THE FUTURE WAS HERE!

Denon Chang May. 2021

CHALLENGES IN HIGHER EDUCATION

BATTLE FOR FUNDING
More researchers competing for level funding dollars. IT budgets not growing to match technology infrastructure demands

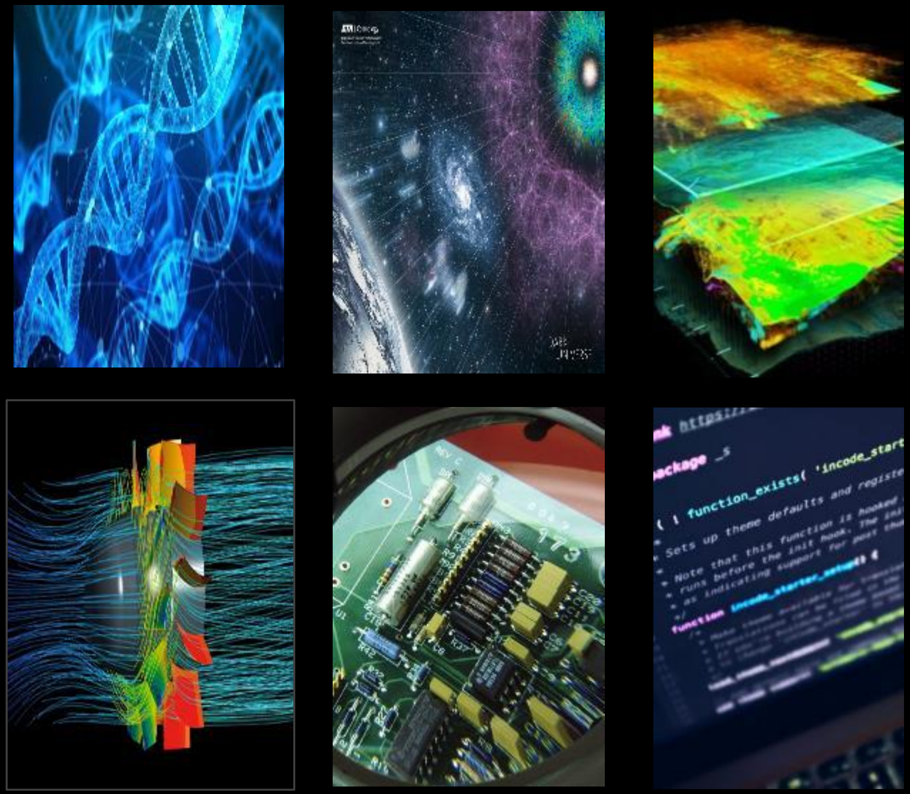
ATTRACTING BEST TALENT
Funds and lab equipment deciding factor for faculty & students

CHANGING CURRICULA
Demand for skilled data scientists & AI expertise requires students to hit the ground running after graduation

COMPLEXITY
Research problems, student projects more complex than ever; data sets & compute requirements growing exponentially

IN-PERSON & REMOTE LEARNING
Distance learning at-scale introduces a new set of challenges for educators, researchers, students and IT-Staff

CURRICULA AND RESEARCH FOR TODAY AND TOMORROW

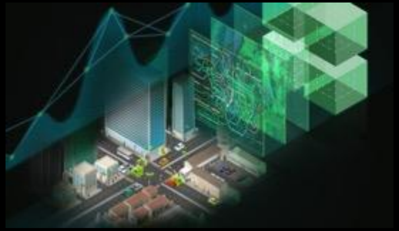


SCIENCE

DATA SCIENCE

AI

DATA SCIENCE IS THE KEY TO MODERN BUSINESS



CONSUMER INTERNET

Ad Personalization
Click Through Rate Optimization
Churn Reduction

FINANCIAL SERVICES

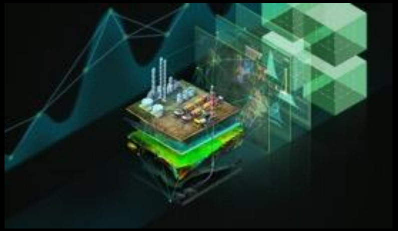
Claim Fraud
Customer Service Chatbots/Routing
Risk Evaluation

HEALTHCARE

Improve Clinical Care
Drive Operational Efficiency
Speed Up Drug Discovery

RETAIL

Supply Chain & Inventory Management
Price Management / Markdown Optimization
Promotion Prioritization And Ad Targeting



