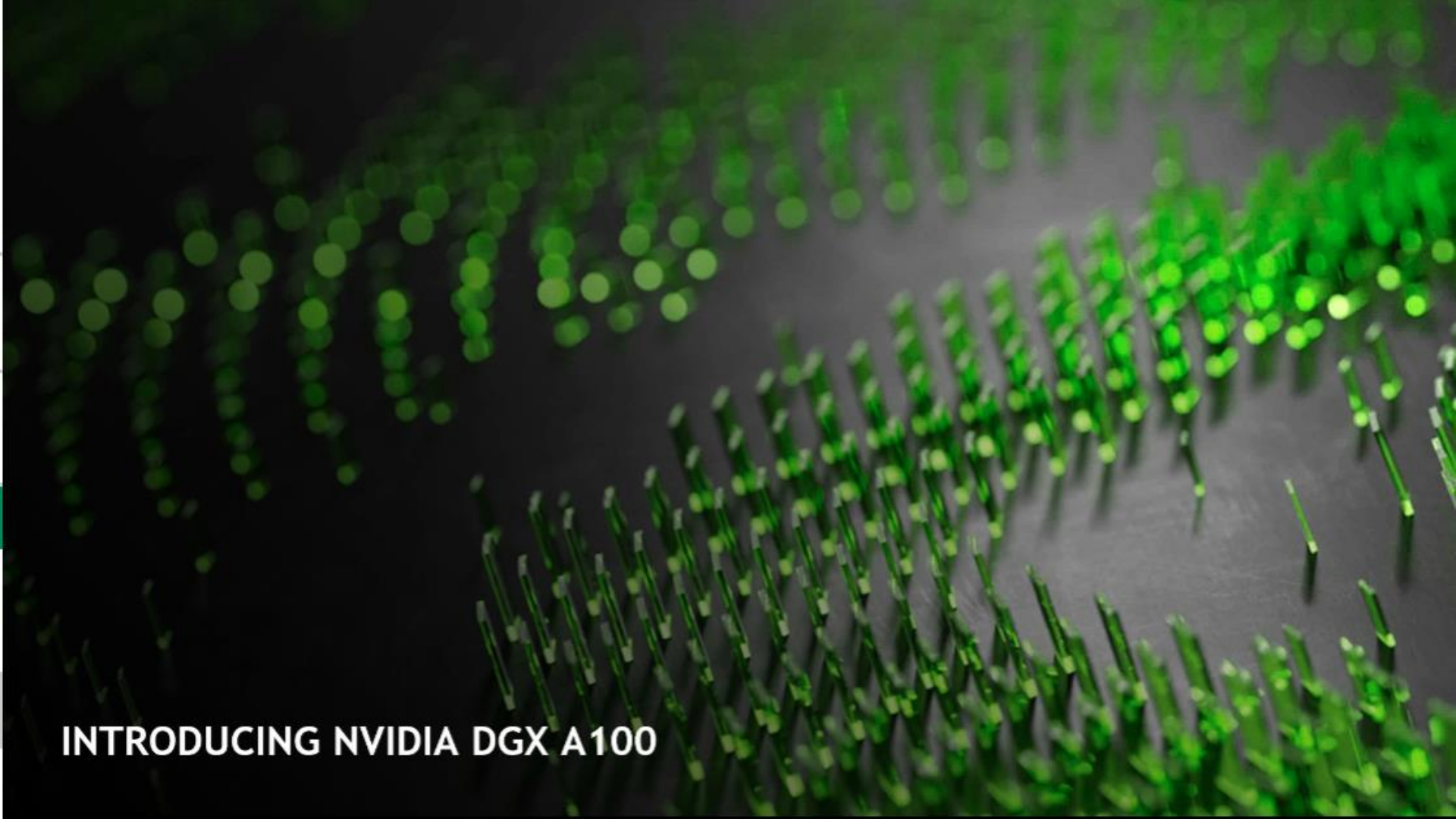

INTRODUCING NVIDIA DGX A100

TOPICS

Introducing NVIDIA DGX A100

Game-changing Performance

DGX A100 Customers

THE CHALLENGES OF AI TRANSFORMATION

Enterprises Need Infrastructure That Supports the Lifecycle of AI Innovation

From Inspiration

AI practitioners need the right tools for exploration:

- Iterating to the best model, with less effort expended
- Fastest time-to-solution for every training run
- Insulation from the bleeding edge of AI open source

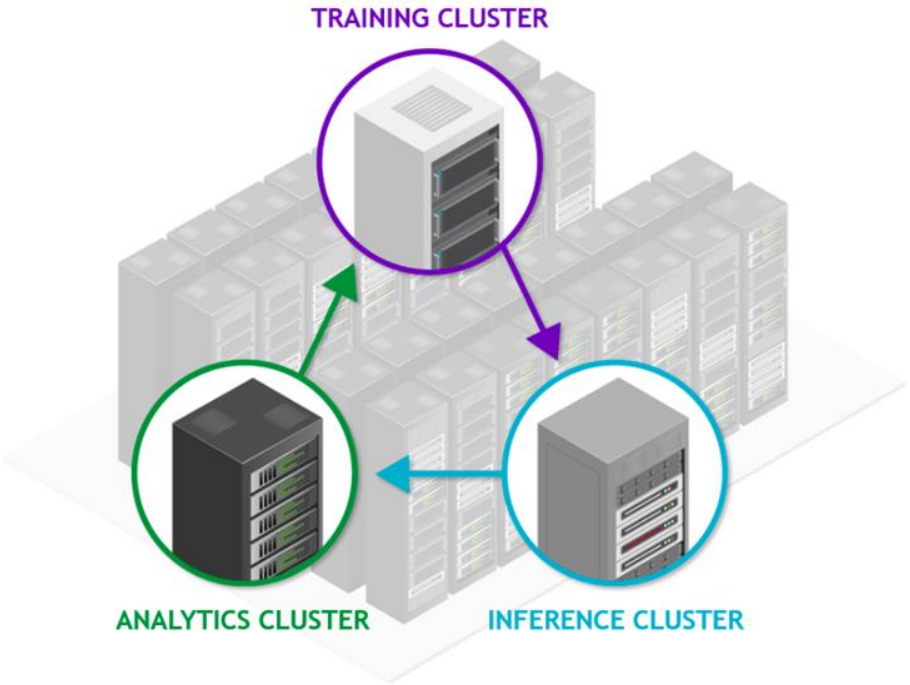
To Production

IT needs a standardized approach for AI infrastructure:

- Simplified infrastructure planning, heterogenous workloads & users
- Security at every layer, operations peace-of-mind
- Linearly predictable performance with scale

SOLVING THE INFLEXIBILITY OF AI INFRASTRUCTURE

Not Optimized, Complex to Manage, Difficult to Scale Predictably

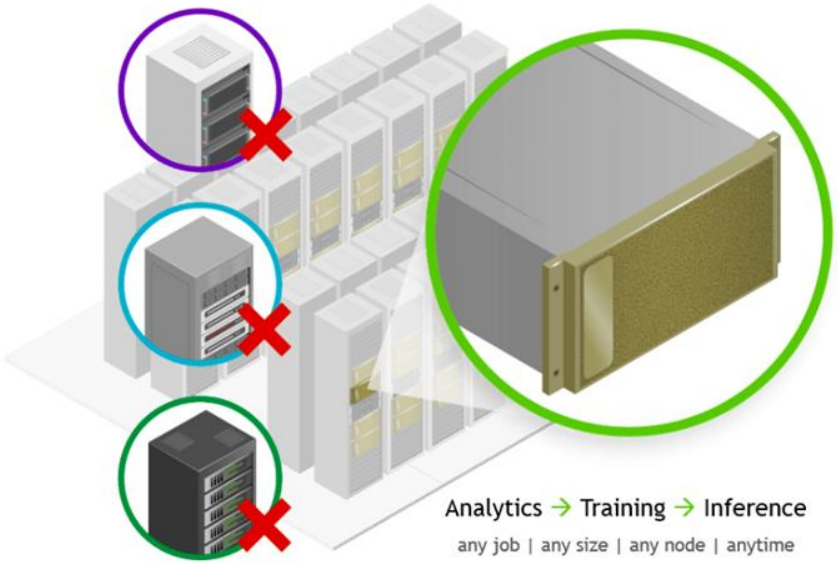


Inflexible infrastructure that was never meant for the pace of AI

- Constrained workload placement by system-level characteristics
- Non-uniform performance across the data center
- Unable to adapt to dynamic workload demands
- Constrained capacity planning

ONE SYSTEM FOR ALL AI INFRASTRUCTURE

AI Infrastructure Re-Imagined, Optimized, and Ready for Enterprise AI-at-Scale



Flexible AI infrastructure that adapts to the pace of enterprise

- One universal building block for the AI data center
- Uniform, consistent performance across the data center
- Any workload on any node - any time
- Limitless capacity planning with predictably great performance with scale

DGX A100: THE UNIVERSAL AI SYSTEM

One System for Every AI Workload

Performance meets utility - analytics, AI training and inference all in one

Integrated Access to Unmatched AI Expertise

Fast-track AI transformation with DGXpert know-how and experience

Game-changing Performance for Innovators

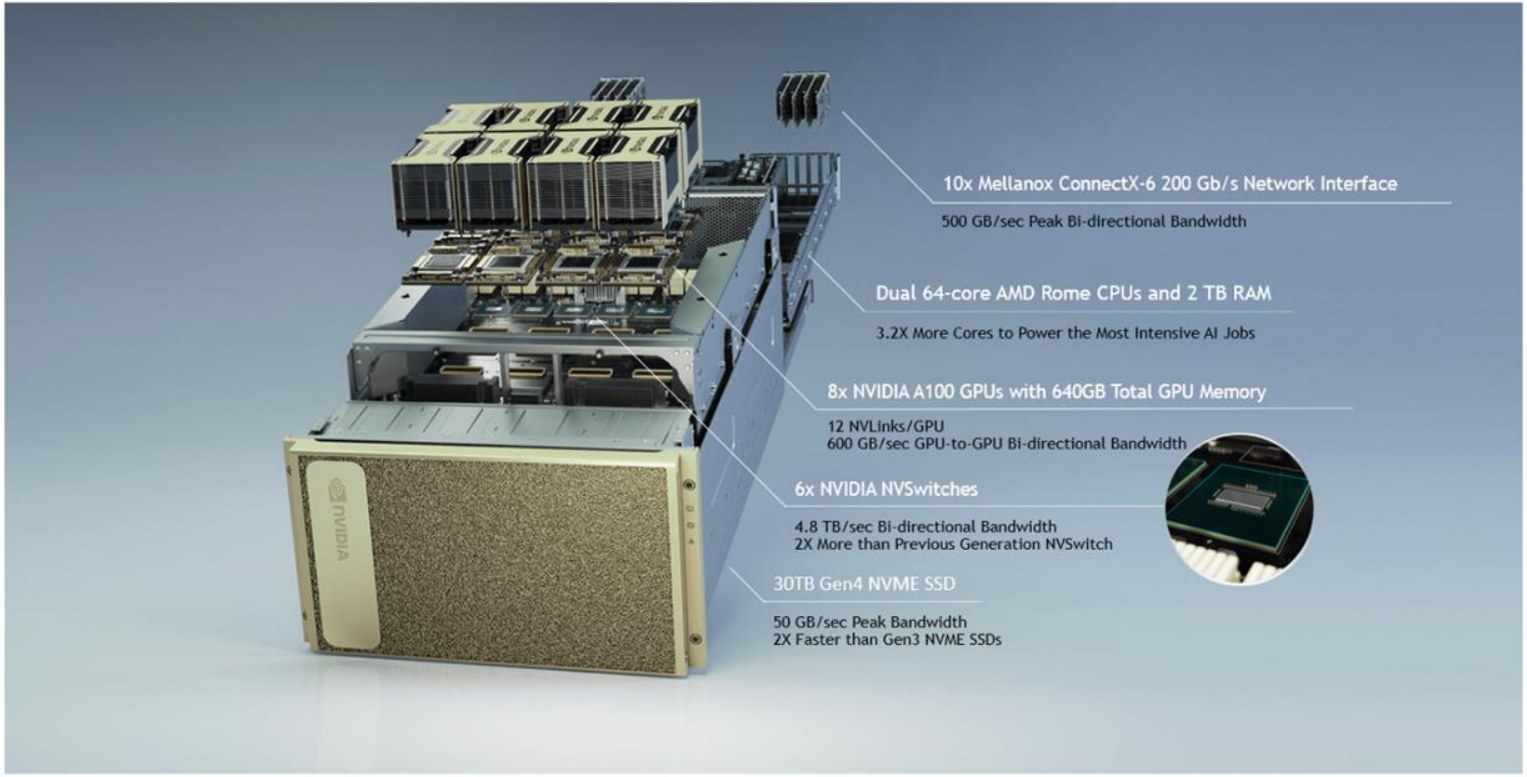
Fastest time-to-solution with the world's first 5 petaFLOPS AI system, built on NVIDIA A100

