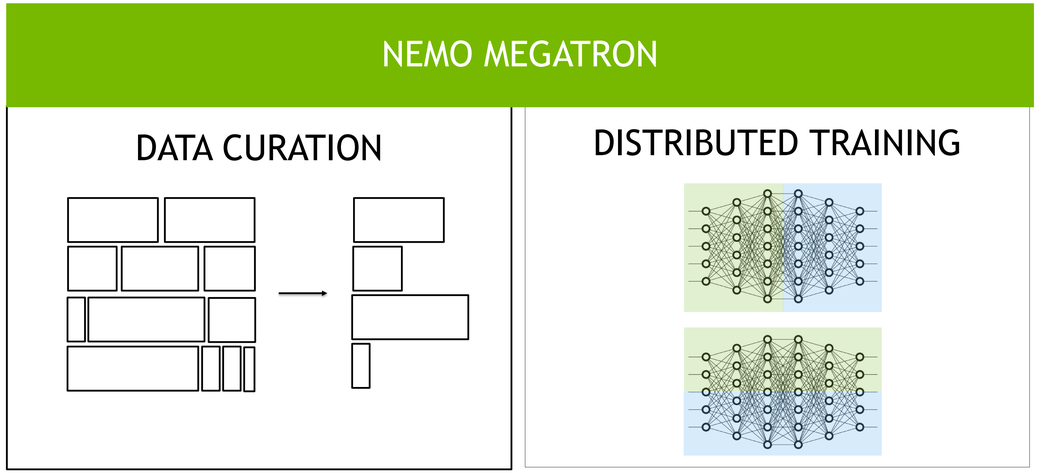


Get started with today with the TAO Toolkit | Sign up for TAO GUI Early Access

NEMO MEGATRON

Framework for Training Large-Scale Language Models





- Language- and industry-specific chatbots, personal assistants, content generation, summarization
- Pipeline, Tensor and Data Parallelism
- Optimized for DGX SuperPOD
- Support Trillions of Parameter Models, Thousands of GPUs

Sign up for EA
developer.nvidia.com/nvidia-nemo

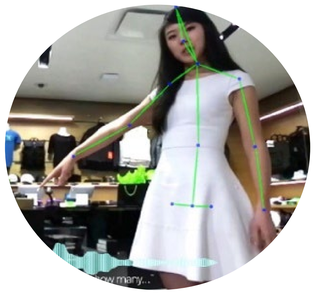
NVIDIA

CONVERSATIONAL AI-RIVA

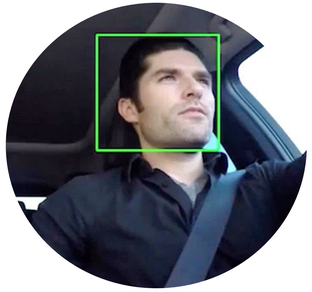
VIDEOCONFERENCE CC,
TRANSLATION, TRANSCRIPTION
200M Meetings per Day