OIL & GAS

Sensor Data Tag Mapping
Anomaly Detection
Robust Fault Prediction

MANUFACTURING

Remaining Useful Life Estimation
Failure Prediction
Demand Forecasting

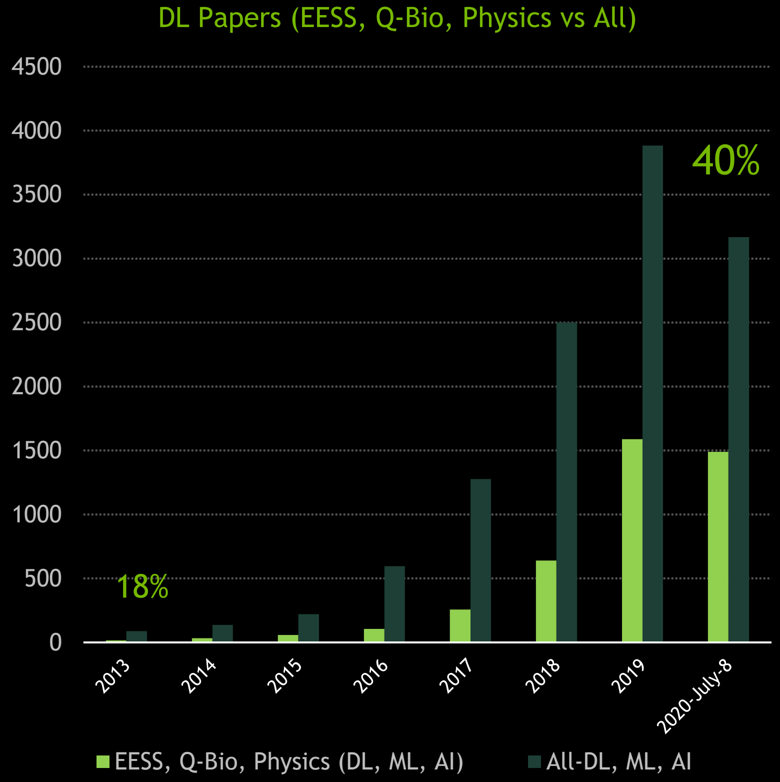
TELECOM

Detect Network/Security Anomalies
Forecasting Network Performance
Network Resource Optimization (SON)

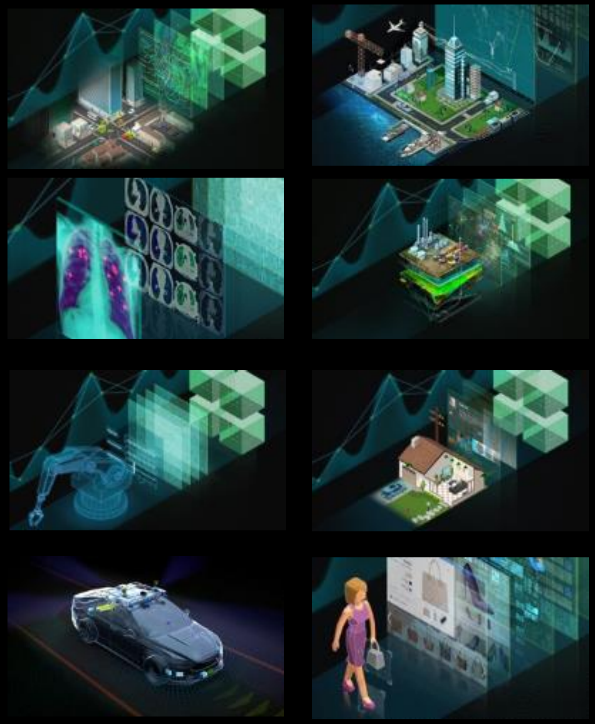
AUTOMOTIVE

Personalization & Intelligent Customer Interactions
Connected Vehicle Predictive Maintenance
Forecasting, Demand, & Capacity Planning