Unmatched Data Center Scalability

Build leadership-class infrastructure that scales to keep ahead of demand

GAME-CHANGING PERFORMANCE FOR INNOVATORS

NVIDIA DGX A100 640GB System



DGX A100: A100 GPUS AND 2X FASTER NVSWITCH

5 PetaFLOPS AI Performance

Eight A100 Tensor Core GPUs, Up to 640GB total GPU Memory

- Twelve NVLinks per GPU, 2x more than V100
- 600GB/s bi-directional bandwidth between any GPU pair
- Nearly 10X PCIe Gen4 bandwidth with third-gen NVLink

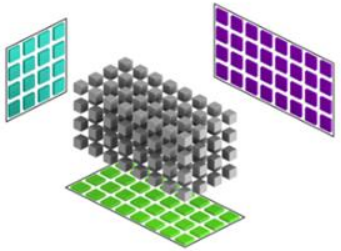
All GPUs fully connected with six second-gen NVSwitch

- 4.8TB/s bi-directional bandwidth
- In one second we could transfer 426 hours of HD video

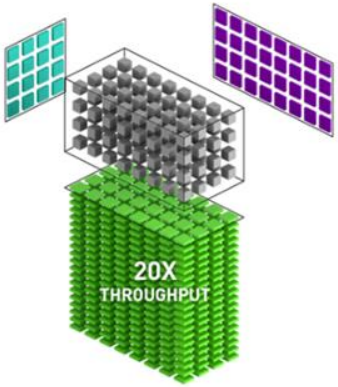
NEW TF32 TENSOR CORES ON A100

20X Higher FLOPS for AI, Zero Code Change

NVIDIA V100 FP32



NVIDIA A100 Tensor Core TF32 with Sparsity



20X Faster than Volta FP32 | Range of FP32 and Precision of FP16
No Code Change Required for End Users | Supported on PyTorch, TensorFlow and MXNet Frameworks Containers

MOST FLEXIBLE AI PLATFORM WITH MULTI-INSTANCE GPU (MIG)

Optimize GPU Utilization, Expand Access to More Users with Guaranteed Quality of Service

- Up To 7 GPU Instances In a Single A100
- Simultaneous Workload Execution With Guaranteed Quality Of Service
  All MIG instances run in parallel with predictable throughput & latency
- Flexibility to run any type of workload on a MIG instance
- Right Sized GPU Allocation
  Different sized MIG instances based on target workloads

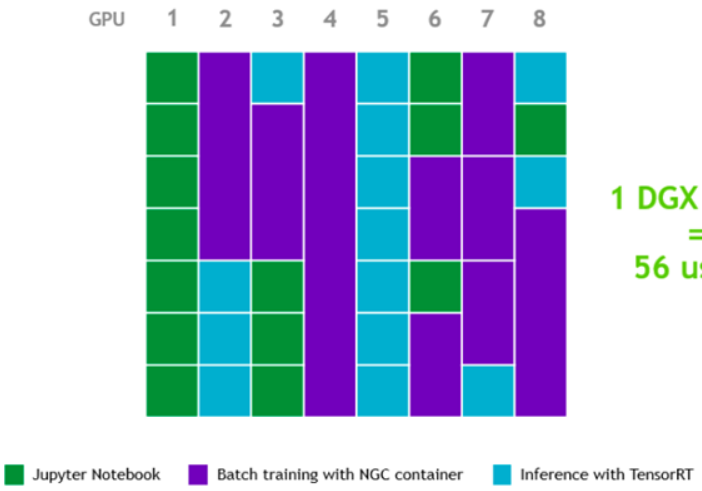
MULTI-INSTANCE GPU (MIG) ON DGX A100 640GB

2X More GPU Memory per MIG Instance with A100 80GB GPUs

GPU Instance Size	Number of GPU Instances Available	GPU Memory
1 GPU Slice	7	10 GB
2 GPU Slice	3	20 GB
3 GPU Slice	2	40 GB
4 GPU Slice	1	40 GB
7 GPU Slice	1	80 GB