CALL CENTER
500M Calls per Day

SMART SPEAKERS
150M Sold per Year

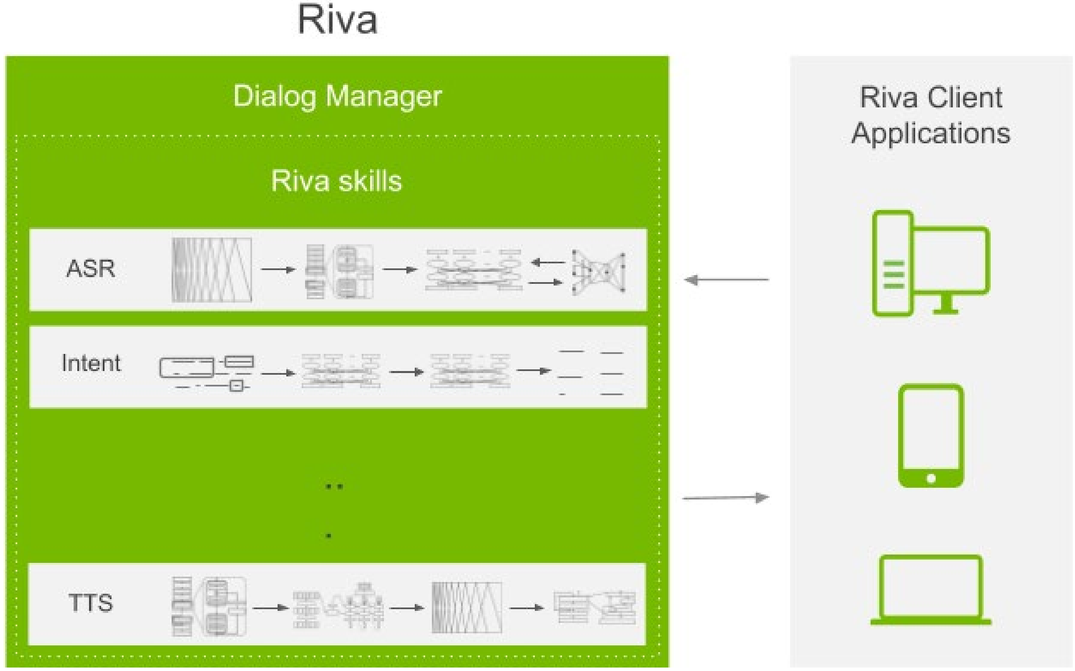


RETAIL ASSISTANTS
12M Retail Stores



IN-CAR ASSISTANTS
75M New Cars per Year

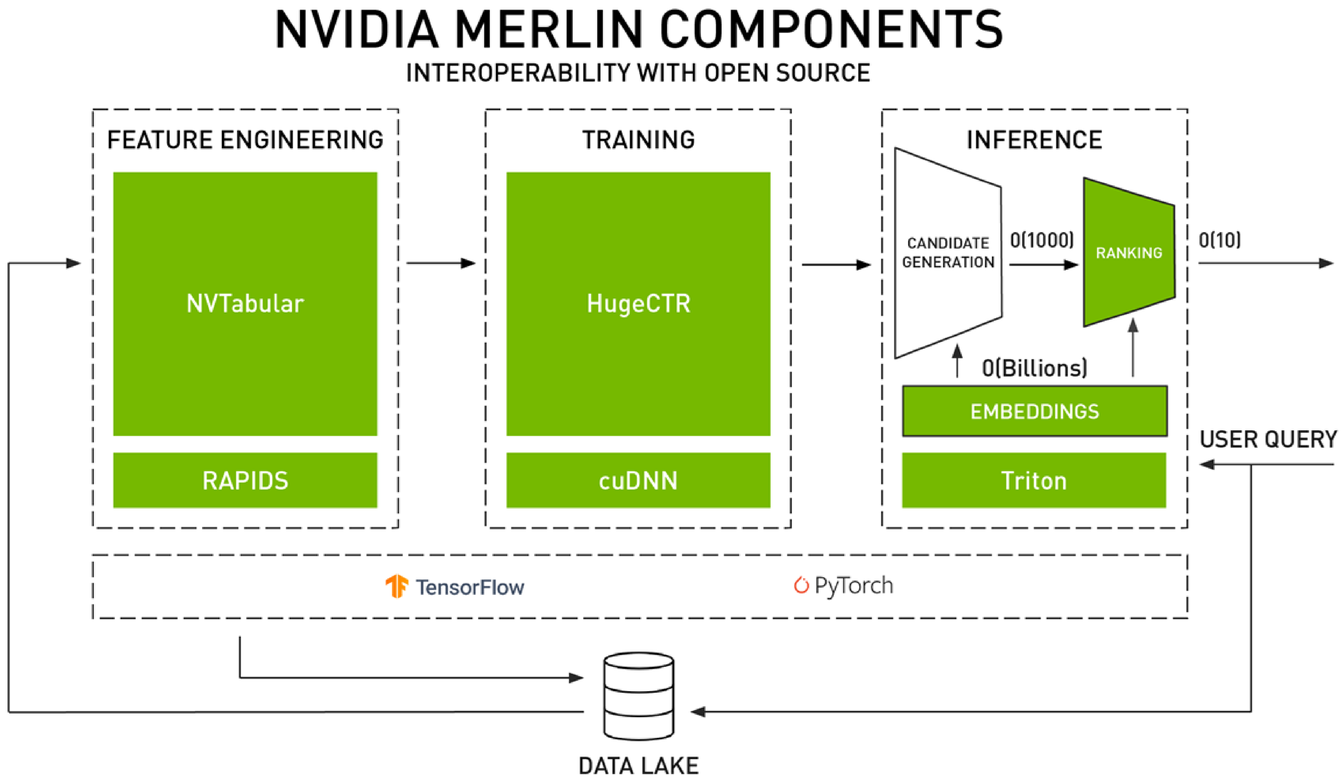
CLIENT APPLICATIONS LEVERAGE Riva SKILLS
TO BUILD NEW USER EXPERIENCES



