NVIDIA DGX STATION™ A100

A WORKGROUP APPLIANCE
FOR THE AGE OF AI

DATA CENTER PERFORMANCE WITHOUT THE DATA CENTER

4X NVIDIA A100 TENSOR CORE GPUs

160 or 320 gigabytes (GB) total GPU memory. Fully interconnected with high-bandwidth, third-generation NVIDIA® NVLink® at 200 GB/s

7.68 TERABYTE (TB) PCIE GEN4 NVME SOLID-STATE DRIVE (SSD)

Delivers 1.4M IOPS storage performance, 2X faster than PCIe Gen3 NVMe SSDs

REFRIGERANT COOLING

Whisper quiet, a perfect solution for your desk while stil being optimized for performance

64-CORE AMD CPU AND PCIE GEN4

3.2X more cores to power multiple users and the most intensive AI jobs, 512GB system memory

NVIDIA DGX™ DISPLAY ADAPTER

4x Mini DisplayPort, 4K resolution

REMOTE MANAGEMENT

Integrated 1Gbase-T Ethernet baseboard management controller (BMC) port

2.5 PETAFLOPS of AI performance

3X FASTER average training performance than prior gen[1]

<1 HOUR from unpacking to up-and-running

2 CABLES and a floor is all you need to operate

0 DATA CENTER requirements; just plug in to any wall socket

[1] Inference: Batch Size=256; INT8 Precision; Synthetic Data; Sequence Length=128, cuDNN 8.0.4

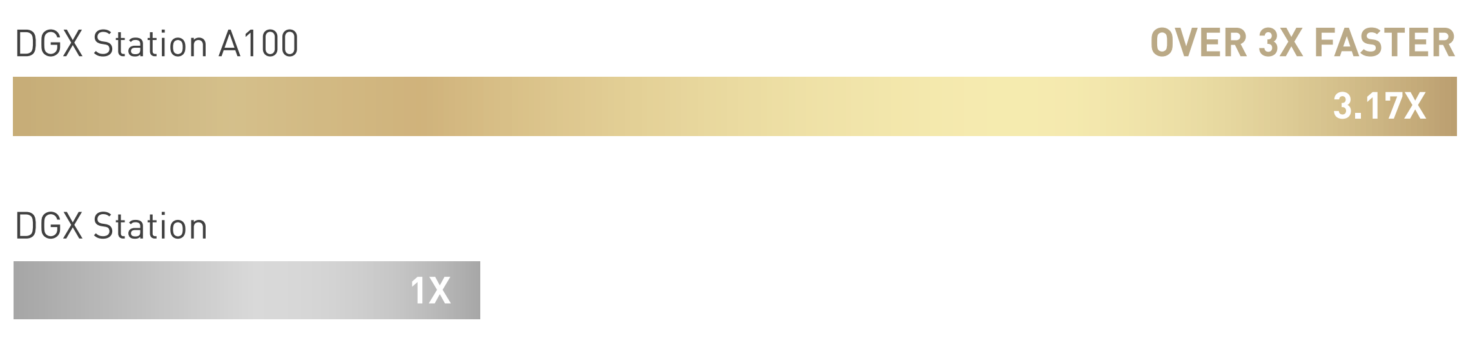
BIGGER MODELS, FASTER ANSWERS

UNPARALLELED AI PERFORMANCE

TRAINING

BERT Large Pre-Training Phase 1 (Relative Performance)

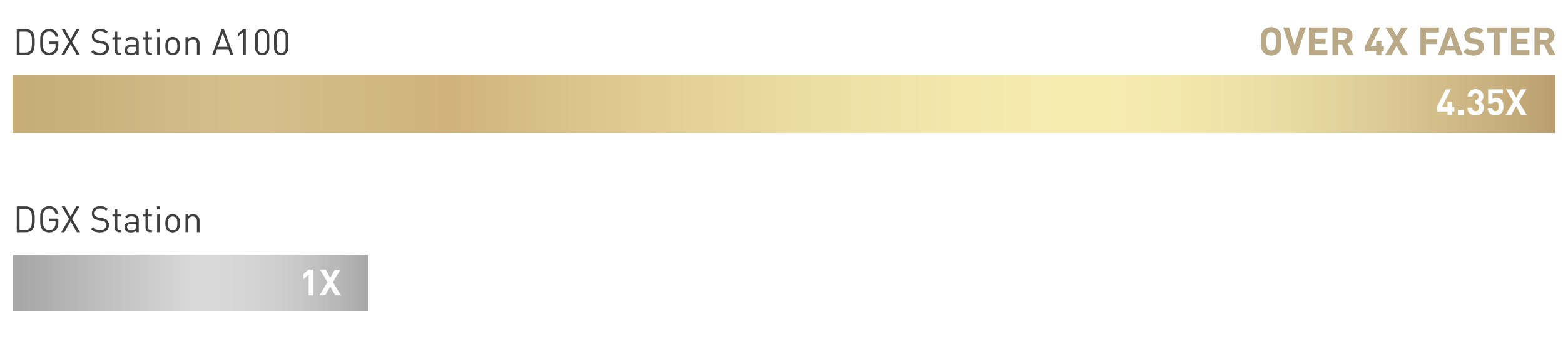
DGX Station A100 320GB; Batch Size=64; Mixed Precision; With AMP; Real Data; Sequence Length=128