DEMAND FOR AI IS GROWING FAST



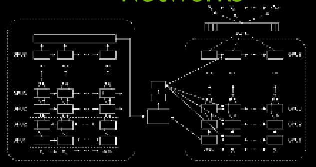
SCIENCE + AI PAPERS IN ARVIX



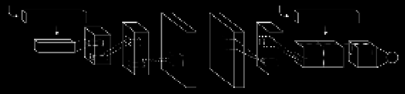
AI IS TRANSFORMING EVERY INDUSTRY

Convolutional Networks

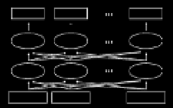
Recurrent Networks



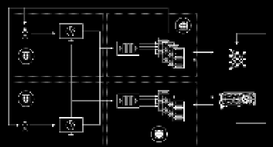
Generative Adversarial Networks



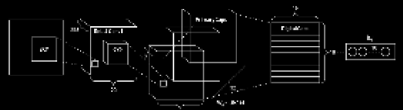
Transformer



Reinforcement Learning



New Species

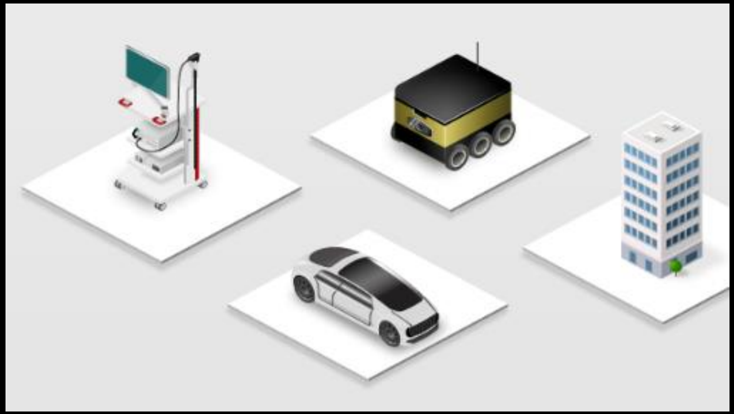


EXPLOSION IN AI NETWORKS

BUILDING AI LEADERSHIP IN HER

ADDRESSES THE CHALLENGES AND DEMANDS

Research that solves society's biggest problems



Partnerships with Government and Industry

Attract Top Talent and Funding

Growing Demand for AI Experts

NVIDIA IN HIGHER EDUCATION AND RESEARCH

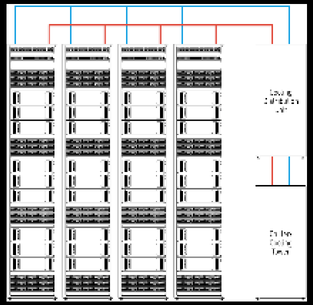
ECOSYSTEM, EXPERTISE AND SOLUTIONS FOR AI LEADERSHIP IN EDUCATION

700+Applications
1.8Million Developers

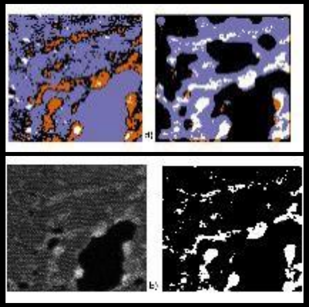
APPLICATION CONTAINERS	AI MODELS	HELM CHARTS
150+	100+	ML, Inference

TOOLKITS & SDKs

Healthcare | Smart Cities | Conversational AI | Robotics | HPC | more



Best Practices



Research Promotion

Teaching Kits

Workshops/DLI/Hackathons

DGX FOR RESEARCHES

Most Powerful Machine for Data Scientists and Researchers

RESEARCH
LAB-IN-A-BOX

BEST IN CLASS
PERFORMANCE

INCREASING
PERFORMANCE

EFFORTLESS
PRODUCTIVITY

END TO END
PORTFOLIO FOR AI

DGX STATION A100

AI APPLIANCE ARE A GAME-CHANGER

IDC Sees the Trend of GPU-based Systems at the Researcher's or Developer’s Office*

“DGX Station just works! It was up and ready within a few hours… Instead of worrying about how to configure many low-level components on the system, I can focus on gathering the right data, training the AI workloads, and working with experts to identify faults accurately.”

NATHALIE RAUSCHMAYR
Machine Learning Engineer, CSEM (Swiss Research and Development Center)

SBB CFF FFS

“It’s no exaggeration to say that we depend on our DGX Station every day, sometimes every hour of every day. It’s an enormous advantage in speeding up our work and getting to market fast.”