NVIDIA MERLIN

BUILD, TRAIN, AND DEPLOY HIGH PERFORMING RECOMMENDERS AT SCALE

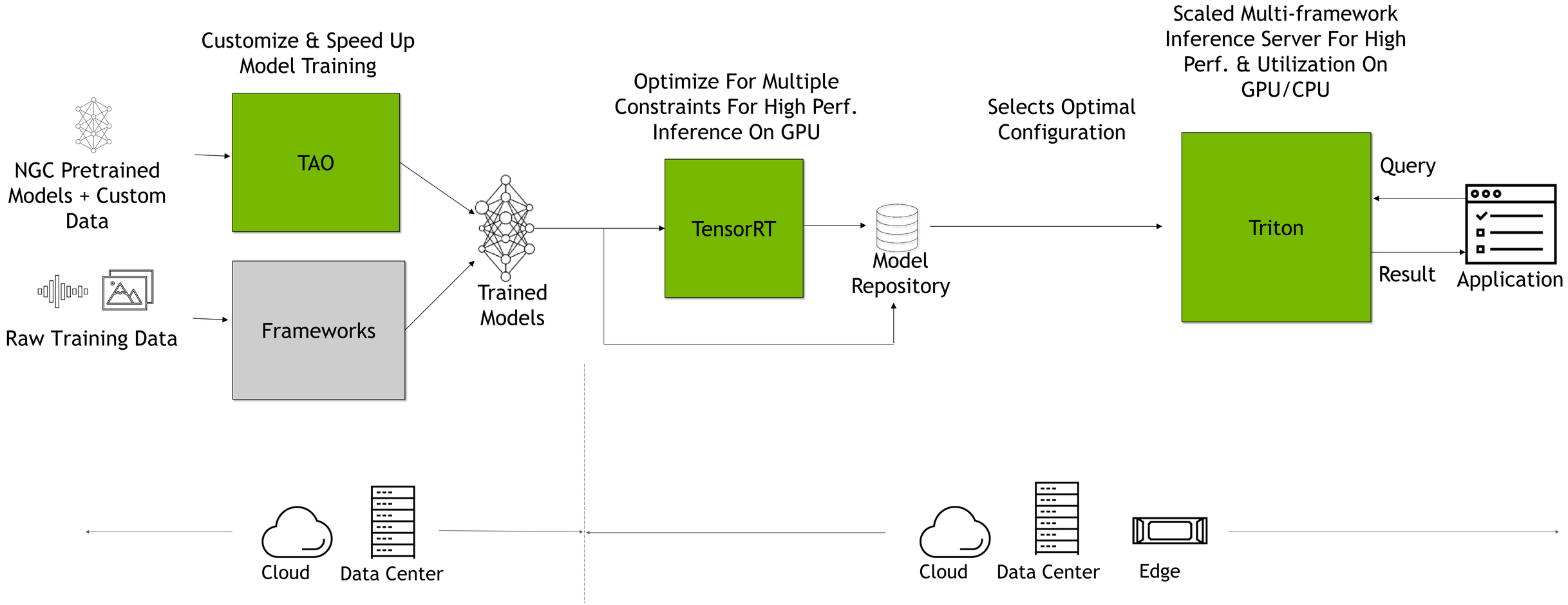
- Designed for Recommender Workflows
- Solves Common Challenges
- Accelerates Entire Pipeline
- Optimized for GPUs



https://developer.nvidia.com/nvidia-merlin

NVIDIA

END-TO-END INFERENCE WITH NVIDIA AI



NVIDIA TensorRT

SDK for High-Performance Deep Learning Inference

Optimize and deploy neural networks in production.

Maximize throughput for latency-critical apps with compiler and runtime.

Optimize every network, including CNNs, RNNs, and Transformers.