Flexible Utilization

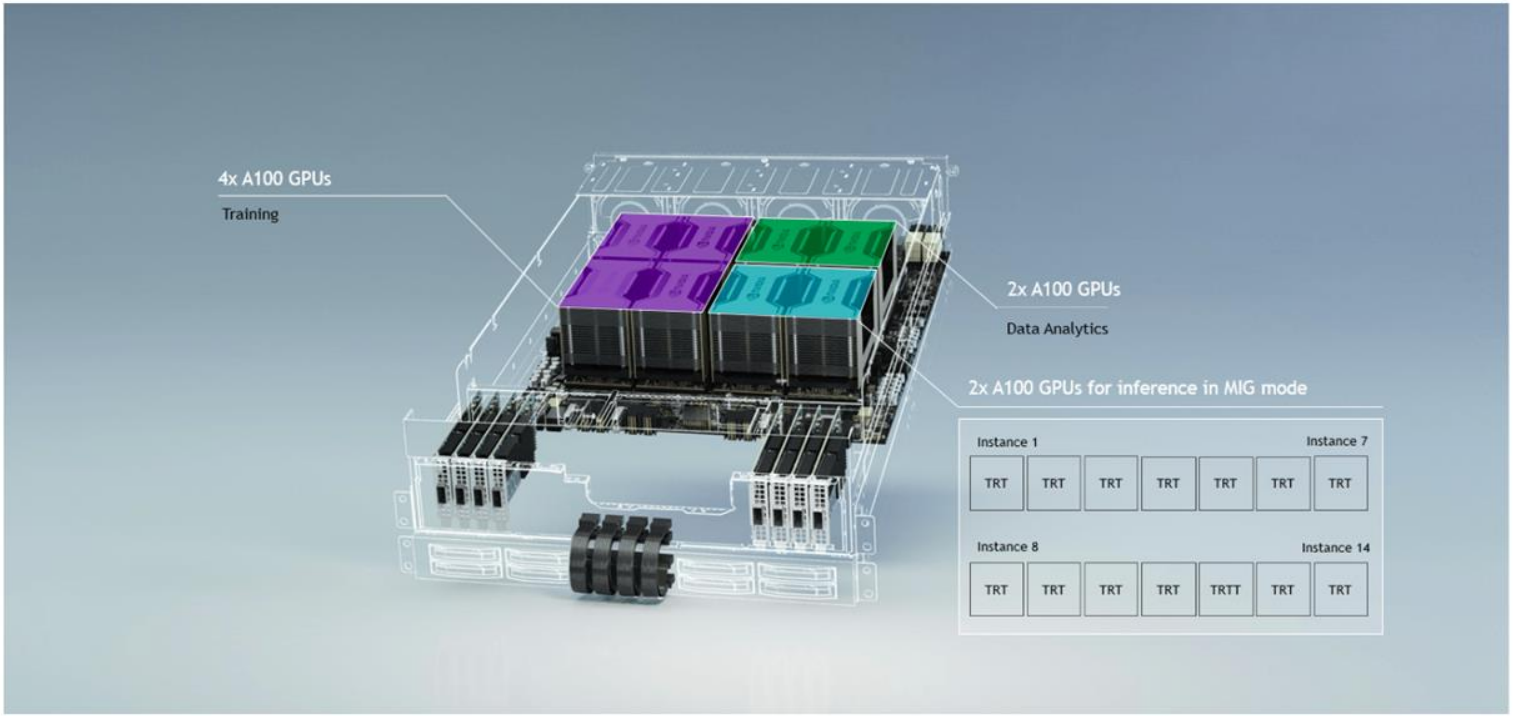
Configure GPUs for vastly different workloads with GPU instances that are fault-isolated



1 DGX A100 = 56 users

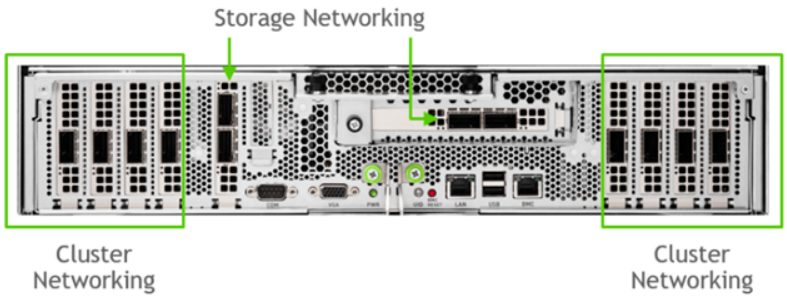
CONSOLIDATING DIFFERENT WORKLOADS ON DGX A100

One Platform for Training, Inference and Data Analytics



UNMATCHED SCALABILITY WITH MELLANOX NETWORKING

Highest Network Throughput for Data and Clustering



For clustering networking:

- Eight Mellanox single-port ConnectX-6
- Supporting HDR/HDR100/EDR InfiniBand default or 200GigE

For data/storage networking:

- Up to Two Mellanox dual-port ConnectX-6
- Supporting: 200/100/50/40/25/10Gb Ethernet default or HDR/HDR100/EDR InfiniBand

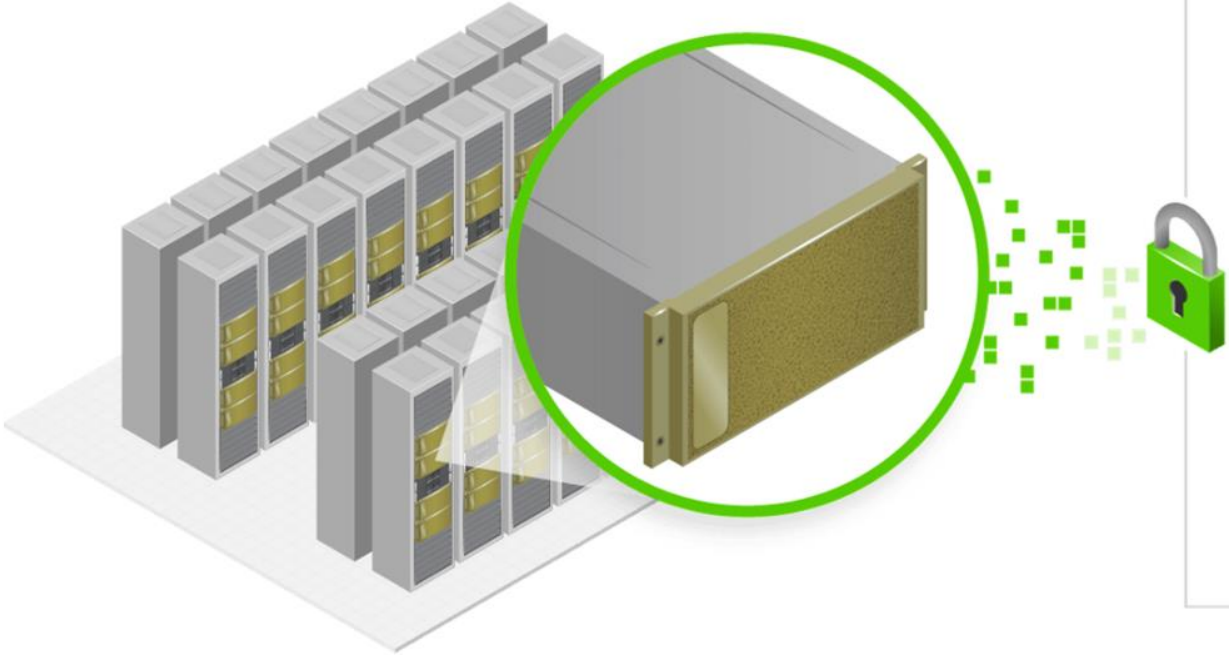
Up to 4 Tb/sec peak bi-directional bandwidth