DANNY ATSMON
CEO, Cognata

cognata

*IDC Technology Spotlight. Accelerated Workstation: Run Deep Learning Workloads at Your Desk. December 2017. IDC#US42998817

AI INNOVATION BOTTLENECKS

LACK OF RESOURCES

Constantly waiting for resource availability

NO DATA CENTER

Lack of AI compute infrastructure

IT CONSTRAINTS

Long IT setup times needing physical data center access

SLOW PERFORMANCE

Lack performance for fast iteration

INTRODUCING NVIDIA DGX STATION A100

The Workgroup Appliance for the Age of AI

INTRODUCING NVIDIA DGX STATION A100

The Workgroup Appliance for the Age of AI

#AI #NVIDIADGX

Introducing NVIDIA DGX Station A100

https://youtu.be/TKtN04z7Q5Q

DGX STATION A100

The Workgroup Appliance for the Age of AI

AI SUPERCOMPUTING FOR DATA SCIENCE TEAMS

A shared system that your team can use without limits for all workloads - training, inference, data analytics, HPC

DATA CENTER PERFORMANCE WITHOUT THE DATA CENTER

A server-grade, plug-and-go AI system that doesn't require data center power and cooling

AN AI APPLIANCE YOU CAN PLACE ANYWHERE

Full server-class remote management, no complicated installation or additional IT infrastructure needed

BIGGER MODELS, FASTER ANSWERS

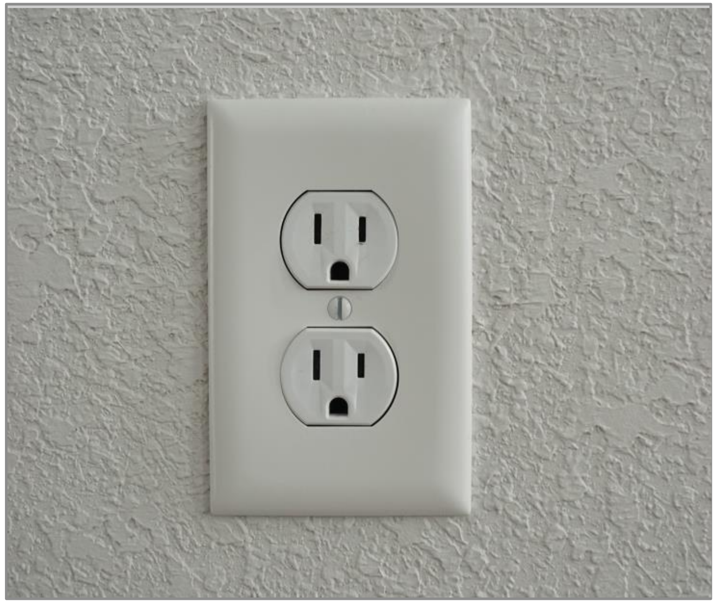
The world's only workstation-style system with four fully interconnected NVIDIA A100 data center GPUs

BREAKING THE DATA CENTER BARRIER

A Supercomputer With Just Two Cables

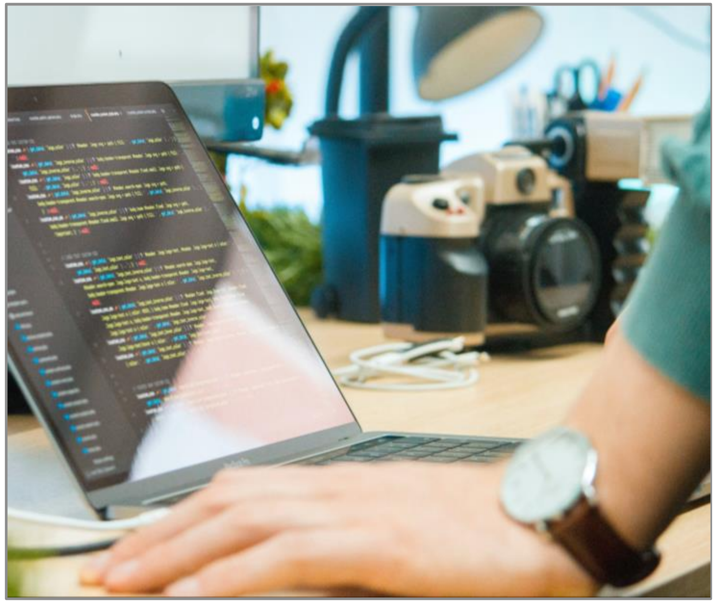
No Data Center, No Problem!

