TRAIN THE TRAINER
NVIDIA AI ENTERPRISE

RSTOCKER@NVIDIA.COM

FEBRUARY 2022

AGENDA

Intro & Challenge

NVIDIA AI Enterprise

GNC & Launchpad

How to position and where to start

FAQ´s & Q&A

AI IN THE INDUSTRY

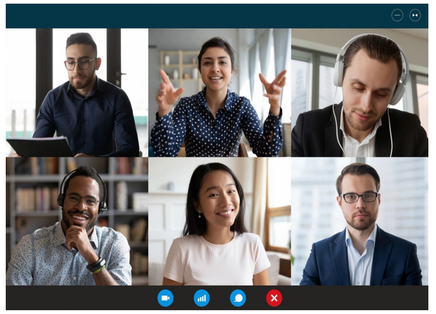
AI IS TRANSFORMING EVERY INDUSTRY

Demand for Fast, Easy Inference Deployment Greater than Ever

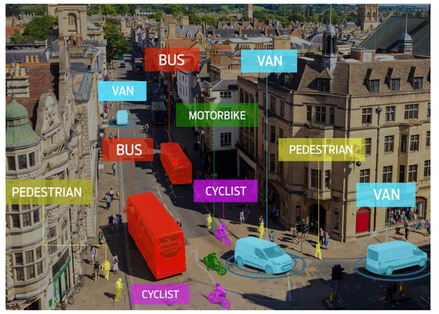
CREDIT CARD FRAUD
1.1B Credit Transactions / Day

CONTACT CENTER AI
500M Calls / Day

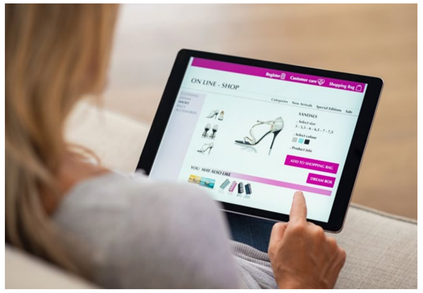
MEETING TRANSCRIPTION
15B Meeting Minutes / Day



PUBLIC SAFETY
> 1B Smart City Cameras Deployed



PRODUCT RECOMMENDATIONS
300M E-commerce Visitors / Day



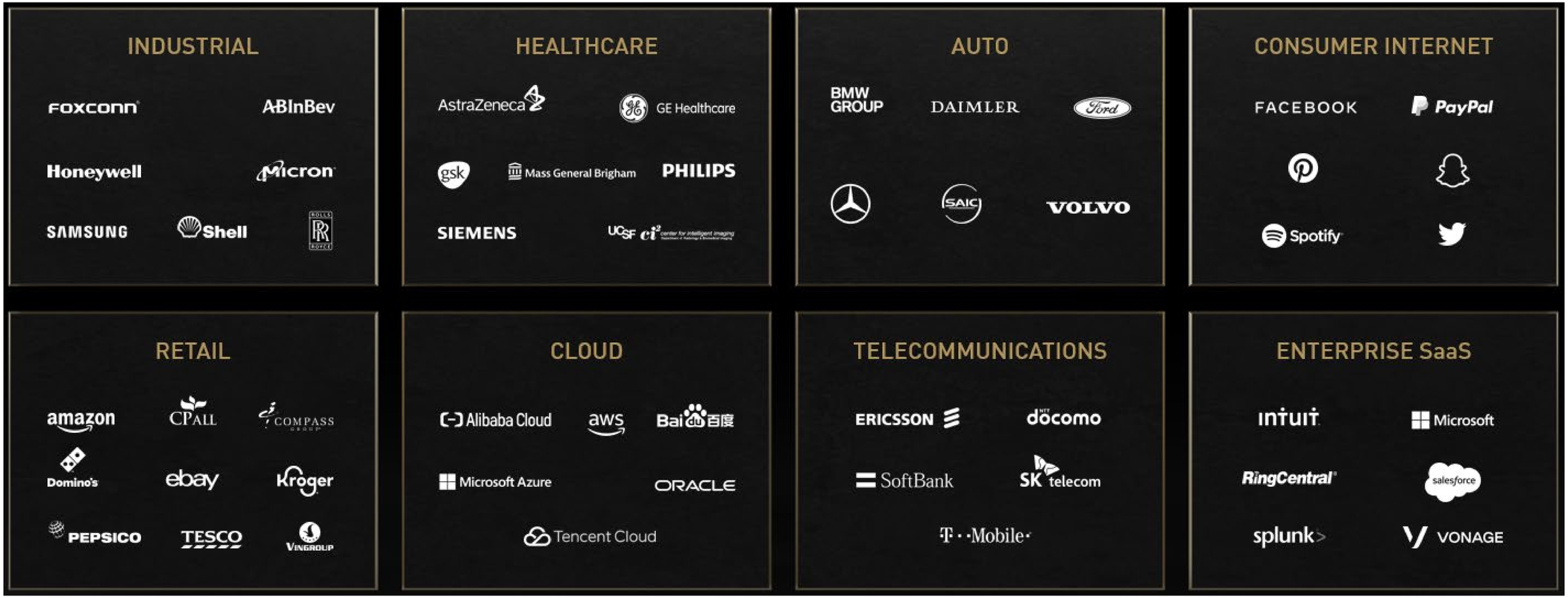
RETAIL ASSET PROTECTION
$275M Inventory Loss / Day