1. Reduced mixed precision: FP32, TF32, FP16, and INT8.
2. Layer and tensor fusion: Optimizes use of GPU memory bandwidth.
3. Kernel auto-tuning: Select best algorithm on target GPU.
4. Dynamic tensor memory: Deploy memory-efficient apps.
5. Multi-stream execution: Scalable design to process multiple streams.
6. Time fusion: Optimizes RNN over time steps.

https://developer.nvidia.com/tensorrt

Trained DNN → TensorRT Optimizer → TensorRT Runtime

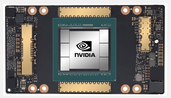
Embedded

Automotive

Data Center

Jetson

Drive



Data Center GPUs

TENSORRT FRAMEWORK INTEGRATIONS

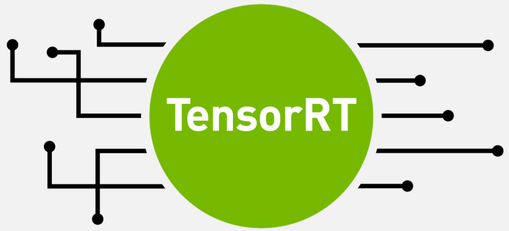
Torch-TensorRT & TensorFlow-TensorRT

Speed Up Native Framework Inference with TensorRT

Accelerate inference with one line of code

- Up to 6x faster inference than framework only on GPUs.
- Optimized to run on every platform from cloud to edge.
- CNNs, RNNs, and Transformers.
- FP32, FP16, INT8.

Available in ready-to-run PyTorch and TensorFlow containers on NGC catalog



PyTorch TensorFlow

TRITON INFERENCE SERVER

Bringing Fast and Scalable AI to Applications

All Major Frameworks, Major Clouds, AI Platforms

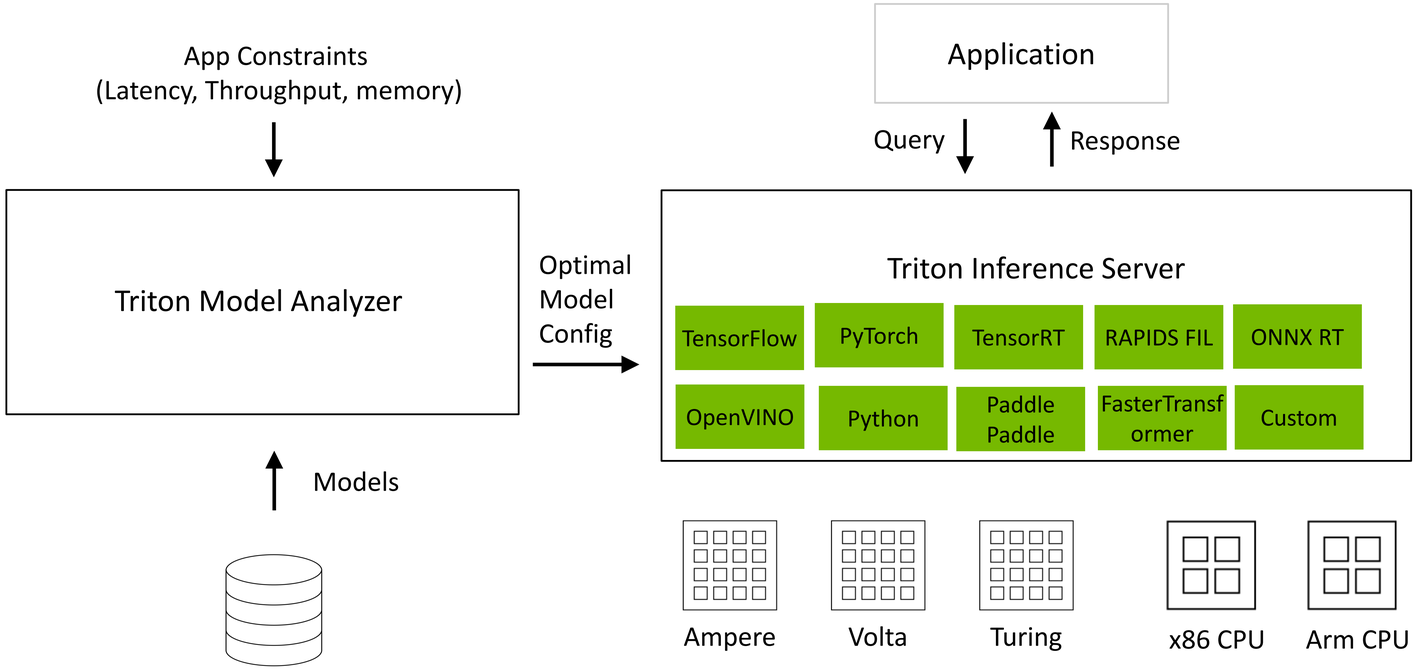
Inference On Every Generation Of GPUs, x86 CPUs And Arm CPUs