All I/O now PCIe Gen4, 2X performance increase over Gen3

Scale up multiple DGX A100 nodes with Mellanox Quantum Switch, the world's smartest network switch

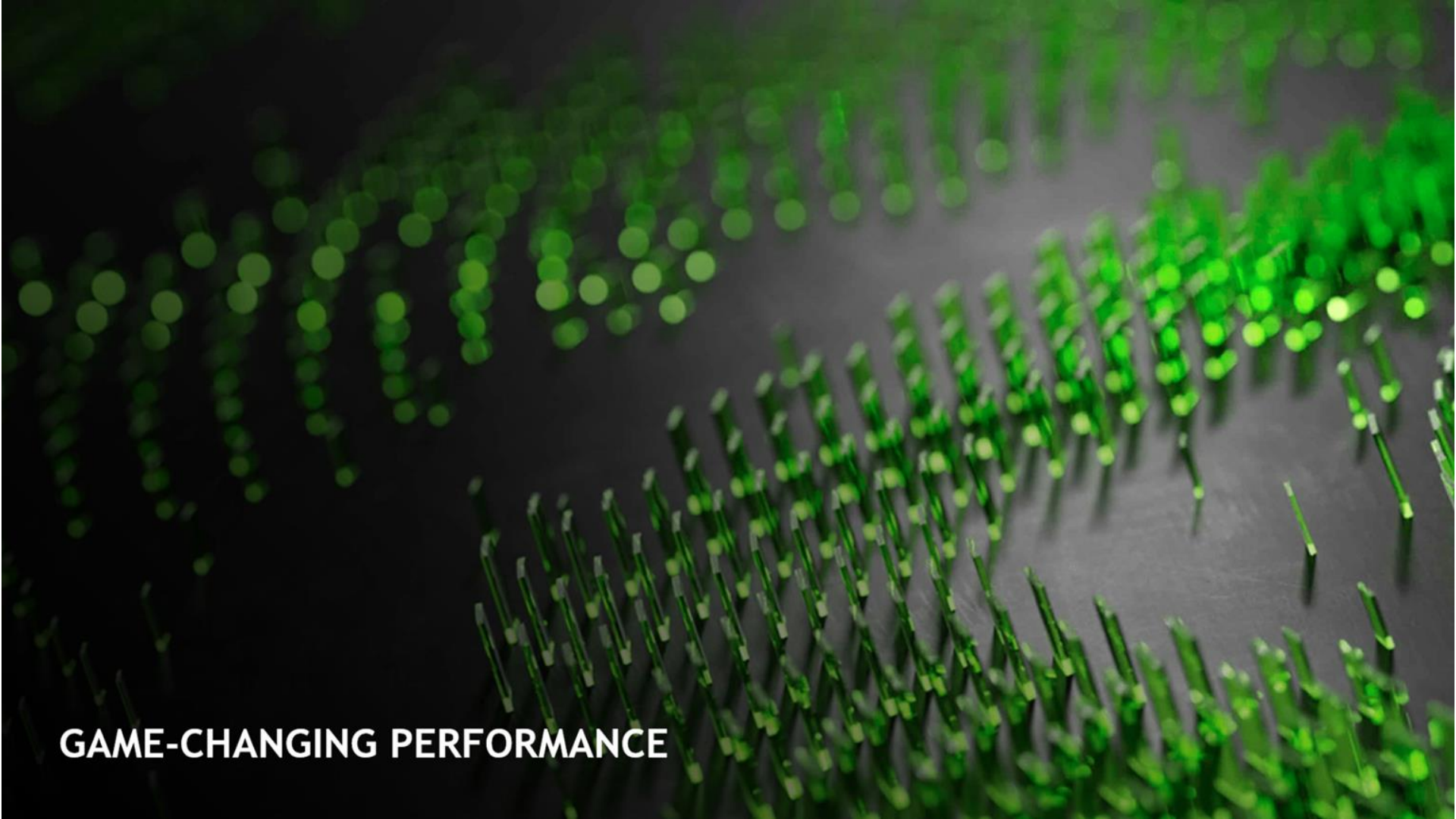
THE WORLD'S MOST SECURE AI SYSTEM FOR ENTERPRISE

Built-In Security: Multi-layered Defense for AI Infrastructure



- Resource sharing with MIG
- Self-Encrypted Drives (SED) to protect data at rest
- GPU Board
- CPU Board
- BMC

Secure Firmware Update


GAME-CHANGING PERFORMANCE

DGX A100 PERFORMANCE

Up to 83X Higher Throughput than CPU

DLRM Training

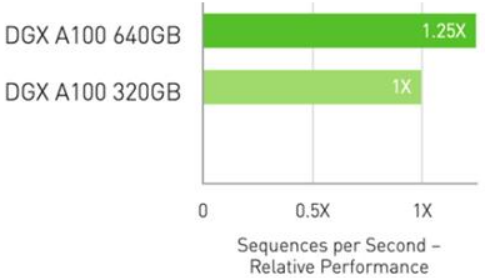
Up to 3X Higher Throughput
for AI Training on Largest Models

Large Model Training

DLRM (Huge CTR framework), FP16 precision | 1x DGX A100 640GB batch size = 48 | 2x DGX A100 320GB batch size = 32 | 1x DGX-2 (16x V100 32GB) batch size = 32. Speedups normalized to number of GPUs