A fully functional AI system out-of-the-box, a whisper-quiet solution



Work from Anywhere AI Appliance

Plug into any standard wall socket, and access resources whether you are in the office, home office, or thousands of miles away



Instant Productivity

Unpack to up-and-running in under an hour, now with server-class remote management capabilities

BIGGER MODELS, FASTER ANSWERS

The Next Evolution in AI platforms for Today's Work from Anywhere Reality

Only workstation-style system with four fully interconnected NVIDIA A100 GPUs

Largest GPU memory available in a workstation, up to 320GB

2.5x*, on average, faster compute compared to previous DGX Station

* Varies, depending on workload, between 1.4x and 4.5x

SERVER-CLASS SOLUTION IN A WORKSTATION PACKAGE

Data Center Technology Outside the Data Center

First and only workstation with 4-way NVIDIA HGX A100

Four A100 Tensor Core GPUs, up to 320GB total HBM2E

3rd generation NVLink

200GB/s bi-directional bandwidth between any GPU pair, almost 3x compared to PCIe Gen4

New Cooling System, Pump Refrigerant 2-Phase Cooling

Maintenance-free, sealed system

No need to check, or refill, water-level

Single loop for CPU and four GPUs

Non-toxic, non-flammable, non-condensing

PURPOSE BUILT FOR AI WORKLOADS

Data Center-Class Technology Inside

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)

	DGX Station A100 320GB	DGX Station A100 160GB
GPUs	4x NVIDIA A100 Tensor Core GPUs	
GPU Memory (total)	320GB	160GB
Performance	2.5 petaFLOPS AI; 5 petaOPS INT8	
System Memory	512GB DDR4 RDIMM, 3200MT/s	
Storage	OS: 1 x 1.92TB M.2 NVME Data:1 x 7.68TB U.2 NVME	
CPU	AMD® Epyc® CPU 7742, 2.25GHz to 3.4GHz, 64 cores/128 threads, PCIe Gen4	
Networking	Dual 10GBASE-T (RJ45)	
Display GPU	4GB, 4x Mini DisplayPort	
Acoustics	<37dB	
Cooling	Custom refrigerant cooling system for GPUs and CPU	
System Power (max)	1,5kW	
Management	AST2500, IPMI, Redfish	
System Dimensions	518 D x 256 W x 639 H (mm)	
Operating Temp.	5°C to 35°C (41°F to 95°F)	

THE WORKGROUP APPLIANCE FOR THE AGE OF AI

A Powerful Tool For Data Science Teams

One DGX Station A100 delivers:

2.5 petaFLOPS of AI training power

5.0 petaOPS INT8 of inference

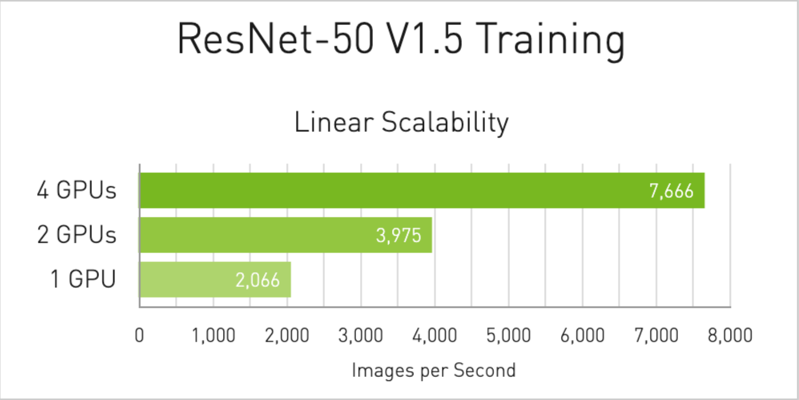
With MIG (Multi-Instance GPU), you can slice up individual GPUs, and a team of, for example, 4, 8, 12 developers can share a DGX Station A100, where:

- Simultaneous workloads can be executed with guaranteed Quality Of Service:
- Flexibility to run any type of workload on a MIG instance
- Different sized MIG instances based on target workloads