Diverse query types – Real time, Offline batch, Video/Audio streaming, Ensembles

Model Analyzer Optimizes For App Constraints

Distributed Multi-GPU Multi-Node Inference

RAPIDS FIL Backend For inference on tree based models (e.g., XGBoost, scikit-learn random forest)

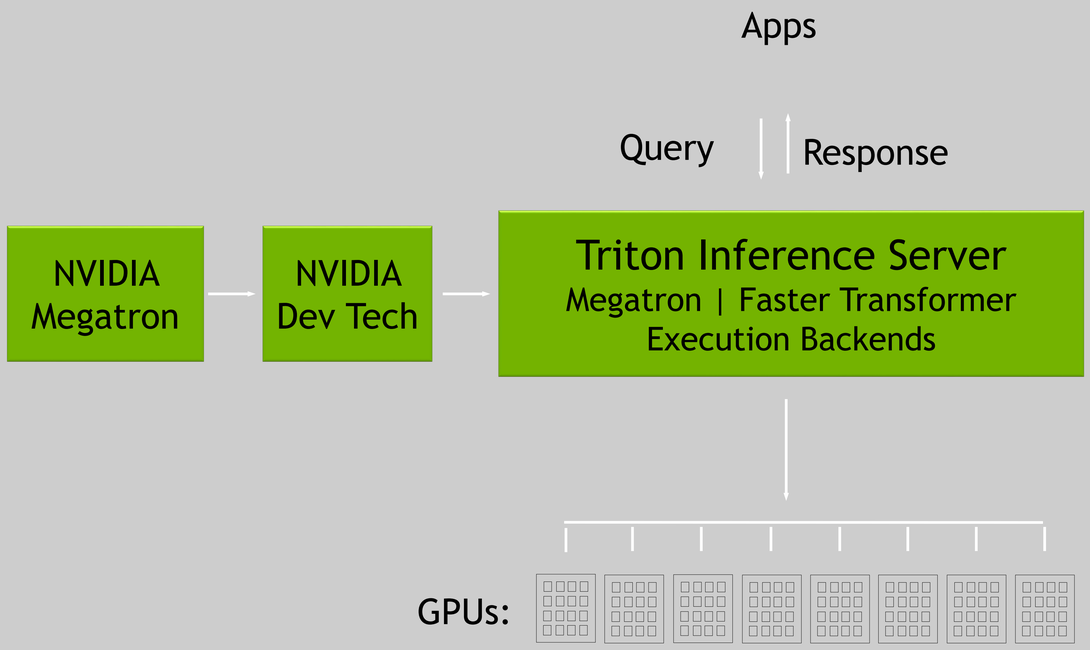


* For full list of backends and functionality, please refer to https://GitHub.com/triton-inference-server/server

REAL-TIME INFERENCE ON GIANT NLP MODELS WITH TRITON

GIANT MODEL INFERENCE

Multi-Node Execution Backends

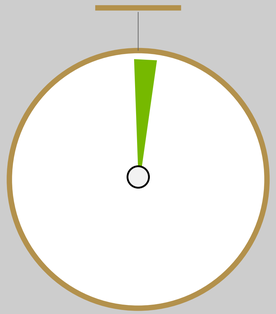


REAL-TIME PERFORMANCE

GPT-3 Based Chatbots

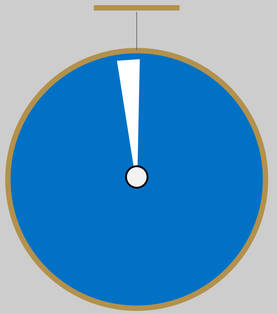
16 QUERIES
1 SECOND

DGX A100



1 QUERY
>1 Minute

Dual Socket CPU Server



Input sequence length=128 tokens (average of 102 words), Output sequence length=8 tokens (average of 6 words)
GPU: Megatron GPT-3 on DGX-A100-80GB, Batch size=16, FP16, FasterTransformer 4.0, Triton 2.6
CPU: OpenAI GPT-3 on Xeon Platinum 8280 2S, 755GB System memory, Batch size=1, FP32, TensorFlow 2.3