RNN-T Inference

Up to 1.25X Higher Throughput
for AI Inference

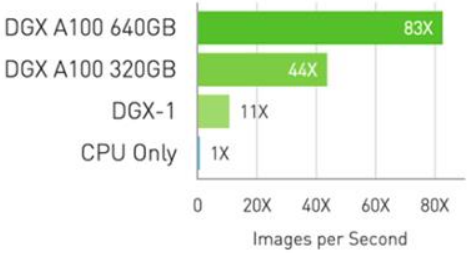


Inference on MIG

MLPerf 0.7 Single stream latency, RNN-T measured with [1/7] MIG slices. Framework: TensorRT 7.2, dataset = LibriSpeech, FP16 precision

Big Data Analytics

Up to 83X Higher Throughput
than CPU



Analyzing Massive Datasets

Big data analytics benchmark | 30 analytical retail queries, ETL, ML, NLP on 10TB dataset | CPU: 19x Intel Xeon Gold 6252 2.10GHz, Hadoop | 16x DGX-1 (8x V100 32GB), RAPIDS/Dask | 12x DGX A100 320GB and 6x DGX A100 640GB, RAPIDS/Dask/BlazingSQL. Speedups normalized to number of GPUs

MOST POWERFUL TOOL FOR A DATA SCIENCE TEAM

Using DGX A100 with MIG to Give Every Developer Power to Explore

One DGX A100 delivers:

- 5 petaFLOPS of AI training power, or
- 10 petaOPS of AI inference power
- With MIG, a team of 25 developers can share a DGX A100

Each developer gets:

- Over 180 teraFLOPS for training
  = (2) reserved cloud V100 instances

or

- Over 357 teraOPS for inference
  = (6) dedicated 28-core dual CPU servers

TODAY'S AI DATA CENTER

- 50 DGX-1 Systems for AI Training
- 600 CPU Systems for AI Inference
- $11M
- 25 Racks
- 630 kW

DGX A100 DATA CENTER

- 5 DGX A100 systems for AI training and inference
- $1M
- 1 rack
- 32.5 KW


NVIDIA
NVIDIA SELENE
Now Featuring NVIDIA DGX A100 640G
#5 Top500 | #1 MLPerf | #1 Industrial System
#1 in Green 500
4,480 A100 GPUs
560 DGX A100 system
850 Mellanox 200G HDR switches
14 PB of high-performance storage
2.8 EFLOPS of AI peak performance
63 PFLOPS HPL @ 24GF/W

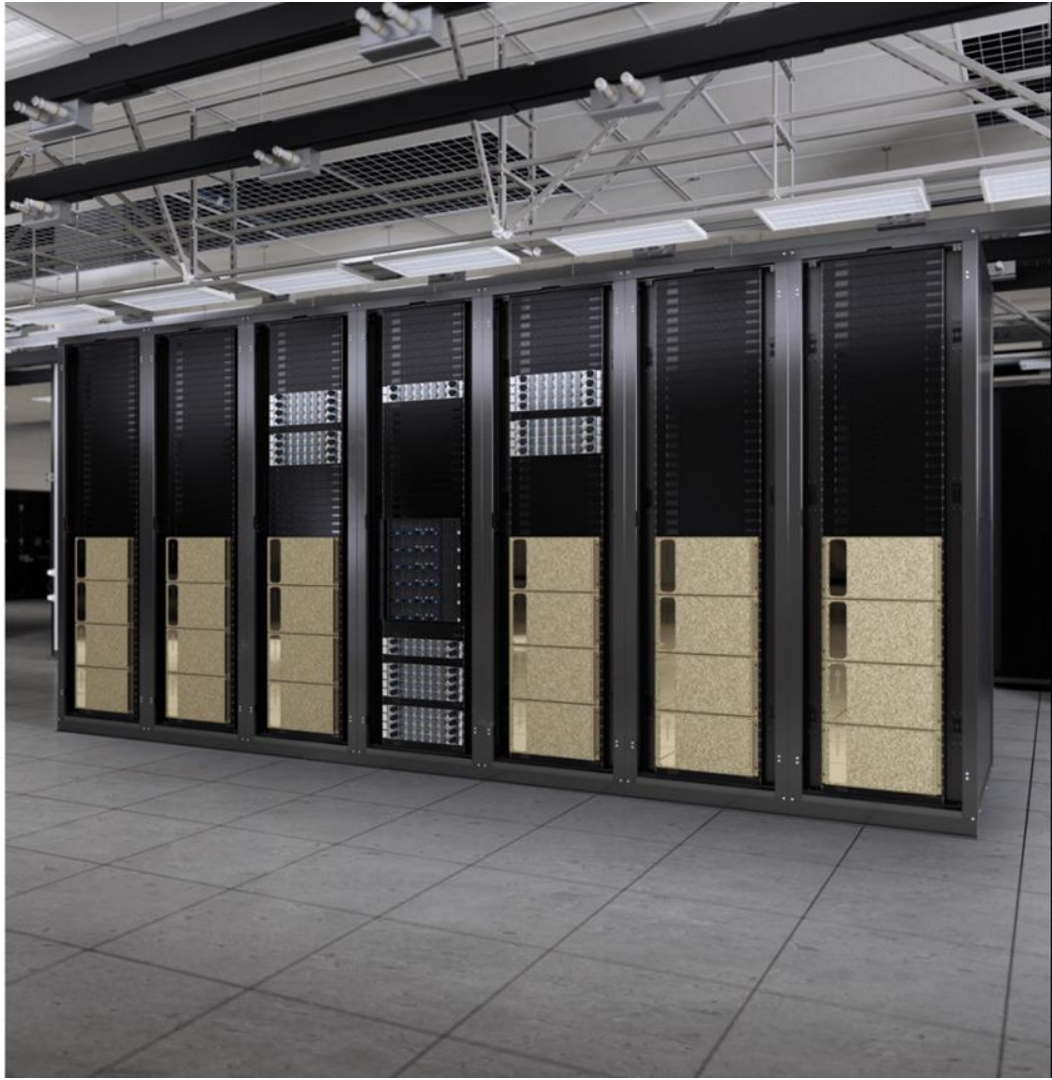
DGX A100 CUSTOMERS

ARGONNE NATIONAL LABORATORY

World's First DGX A100 Supercomputer Fighting COVID-19

- 24-node Cluster of DGX A100 Systems
- 192 A100 GPUs
- Mellanox High-Speed Low-Latency Network Fabric
- 120 PetaFLOPS of AI Computing Power for Scientific Research