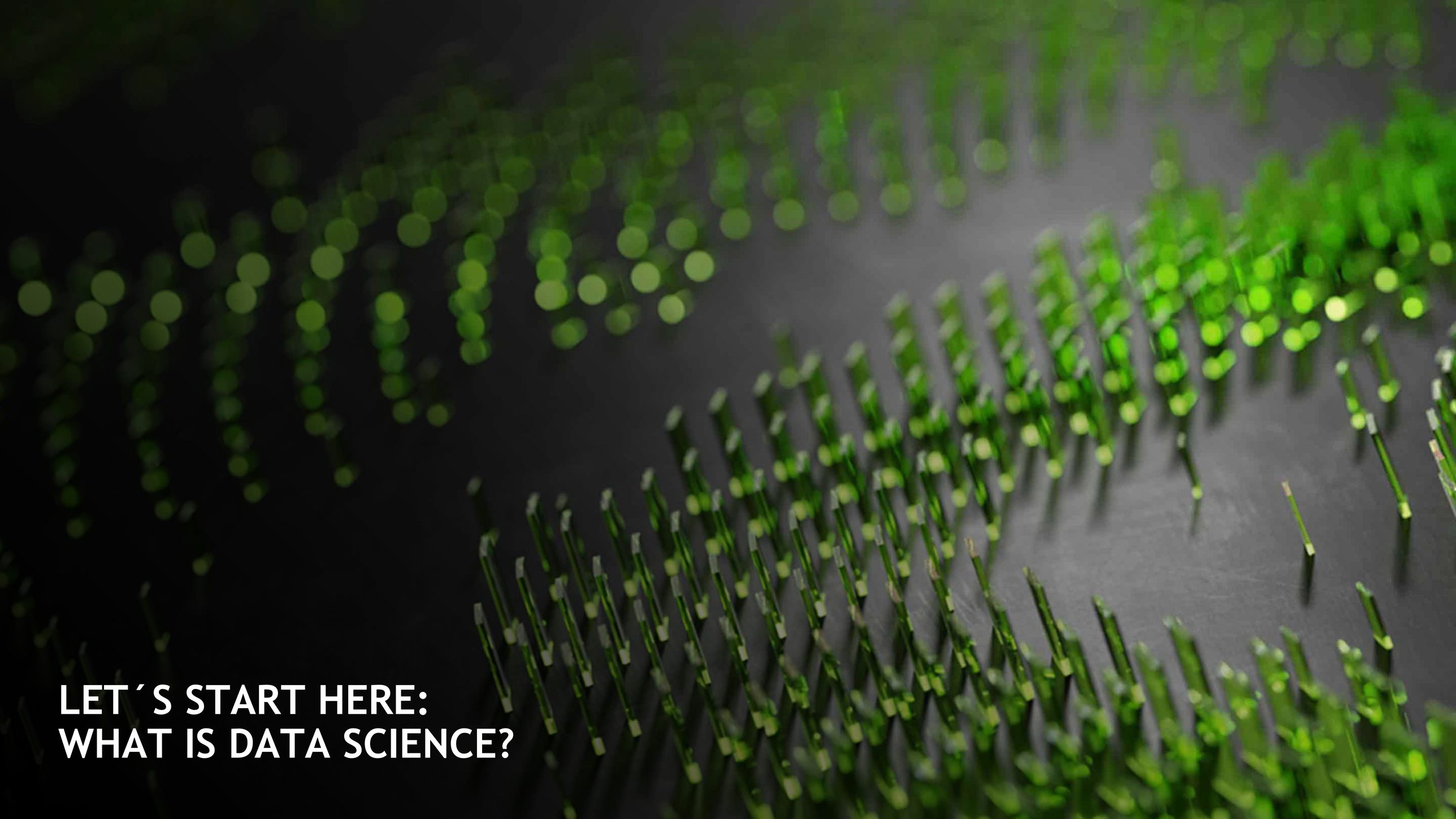
MEDICAL IMAGING
10M Diagnostic Scans / Day

INDUSTRIAL INSPECTION
94M Vision Sensors Installed by 2025

NVIDIA

LEADING COMPANIES RUNNING NVIDIA AI





LET´S START HERE:
WHAT IS DATA SCIENCE?

OVER 2 QUINTILLION (21X0) BYTES OF DATA IS PRODUCED EVERY DAY
WHERE IS IT?

DATABASES, SPREADSHEETS, PAPER, ARCHIVES, CD'S….ETC ETC ETC

- WHAT CAN WE DO WITH IT?
- IS IT USEFUL?
- CAN WE MONETIZE IT?
- DO WE NEED TO USE IT REAL TIME?

WHAT IS DATA SCIENCE?

It is the Extraction of insights and knowledge from any type of data

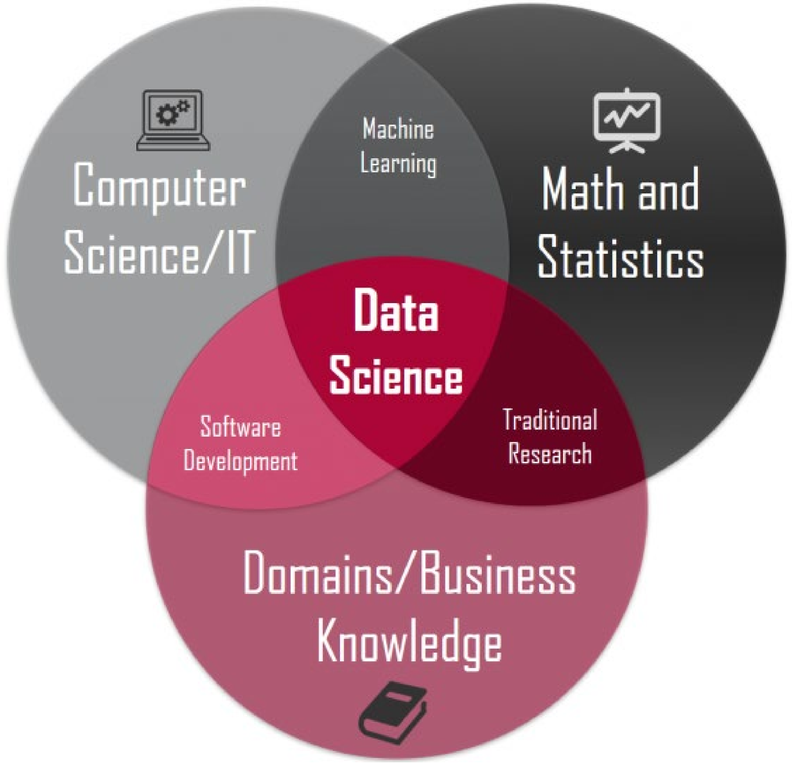
….for business…. for research…. for technological advances….
for global warming…. for humanity