NVIDIA

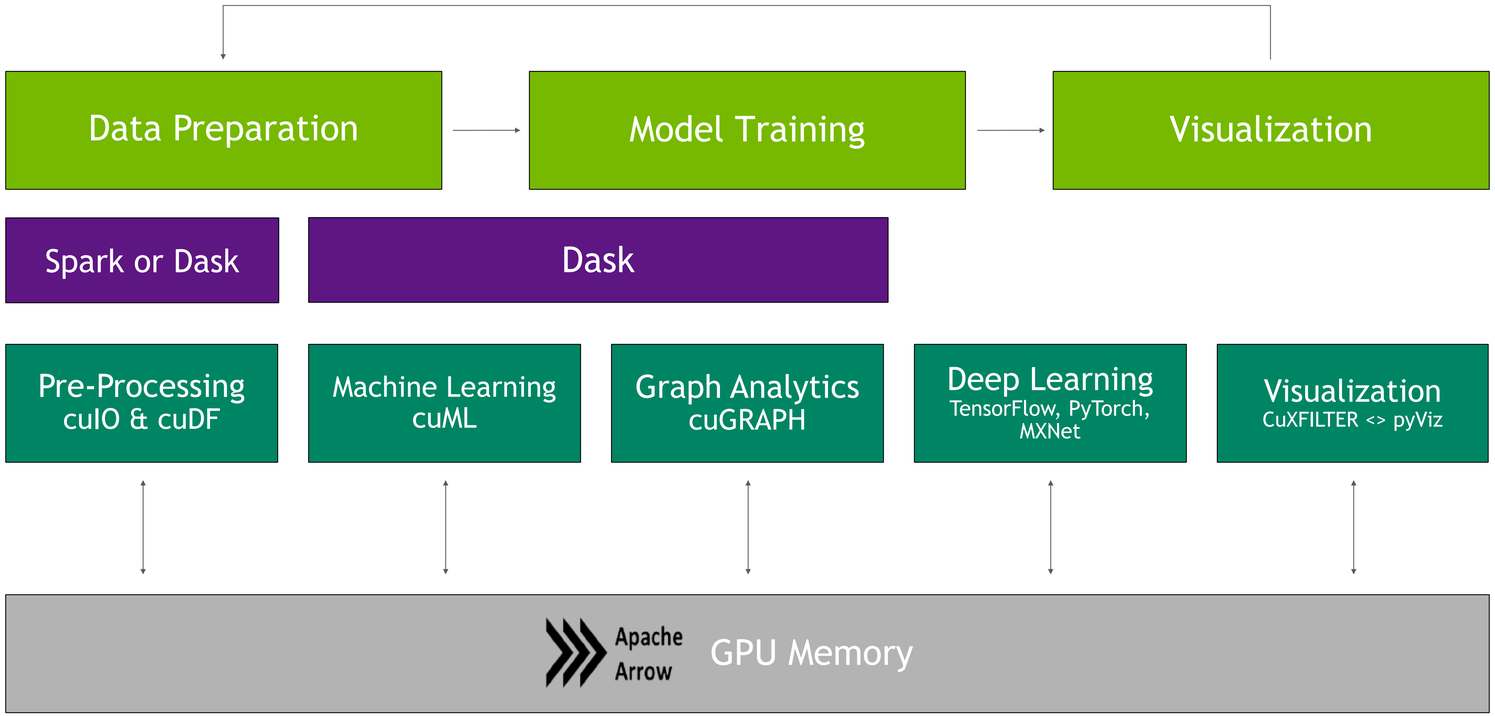
RAPIDS ACCELERATES POPULAR DATA SCIENCE TOOLS

DELIVERING ENTERPRISE-GRADE DATA SCIENCE SOLUTIONS

The RAPIDS suite of open source software libraries gives you the freedom to execute end-to-end data science and analytics pipelines entirely on GPUs.

RAPIDS utilizes **NVIDIA CUDA** primitives for low-level compute optimization and exposes GPU parallelism and high-bandwidth memory speed through user-friendly interfaces like Apache Spark or Dask.

With Spark or Dask, RAPIDS can scale out to multi-node, multi-GPU cluster to power through big data processes.