Argonne NATIONAL LABORATORY



> *“ Our AI medical imaging workflow powered by NVIDIA DGX A100 and NVIDIA Clara provides pre-trained models, real-time AI-Assisted Annotation for data labeling, and hyperparameter finetuning which helps us build our AI radiology models quickly to ultimately save physician's time on reading studies and bring information to patients faster. “*
>
> — Lin Xinrong, President, Hualien Tzu Chi Hospital

佛教慈濟醫療財團法人
Buddhist Tzu Chi Medical Foundation

TURNKEY AI-ASSISTED SOLUTION SPEEDS DIAGNOSIS

Challenge

Hualien Tzu Chi Hospital serves over 550,000 patients and is the only medical center in Eastern Taiwan.

Hospital needed to build workflows for quickly reading radiology studies and sharing results with patients.

CT studies for pneumonia and liver lesions required a lot of accurate annotations for model training, which can be tedious and time-consuming.

Solution

Global Enterprise and Mobile PACS company EBM Technologies and NVIDIA developed a solution that provides radiology images on mobile devices that can be shared with patients in exam rooms and has built in AI-Assisted annotation tools for easy data labeling of DICOM images for AI model creation.

DGX A100 helped to speed training time, enabling faster AI model creation.

Clara's pre-trained Liver Segmentation Model was downloaded and fine-tuned by hospital's own DICOM images for quick model development.

NVIDIA DGX A100
for Training

NVIDIA Clara
for AI Assisted Annotation

NVIDIA NGC
for pre-trained Liver model

1 hour vs 3 months

to complete the AI workflow- from image labeling to model training.

“ We use DGX A100 for language-related model development process and have seen a remarkable improvement in training speed and efficiency. ”

— Youngjoon Kim, Vice President, SK Telecom

SKT BUILDS AN AI ECOSYSTEM, INFUSING AI IN EVERYDAY LIFE

Challenge

SK Telecom’s AI voice assistant 'NUGU’ has over 700M monthly active users.

Needed more interactive experiences across devices and services where NUGU is enabled, including in car navigation, mobile phones, home appliances and set-top boxes.

Solution

Trained 110M parameter models on NVIDIA DGX with linear performance and increased accuracy.

The scalability of DGX systems enabled SKT to expand the number of services leveraging 'NUGU' such as interaction with human avatars, applying a K-pop singer's voice to smart speakers, and gesture recognition games.

NVIDIA DGX Systems, including DGX A100

NVIDIA DGX Station

Reduction in training time from **weeks to hours**

> *“ DGX SuperPOD is helping NAVER CLOVA build state-of-the-art language models for Korean and Japanese markets and evolve into a strong AI platform player in the global market. ”*

— Suk Geun SG Chung, Head of NAVER CLOVA CIC, NAVER Corp

CLOVA
NAVER LINE

AI ADVANCES THAT POWER EVERYDAY LIFE

Challenge

NAVER, Korea’s leading web search engine, and LINE, the world’s fastest growing messenger, created an AI technology brand and AI platform, CLOVA.

With CLOVA embedded in cars, home appliances, smart speakers, mobile apps, and B2B/B2C AI solutions, NAVER wanted to ensure scalable AI services to their users to achieve their vision, “AI for Everyone.”

Solution

NAVER is using DGX SuperPOD to create new conversational AI services and to enhance its chatbot and call center solution with improved accuracy.

Their researchers use the infrastructure to develop disruptive technologies for CLOVA.

Recent AI innovations include a smart lamp that converts text from a book into speech, AI speakers that recreate the voice of a celebrity, and apps that translate menus and signs when you travel.

- NVIDIA DGX SuperPOD
- 140 DGX A100 Systems
- DDN A^{3}I Storage
- NVIDIA TensorRT SDK

2.7X faster processing speed over legacy systems

> *"The 210 Petaflops PARAM SIDDHI-AI build using DGX SuperPOD, C-DAC HPC-AI Software Stack and Cloud Platform will accelerate experiments for solving India-specific grand challenges using science and engineering."*

— Dr Hemant Darbari, Director General, C-DAC

HPC-AI INFRASTRUCTURE FOR THE NATION

Challenge

- C-DAC is a premier R&D organization under the Ministry of Electronics and Information Technology in India
- Desired to unite the scientific community and provide AI computing resources to academicians, researchers, enterprises, MSME, industry and start-ups to solve India's specific grand challenges

Solution

Turned to NVIDIA DGX SuperPOD architecture along with CDAC Software Stack and Cloud Platform to develop India's fastest HPC-AI supercomputer, the PARAM Siddhi-AI.

The infrastructure will support thousands of users through cloud, and speed deployment of many projects.

The HPC-AI infrastructure will be used in area such as the discovery new drugs capable of fighting COVID-19, weather prediction, education, agriculture, space, cybersecurity, NLP, and smart cities.