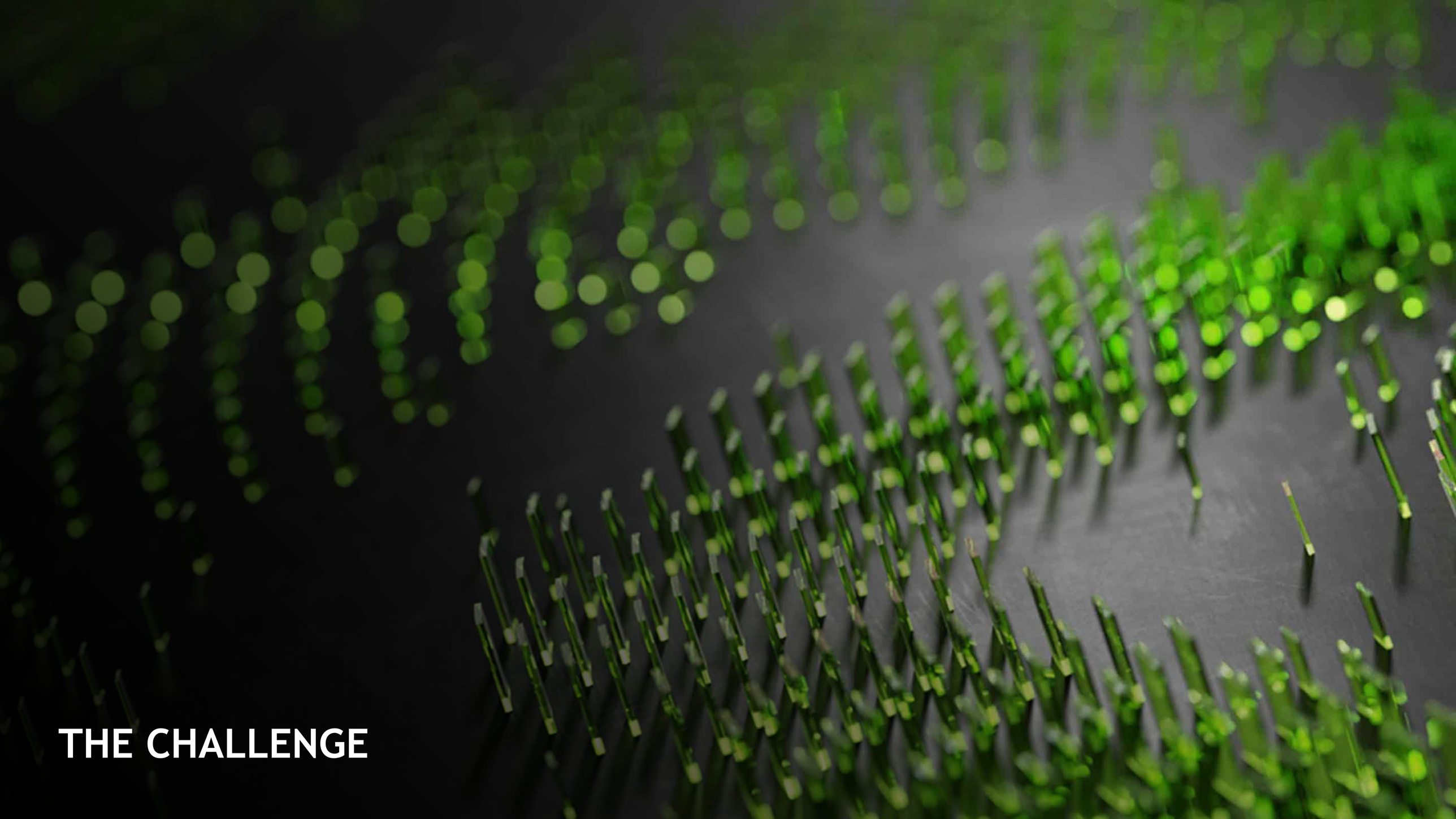
WHAT DOES A DATA SCIENTIST LOOK LIKE?

COMBINING - MATHS & STATISTICS, CODING, RESEARCH, ALGORITHMS……

SOLVING PROBLEMS WITH DATA, MANIPULATING DATA, EXPLORING DATA

- Degrees & PhDs In Computer Science/Mathematics/Programming/Economics etc
- Access high performance computing platforms
- Work with iterative and experimental workflows
- Not generally part of IT, part of the business near lots of data
- Create their own software stack and applications





THE CHALLENGE

MARKET TRENDS BY THE NUMBERS

AI Adoption

67%

of enterprises are using machine learning today - 97% are planning to use it in the next year*

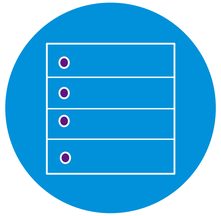
60%

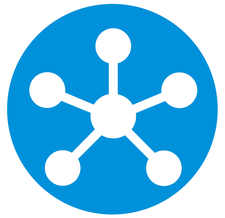
use enterprise software with AI baked in*

Compute Infrastructure Design

80% - 90%

of data centers rely on server virtualization for manageability, security and flexibility**





80%

of organizations will be using hyperconverged solutions by 2025 – double the percent in 2020***

*Deloitte. 2020 State of AI in the Enterprise, 3rd Edition. July 14, 2020.
**Gartner. 2020 Strategic Roadmap for Compute Infrastructure. September 16, 2020. ID G00723203
***Gartner. Market Guide for Server Virtualization. November 3, 2020. ID G00725264

NVIDIA.

AI CHALLENGES AND BARRIERS

53%

Average proportion of projects that make it from pilot to production

n = 603 All Respondents, Excluding Unsure
Q13. What is your best guess of the proportion of AI prototypes that make it to production?

30%

Cite complexity of AI solution integration with existing infrastructure as a top barrier

30%

Cite security or privacy concerns as a top barrier

n = 601 All Respondents, Excluding Not Sure
Q18. What are the top 3 barriers to the implementation of AI techniques within your organization?

Source: Gartner “P-19019 AI in Organisations”, Claudia Ramos, Erick Brethenoux, 2020