INFERENCE

BERT Large Inference (Relative Performance)

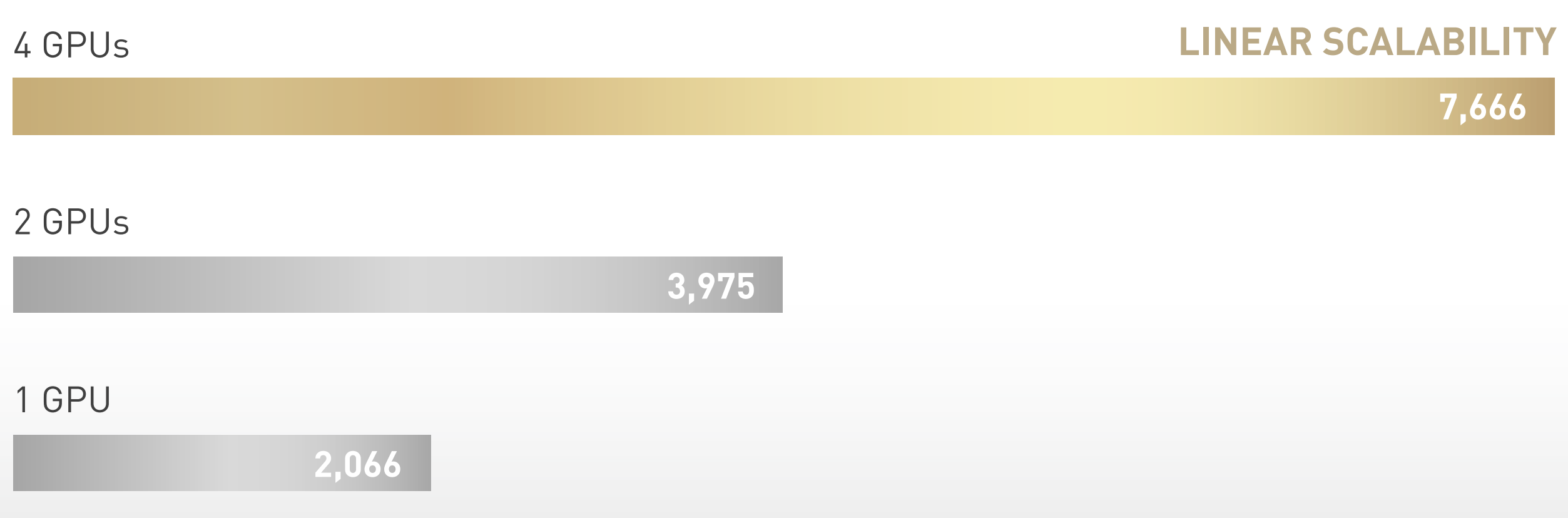
DGX Station A100 320GB; Batch Size=256; INT8 Precision; Synthetic Data; Sequence Length=128, cuDNN 8.0.4



MULTI-GPU SCALABILITY

ResNet-50 V1.5 Training (Images per Second)

DGX Station A100 320GB; Batch Size=192; Mixed Precision; Real Data; cuDNN Version=8.0.4; NCCL Version=2.7.8; NGC MXNet 20.10 Container