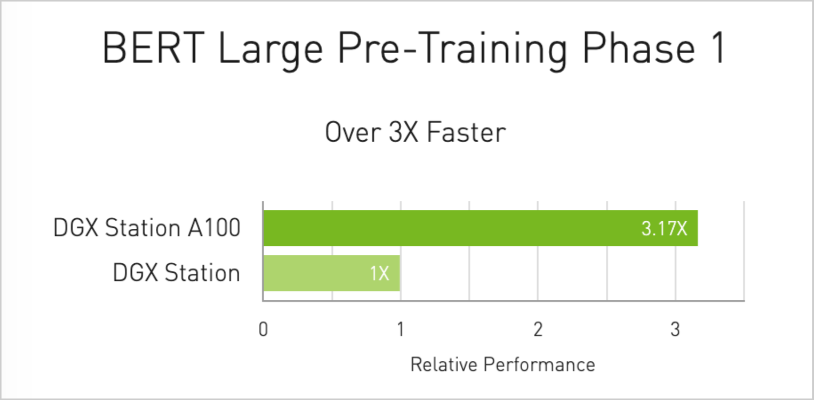
The only workstation-style system with support for MIG!

A DATA CENTER IN-A-BOX

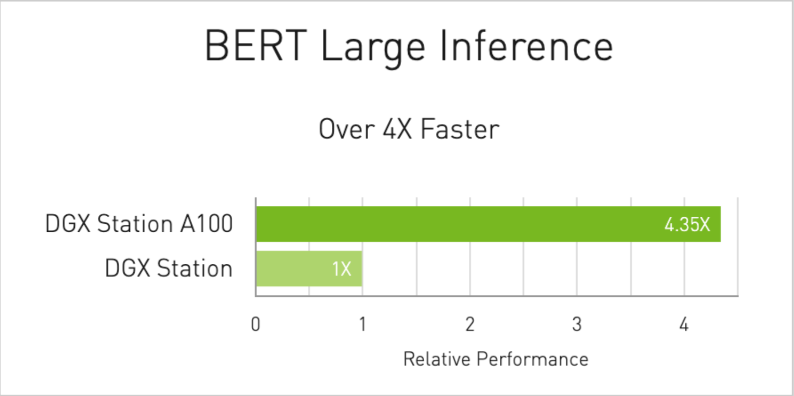
DGX Station A100 is More Than 4X Faster



SCALABILITY

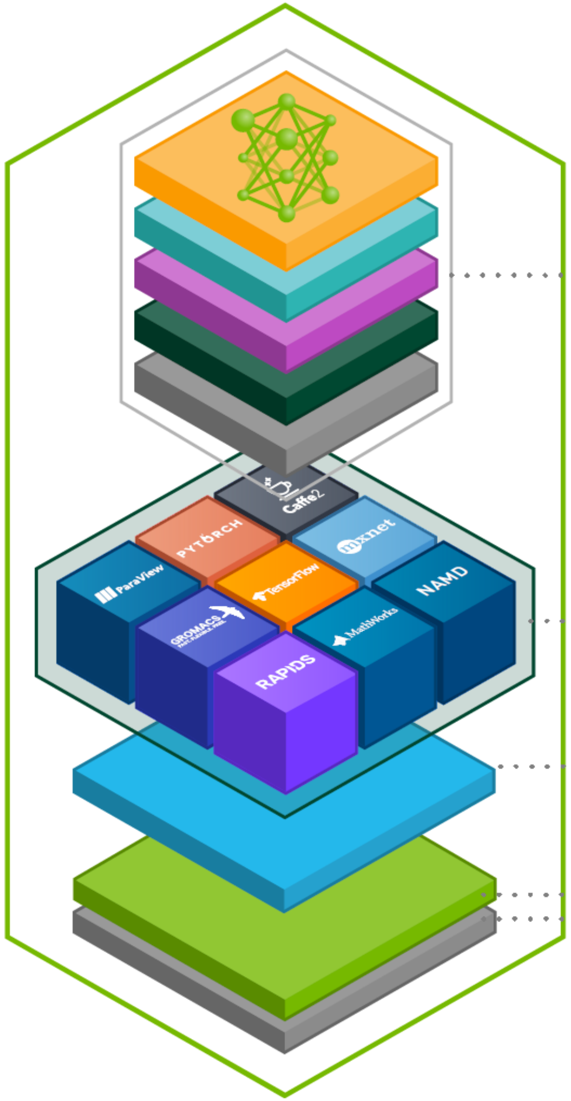


TRAINING



INFERENCE

Training: Batch Size=64; Mixed Precision; With AMP; Real Data; Sequence Length=128
Inference: Batch Size=256; INT8 Precision; Synthetic Data; Sequence Length=128, cuDNN 8.0.4
HPC: FP32 Precision; Dataset/Input=Cellulose (h-bond) | Best value listed. Average is 1.5X across all these inputs: ADH Dodec (h-bond); Cellulose (h-bond); STMV (h-bond)