AI DEPLOYMENT CHALLENGES

Risk

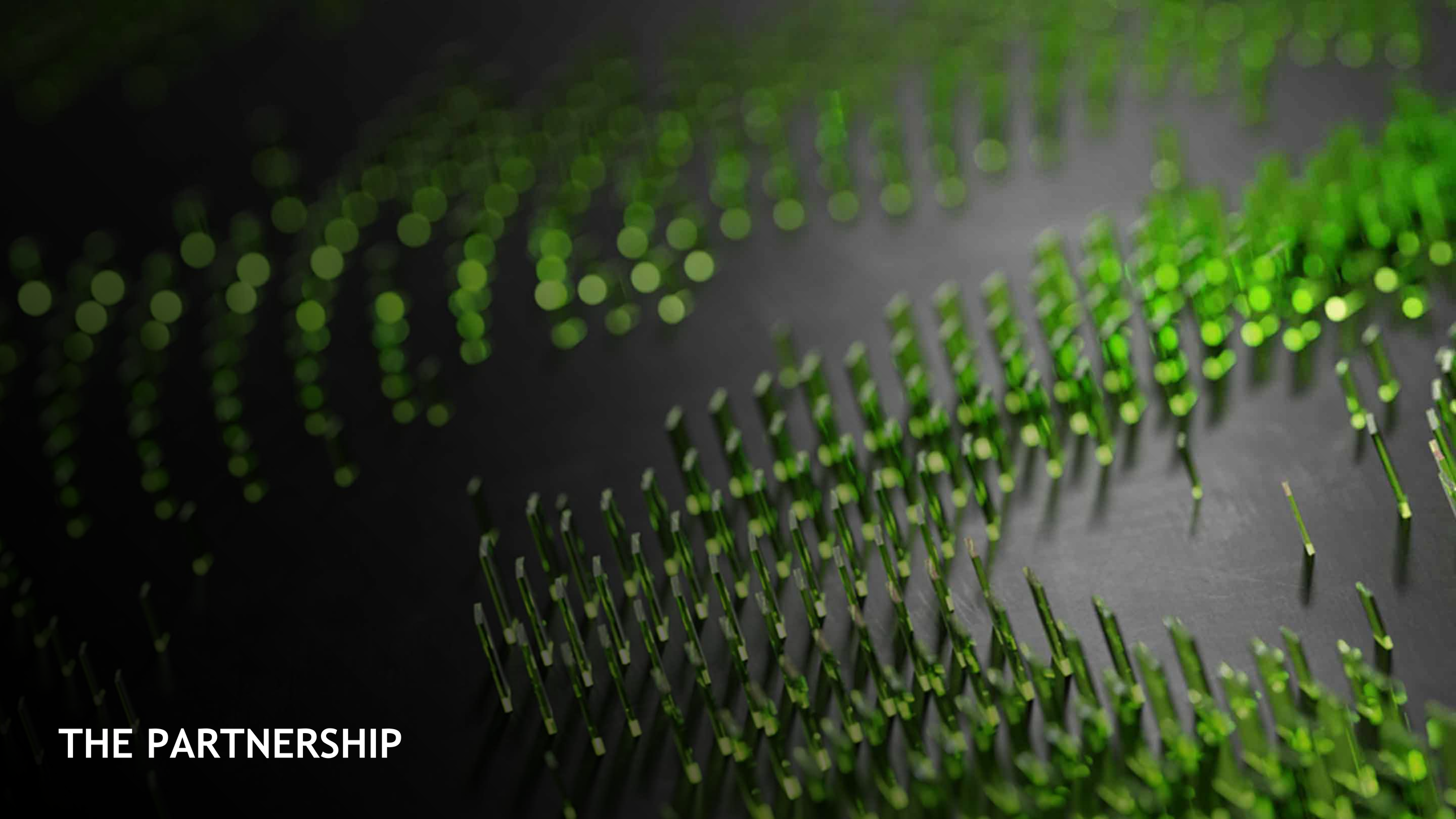
- Pulling together piece-parts
- Integrating with existing infrastructure

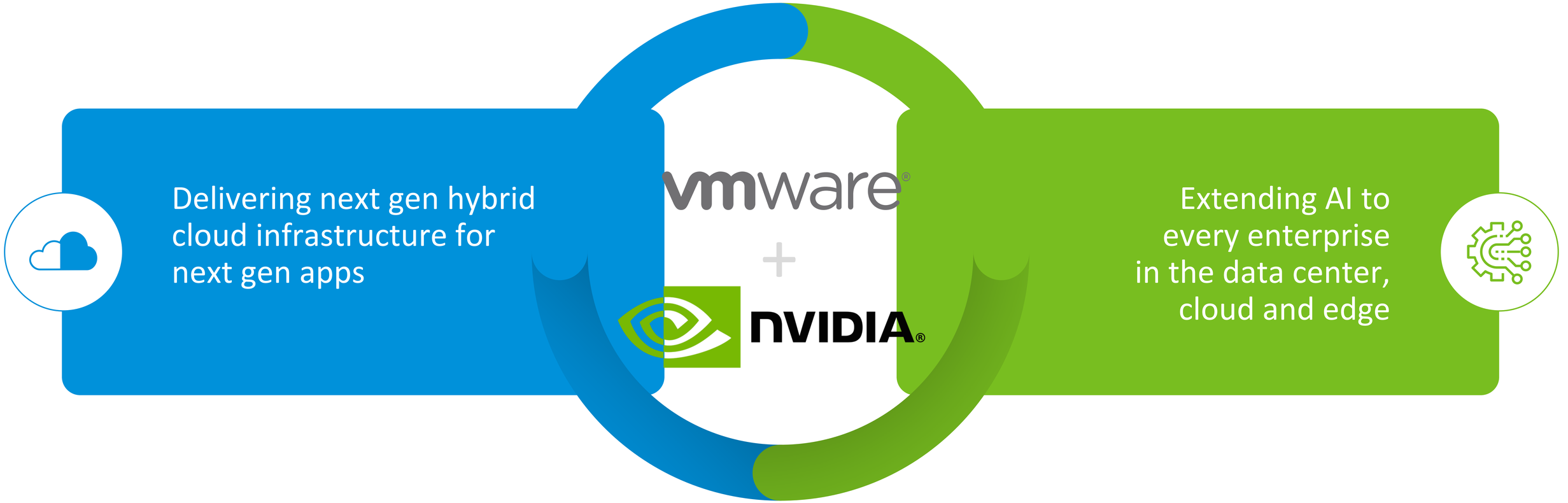
Performance

- Compute performance crunching through data
- Time to train models

Scaling

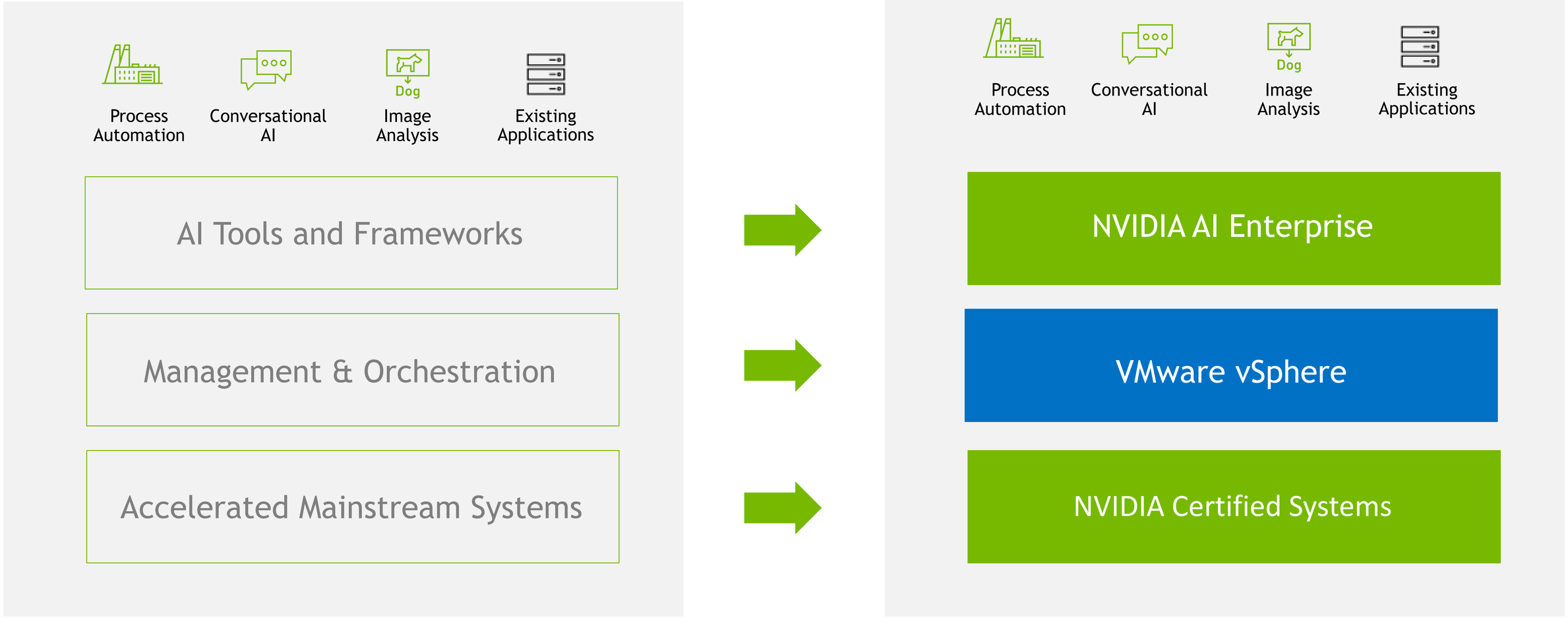
- Sharing GPU resources
- Maintaining manageability and availability