A POWERFUL TOOL FOR DATA SCIENCE TEAMS

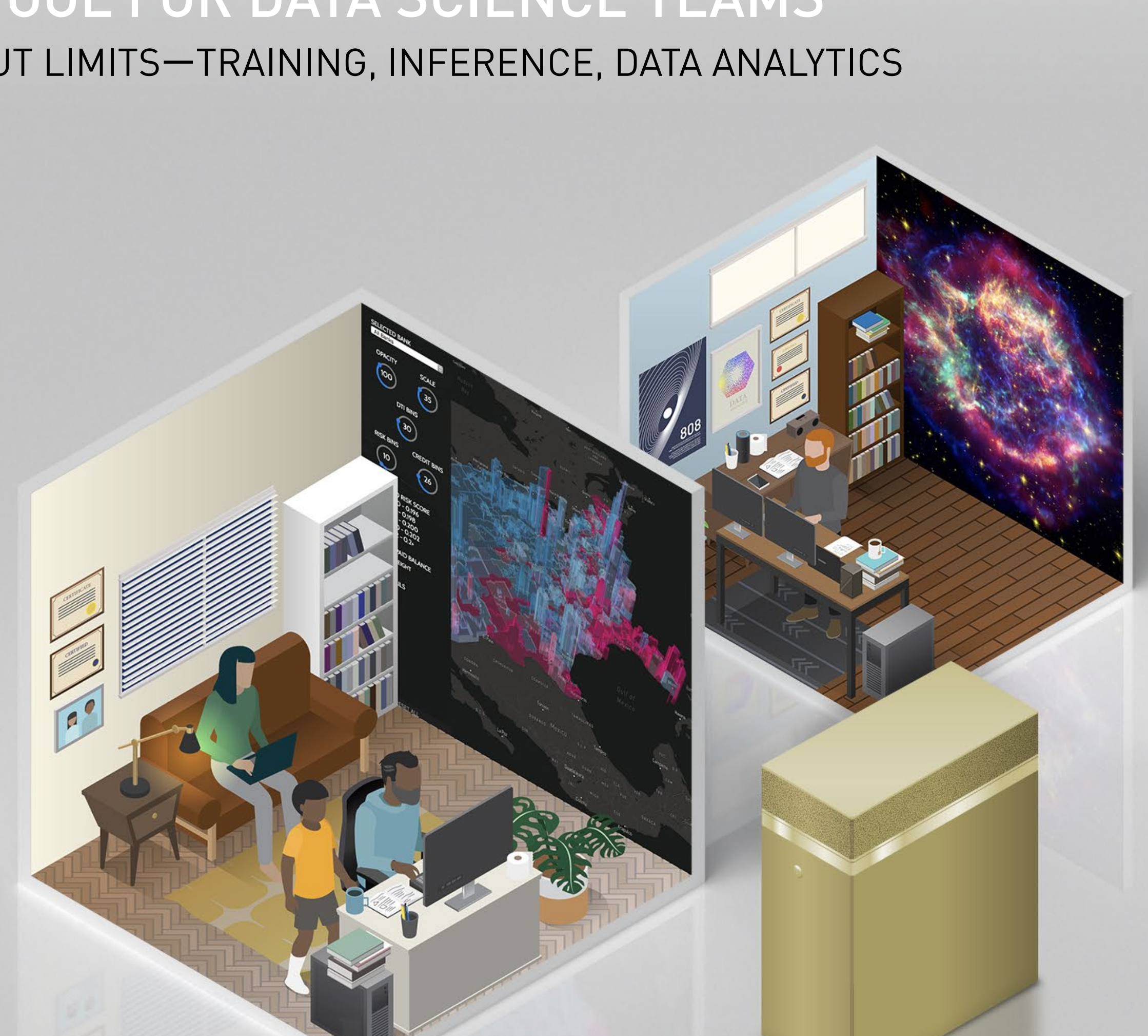
A SHARED SYSTEM WITHOUT LIMITS—TRAINING, INFERENCE, DATA ANALYTICS

Multi-Instance GPU (MIG) in a single NVIDIA DGX™ Station A100 gives

12 developers the performance equivalent to

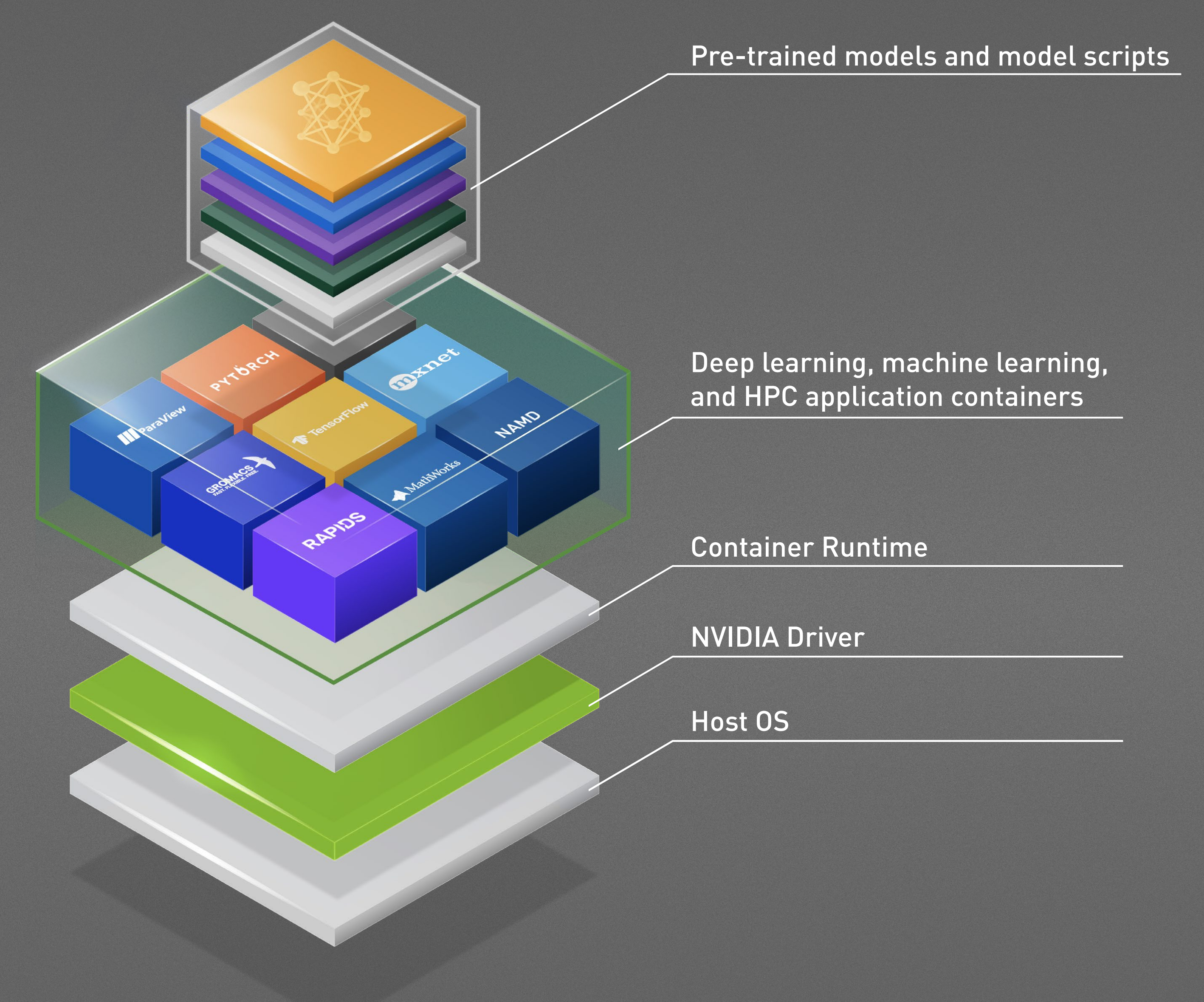
2 dedicated NVIDIA V100 Tensor Core GPUs each or

6 dedicated 28-dual core CPU servers each



FASTEST TIME TO INSIGHTS WITH NVIDIA AI OPTIMIZED SOFTWARE

FULLY INTEGRATED SOFTWARE STACK FOR INSTANT PRODUCTIVITY


Pre-trained models and model scripts
Deep learning, machine learning, and HPC application containers
PYTORCH
mxnet
ParaView
TensorFlow
NAMD
GROMACS
MathWorks
RAPIDS
Container Runtime
NVIDIA Driver
Host OS

Developed and Tested on DGX

Run your AI projects on the exact same platform NVIDIA engineers use to develop and test optimized AI software.

Always the Best Performance

Monthly updates to key AI tools and stack optimizations deliver better performance over time on the exact same hardware.

Get Results Sooner

Pre-trained models, scripts, and more translate to better results sooner over do-it-yourself problem solving.

Consistency Across DGX Systems

The same base operating system and quality-assurance testing ensure easy and predictable interoperability.

DIRECT ACCESS TO A GLOBAL TEAM OF NVIDIA DGXPERTS

GET UNMATCHED AI EXPERTISE WITH EVERY DGX SYSTEM

NVIDIA With You Every Step of the Way

Design | Plan | Build | Test | Deploy | Operate | Monitor

10+ years of AI innovation

100+ GPU-optimized software and tools on NGC™

10,000+ of AI-fluent practitioners around the globe

Experiment, Prototype, Develop. From Anywhere.