RAPIDS puts the power of GPUs in the hands of all Data Scientists with drop-in replacement of existing libraries

CUNUMERIC

Automatic NumPy Acceleration and Scalability

cuNumeric

CuNumeric transparently accelerates and scales existing Numpy workloads

Program from the edge to the supercomputer in Python by changing 1 import line

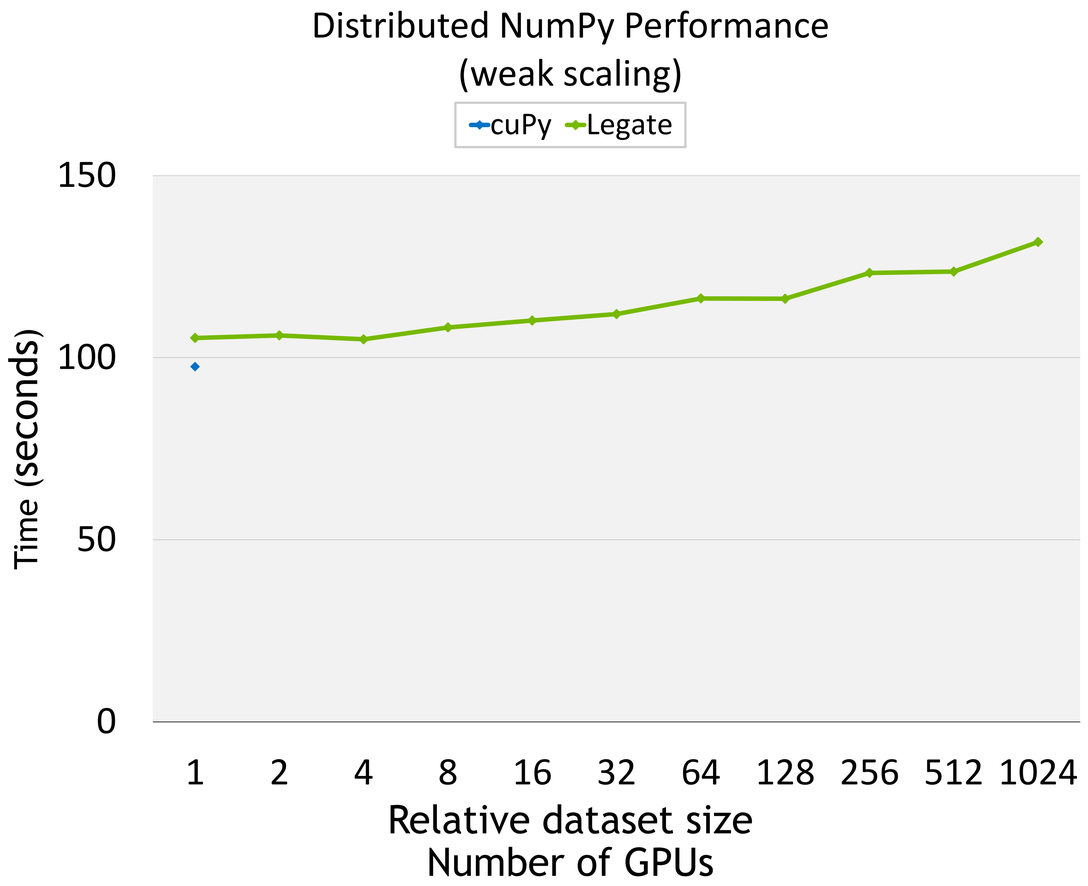
Pass data between Legate libraries without worrying about distribution or synchronization requirements

Alpha release available at github.com/nv-legate

```
for _ in range(iter):
    un = u.copy()

    vn = v.copy()
    b = build_up_b(rho, dt, dx, dy, u, v)
    p = pressure_poisson_periodic(b, nit, p, dx, dy)
```