THE PARTNERSHIP


Delivering next gen hybrid cloud infrastructure for next gen apps
vmware
+
NVIDIA
Extending AI to every enterprise in the data center, cloud and edge

AI OPTIMIZED FOR VIRTUALIZED DATA CENTERS

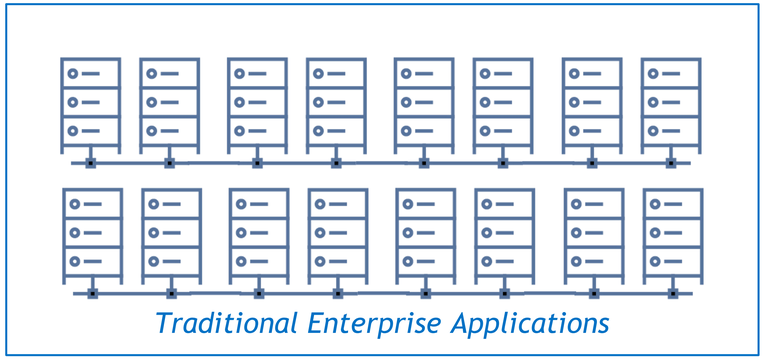
VMware vSphere is the Virtualization Platform for 70% of Enterprise Data Centers



PATH TO A SIMPLIFIED ACCELERATED DATA CENTER

Prepare for the Future while Driving Down Data Center Costs

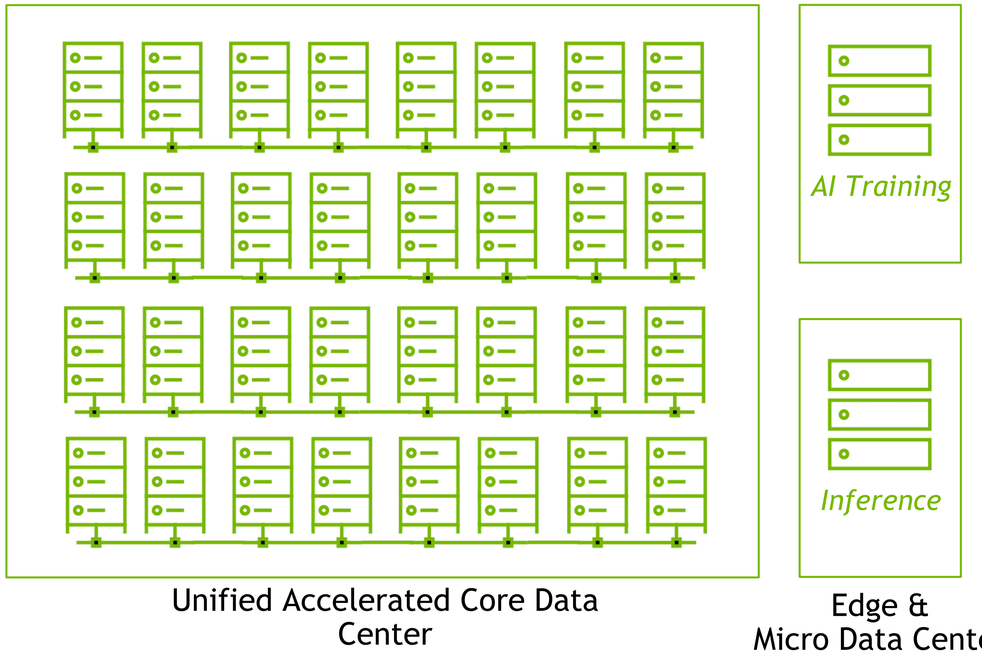
TODAY



CPU-only Data Center

Single-purpose Silos

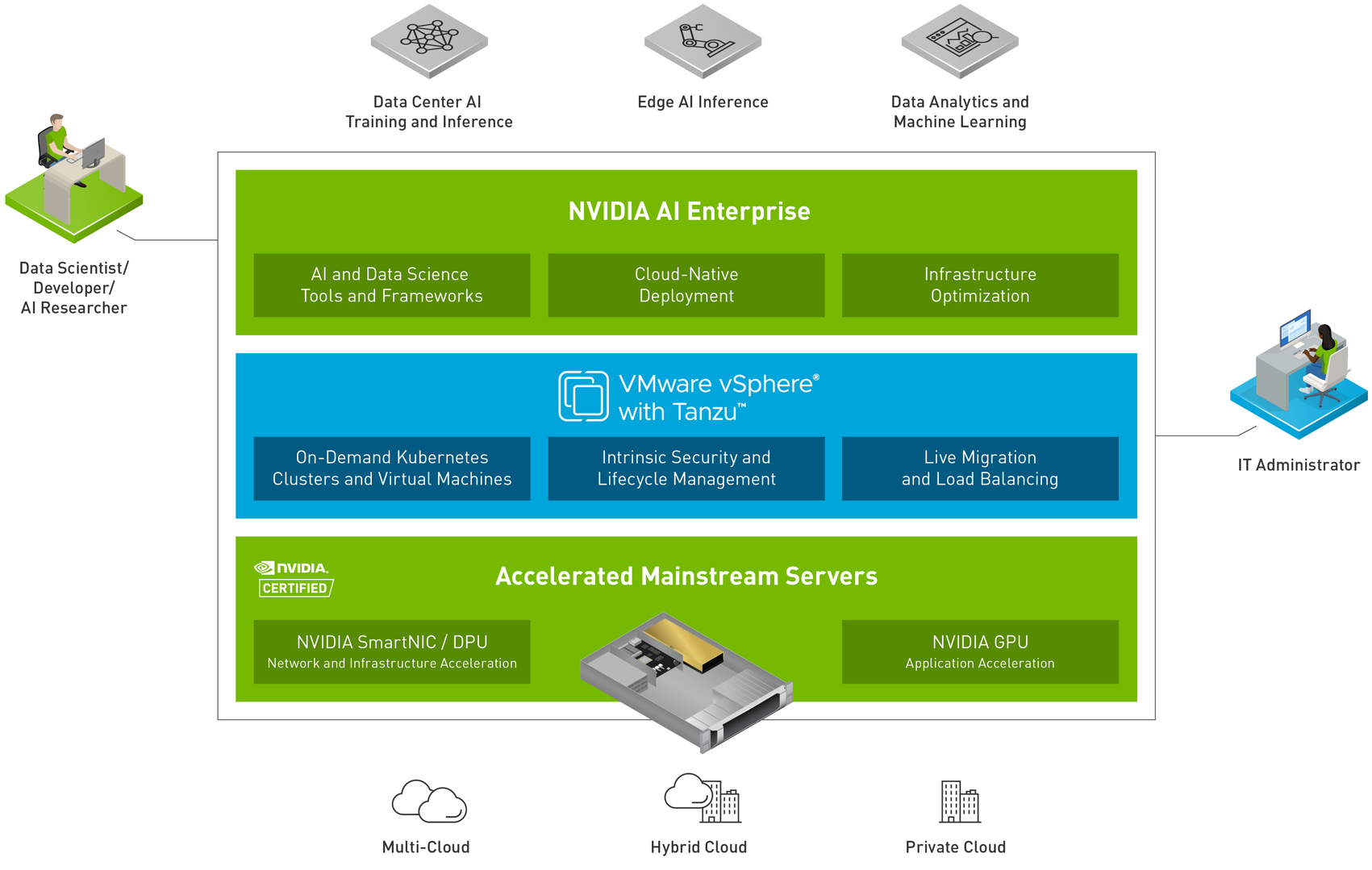
TOMORROW



Unified Accelerated Core Data Center

Edge & Micro Data Center