- NVIDIA DGX SuperPOD
- 42 DGX A100 Systems
- DDN A^{3}I Storage

#1 Supercomputer in India

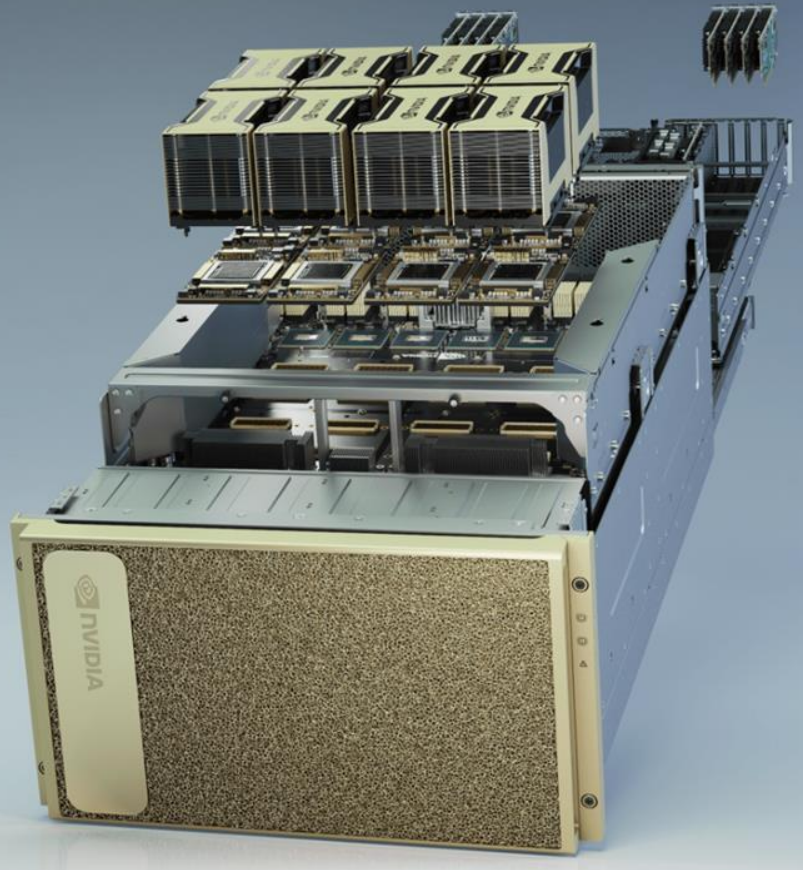
India's largest and greenest HPC-AI supercomputer

LEADING ORGANIZATIONS ARE ACHIEVING AI-AT-SCALE WITH DGX SUPERPOD

Argonne NATIONAL LABORATORY · CLOVA NAVER LINE · UF FLORIDA · SONY · MTS · AstraZeneca · KINGS COLLEGE LONDON · NHS Guy's and St Thomas' NHS Foundation Trust · gsk · LINKÖPING UNIVERSITY · VinAI

AND USING EASY TO DEPLOY AI INFRASTRUCTURE BUILT ON DGX A100

Martinos Center for Biomedical Imaging · St. Jude Children's Research Hospital · GE Healthcare · dkfz. GERMAN CANCER RESEARCH CENTER IN THE HELMHOLTZ ASSOCIATION · Kroger · SUBARU · BMW GROUP · NTT PC · SK telecom · TELCONET · CyberAgent · LOCKHEED MARTIN · Square · DFKI Deutsches Forschungszentrum für Künstliche Intelligenz GmbH · CMU · zenseact · UTSA The University of Texas at San Antonio · KIT · 國立臺灣大學 National Taiwan University · hhu · UNIVERSITÀ DI PISA · Mila · TECHNISCHE UNIVERSITÄT DARMSTADT · tu technische universität dortmund · 佛教慈濟醫療財團法人 Buddhist Tzu Chi Medical Foundation · CEDIA · 慈濟大學 Tzu Chi University · Riskfuel · Sandia National Laboratories · Pacific Northwest NATIONAL LABORATORY · NERSC National Energy Research Scientific Computing Center · TACC · ILLINOIS NCSA | National Center for Supercomputing Applications



ENTERPRISE IT AI CENTER OF EXCELLENCE

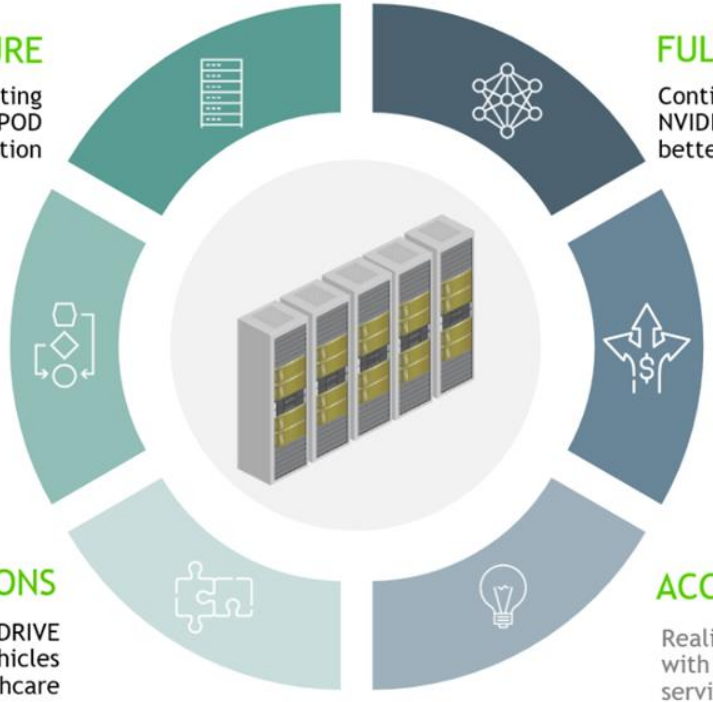
DGX: It's Much More Than A Box

APPLIANCE OR INFRASTRUCTURE

Deploy as a standalone computing appliance or full DGX POD infrastructure solution

FULLY OPTIMIZED AI STACK

Continuous software engineering from NVIDIA that delivers performance that gets better over time



SOFTWARE AND TOOLS

Pre-trained models, scripts, and more from NGC enable out-of-box productivity, and certified AI workflow solutions enable MLOps

FLEXIBLE DEPLOYMENT

Deployed in DGX-Ready Data Centers or your own data centers, with options for CAPEX or OPEX with leasing or -aaS

INDUSTRY SOLUTIONS

Solutions built on DGX such as DRIVE Constellation for autonomous vehicles or Clara for healthcare

ACCESS TO AI EXPERTISE

Realize results sooner with NVIDIA DGXperts, along with domain-specific knowledge from consulting services from our partners