PRE-TRAINED MODELS AND MODEL SCRIPTS

DEEP LEARNING, MACHINE LEARNING, HPC APPLICATION CONTAINERS

CONTAINER RUNTIME

NVIDIA DRIVER

HOST OS

NGC SOFTWARE STACK

DGX STATION SOFTWARE STACK

Pre-Installed, Integrated Software Built for Instant Productivity

Advantages:

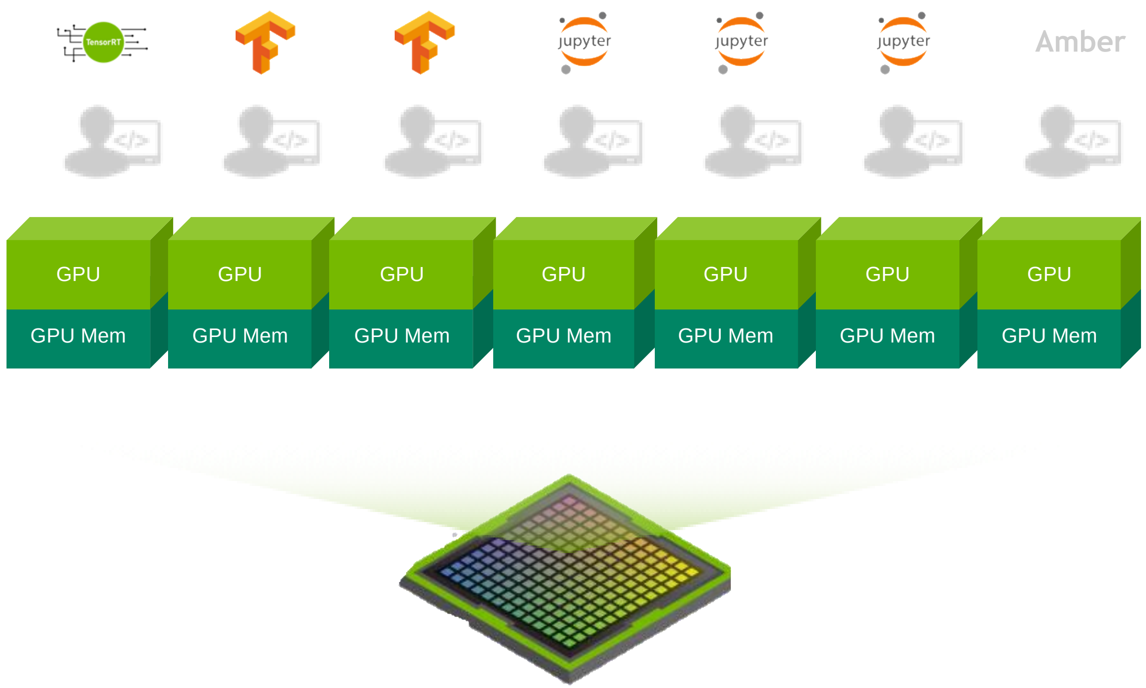
Fully tested and optimized DGX software stack, including an AI-tuned base operating system, all necessary system software, GPU driver, CUDA, libraries

Faster Time-to-Insight with pre-built, tested, and ready to run containers from NGC

- Containers for DL training & inference, HPC, analytics, and industry-specific applications
- Container portability, flexibility, repeatability
- Continuous (monthly) performance improvements
- Pre-trained models and model scripts
- Private Registry for DGX customer
- Scalable with support for multi-GPU and multi-node systems

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

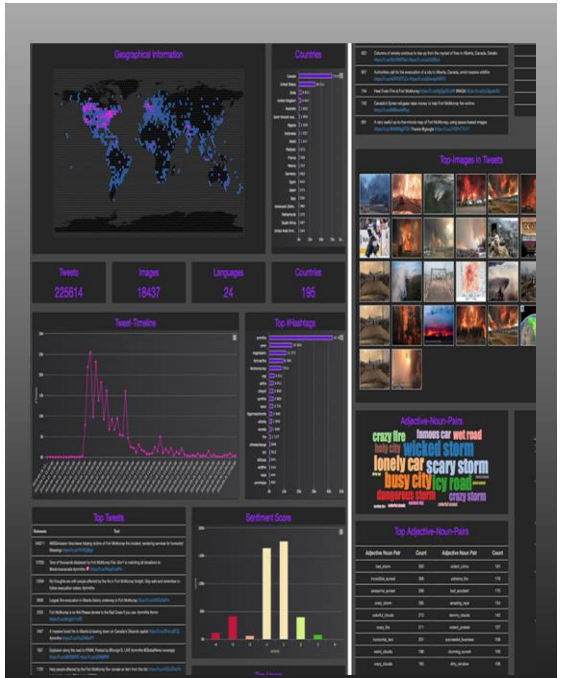
DGX Station A100 is the only workstation-style system that supports MIG