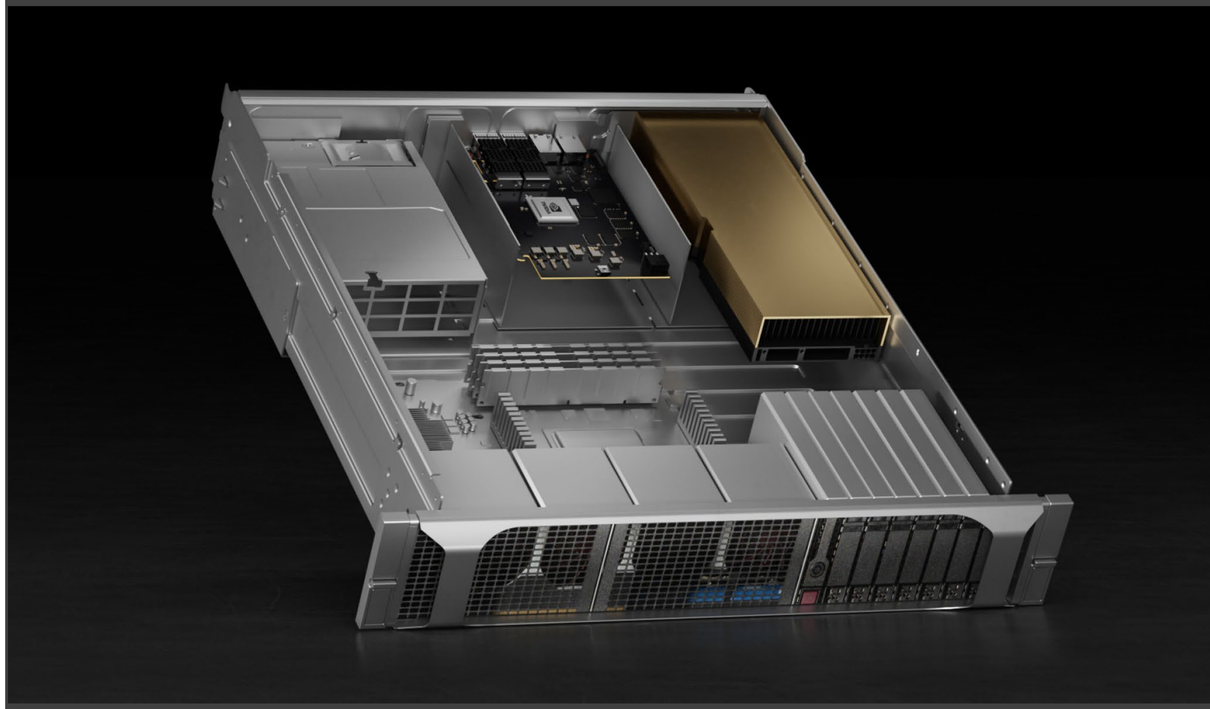
AI-READY ENTERPRISE PLATFORM



INTRODUCING NVIDIA-CERTIFIED FOR MAINSTREAM SERVERS

Common Accelerated Infrastructure for AI and Existing Enterprise Applications

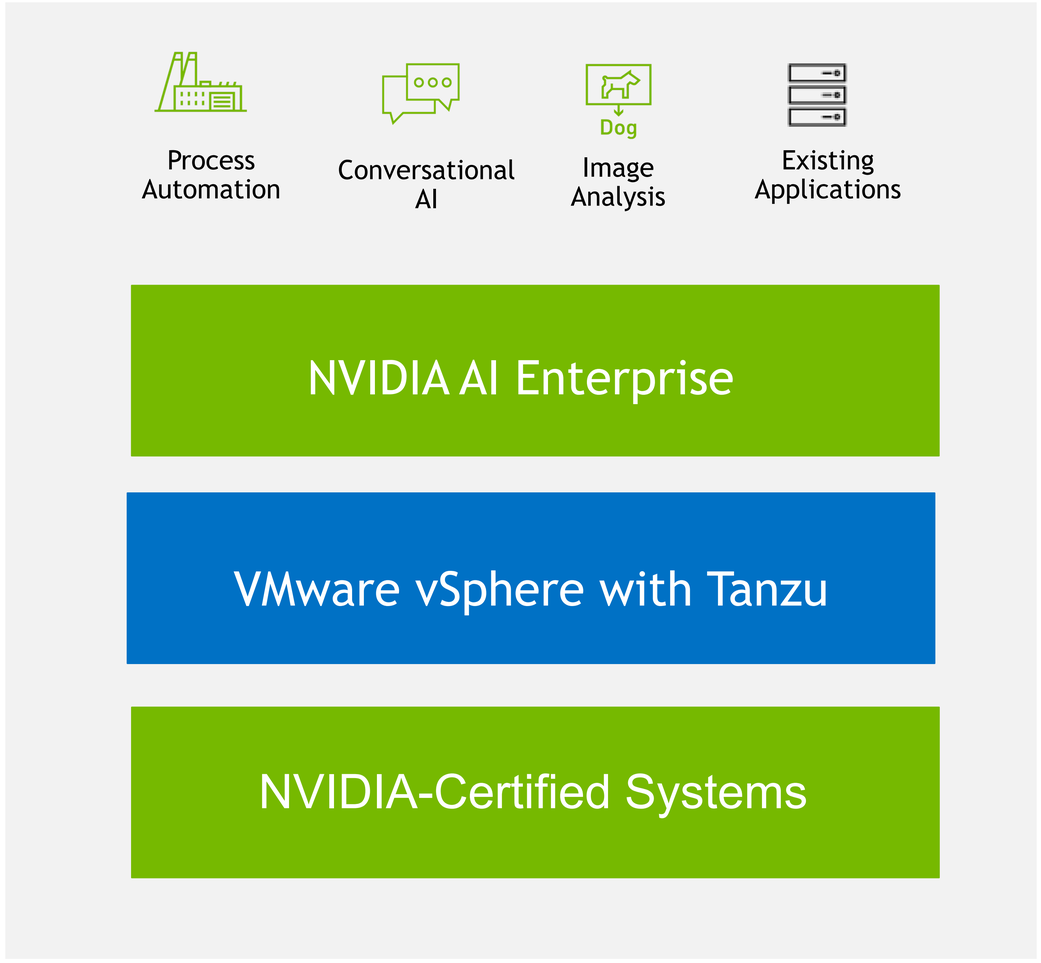
- Validated performance for both AI training and inference
- Optimized for NVIDIA AI Enterprise software, tools and frameworks
- Systems available from Dell, HPE, Lenovo, Supermicro, Cisco, Fujitsu and others
- Broad ecosystem including Equinix, Red Hat, VMware



NVIDIA CERTIFIED

TANZU KUBERNETES GRID FOR NVIDIA AI

Unified orchestration streamlines application management for containers and virtual machines