…

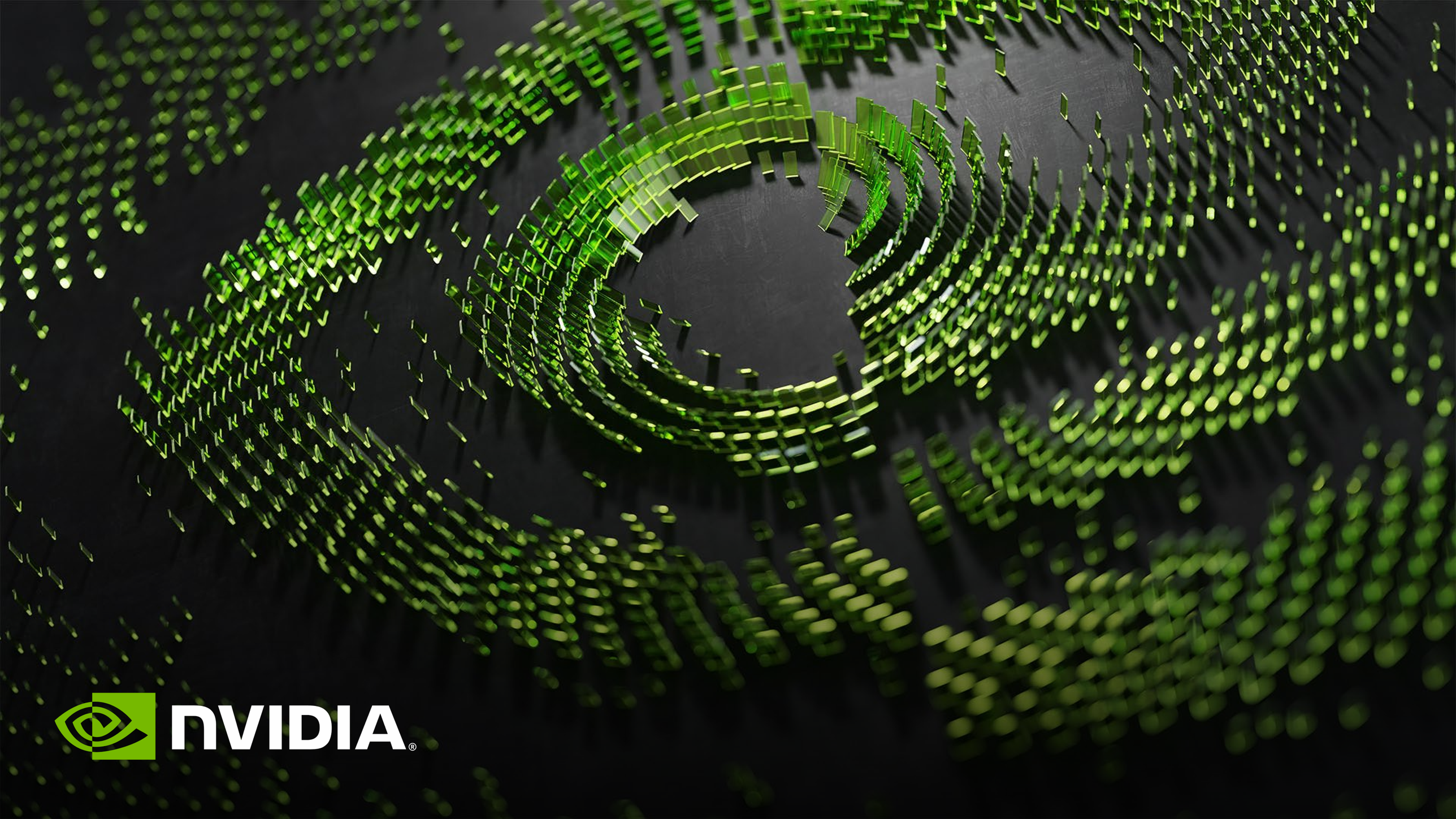


Extracted from “CFD Python” course at https://github.com/barbagroup/CFDPython
Barba, Lorena A., and Forsyth, Gilbert F. (2018). CFD Python: the 12 steps to Navier-Stokes equations. *Journal of Open Source Education*, **1**(9), 21, https://doi.org/10.21105/jose.00021

NVIDIA

SUMMARY AND KEY TAKEAWAYS

- NVIDIA is a full-stack computing company, covering GPUs, DPUs, SOFTWARE and ORCHESTRATIONS
- NVIDIA's layered offering with GPU+DPUs, orchestration platform software such as NVIDIA AI Enterprise, the CUDA-X middleware layer, deep learning and AI SDKs, as well as the industry vertical specific and use-case specific SDK and APIs allows NVIDIA AI that not only runs everywhere, but also scales beautifully.
- NVIDIA's platform offering continues to be the best-in-class in performance for Mlperf in both AI training and inference
- NVIDIA releases a row of new Software SDKs to continue the improvement Data Science, AI, and Inference Performance:
  - RAPIDS as drop-in replacement for DS workload by replacing popular libraries like PANDAS, Scikit
  - cuNumeric provides replacement and scalability for the popular library of NumPy
  - NVIDIA RIVA, NemoMegatron and TAO toolkit enables new possibly of building the domain-specific NLP and Speech offerings
  - TensorRT and Trition allows easy implementation and deployment for AI inference workflow across different platforms


NVIDIA