NVIDIA DGX STATION CASE STUDIES

AI workstation for leading-edge innovative development

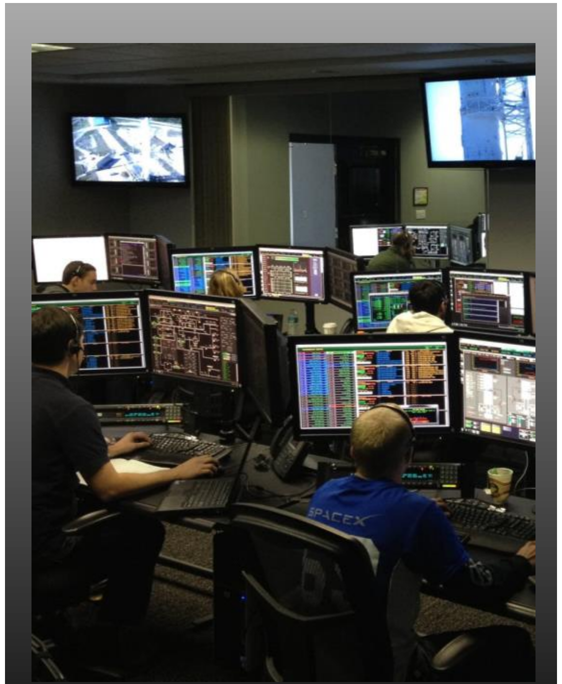
BMW

Explore insights faster in the development and deployment of AI models that improve operations



German Research Center for Artificial Intelligence

Build AI models that include computer vision which help emergency services respond rapidly to natural disasters



Pacific Northwest NATIONAL LABORATORY

Conduct federally funded research in support of national security

NTT docomo

Develop innovative AI-driven services such as its image recognition solution for over 79 million subscribers

ADOPTED BY LEADING COMPANIES ACROSS INDUSTRIES

DGX Station Delivers AI Supercomputing to More Teams, From Anywhere

6
Of the Top 10
US Government
Institutions

6
Of the Top 10
Global Car
Manufacturers

7
of the Top 10
US Hospitals

10
Of the Top 10
Aerospace & Defense
Companies

Argonne NATIONAL LABORATORY · BMW · Carnegie Mellon University · DFKI German Research Center for Artificial Intelligence · NTT docomo · GE Healthcare · HONDA The Power of Dreams · LOCKHEED MARTIN · Martinos Center

MIT Massachusetts Institute of Technology · MUSASHI Power to Value · Microsoft · NUANCE · Pacific Northwest NATIONAL LABORATORY · P&G · SK hynix · SUBARU · Tencent

DATA SCIENCE TEAMS ON DGX STATION

Sharing Their Excitement

Adrià Romero
@iamrosmarin

Can anyone imagine a better thing to pack inside a cardboard box 📦? #DGXStation has arrived to @TriageAI @nvidia @NvidiaAI

11:03 AM - 26 Nov 2017

5 Retweets 24 Likes

Serge Lemonde
AI/Deep learning Startups Business director, EMEA and India at NVIDIA

Luca Baldassarre, Head of Data Science @Gamaya in Switzerland, unpacking the DGX Station SuperComputer they won at the Inception awards during GTC Europe. Great team, can't wait to see how it impacts their research!

5 Likes

239 views of your post in the feed

2 NOVEMBER 2017 / DGX

The DGX has arrived!

DGX STATION A100

Workgroup Appliance for the Age of AI

AI Supercomputing for Data Science Teams

Data center performance without the data center

An AI appliance you can place anywhere

Bigger models, faster answers

2.5 PFLOPS AI

320 GB GPU MEMORY

Only workstation with 4-way NVLink and Multi-instance GPU (MIG)