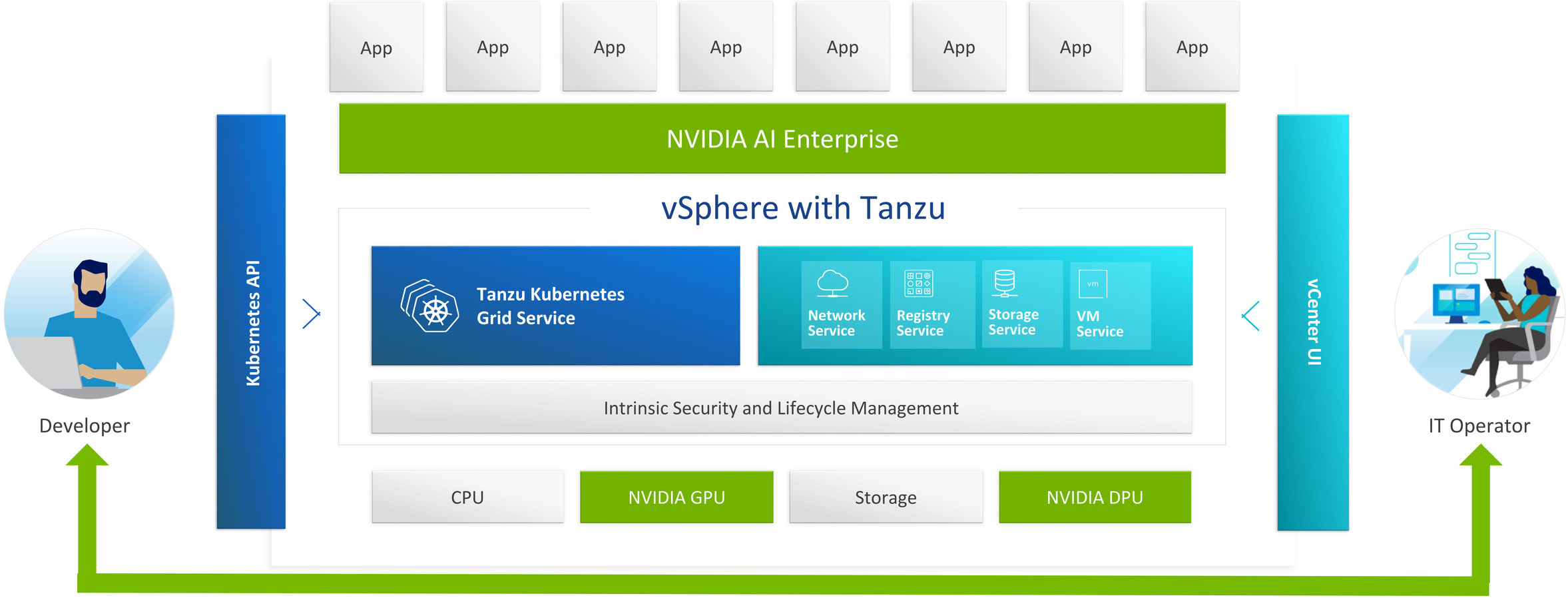
- By 2025, more than 80% of independent software vendors (ISVs) will offer their application software in container format[1]
- Use existing vSphere environment to rapidly deliver Kubernetes with NVIDIA AI and GPU acceleration to AI and data science teams
- Manage multiple accelerated Tanzu Kubernetes clusters alongside virtual machines through VMware vCenter Server

1. Gartner (July 2020): CTO's Guide to Containers and Kubernetes: 10 Frequently Asked Questions

NVIDIA

FULL KUBERNETES AUTOMATION

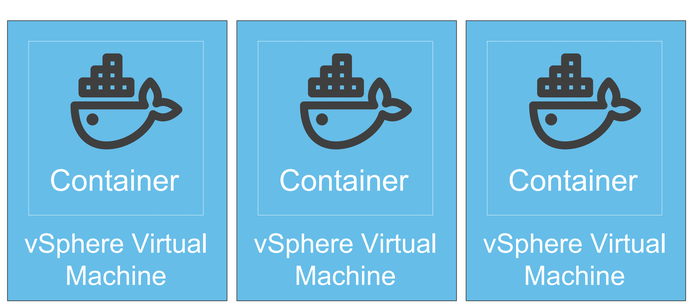
With NVIDIA AI Enterprise Software and NVIDIA GPUs and DPUs

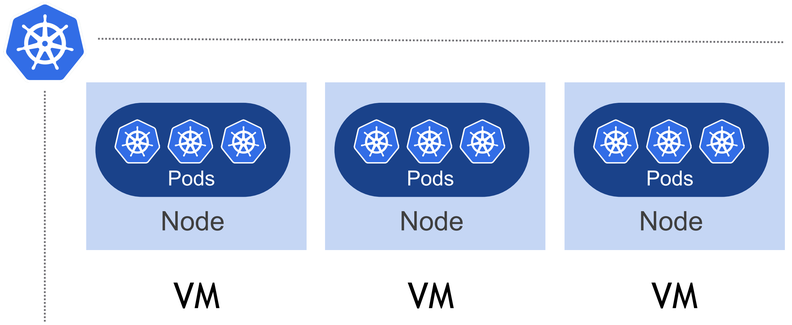


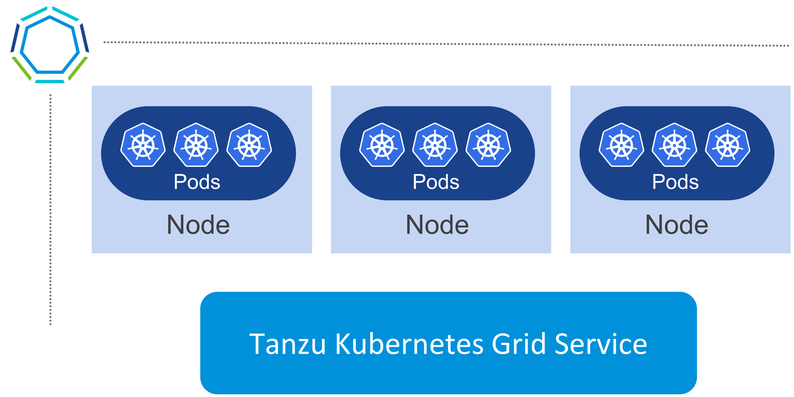
Building the bridge between these silos!

DELIVERING AI WORKLOADS WITH NVIDIA AI ENTERPRISE

Methods for Orchestration with VMware vSphere







vSphere + NVIDIA AI Enterprise

Run AI/ML Containers in VMs

vSphere + NVIDIA AI Enterprise

Run Kubernetes in VMs

vSphere + NVIDIA AI Enterprise

Run AI/ML in vSphere 7 w/ Tanzu Tanzu Kubernetes Grid Service

DATA CENTER READY ENTERPRISE MACHINE LEARNING OPERATIONS

Domino Data Labs

Domino Data Lab's validation for NVIDIA AI Enterprise pairs the Enterprise MLOps benefits of workload orchestration, self-serve infrastructure, and collaboration with cost-effective scale from virtualization on mainstream accelerated servers.

For Data Scientists & AI Researchers

Focus on research instead of dev ops. Launch Domino Workspaces on-demand, with docker images configured with the latest data science tools, frameworks, and NVIDIA GPUS – with automatic storing and versioning of code, data, and results.

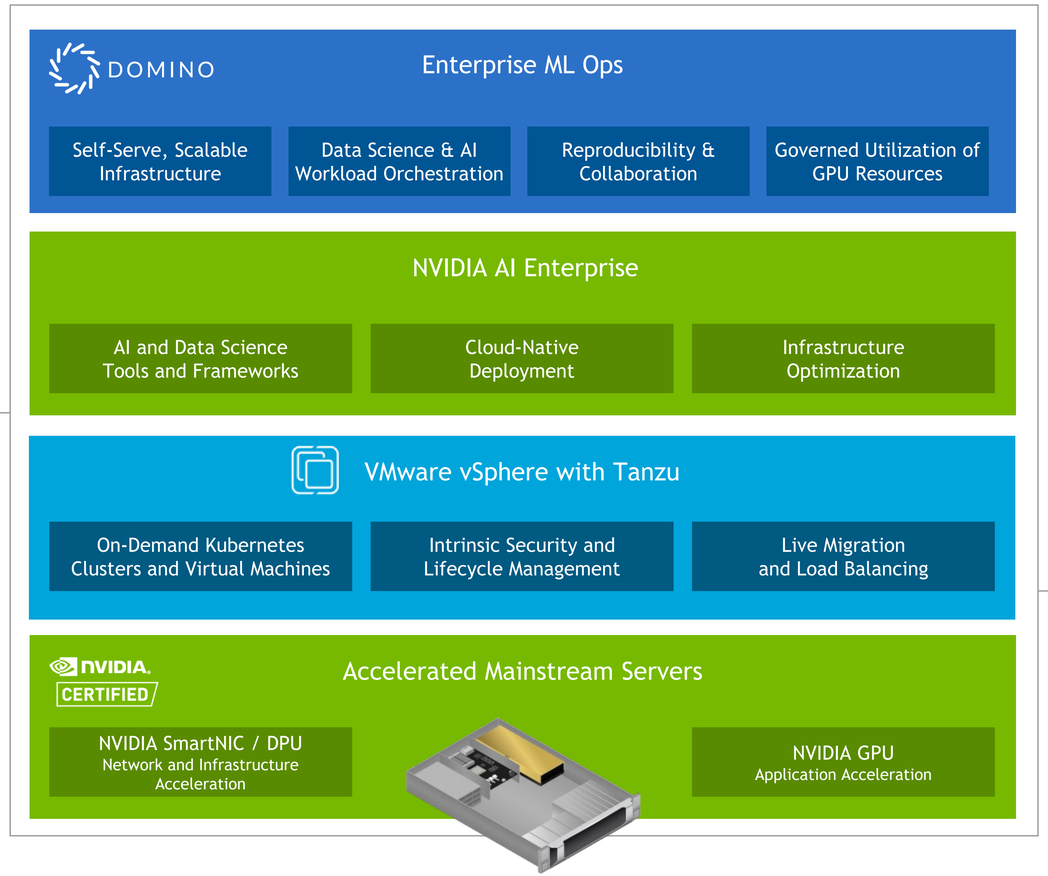
For IT

Get the confidence of enterprise-grade security, manageability, and support. Domino is validated to run on VMware vSphere with Tanzu – all deployed on industry-leading, NVIDIA-Certified™ systems.

Engage with Domino!

Email NVIDIA@dominodatalab.com

Data Scientist/
Developer/
AI Researcher

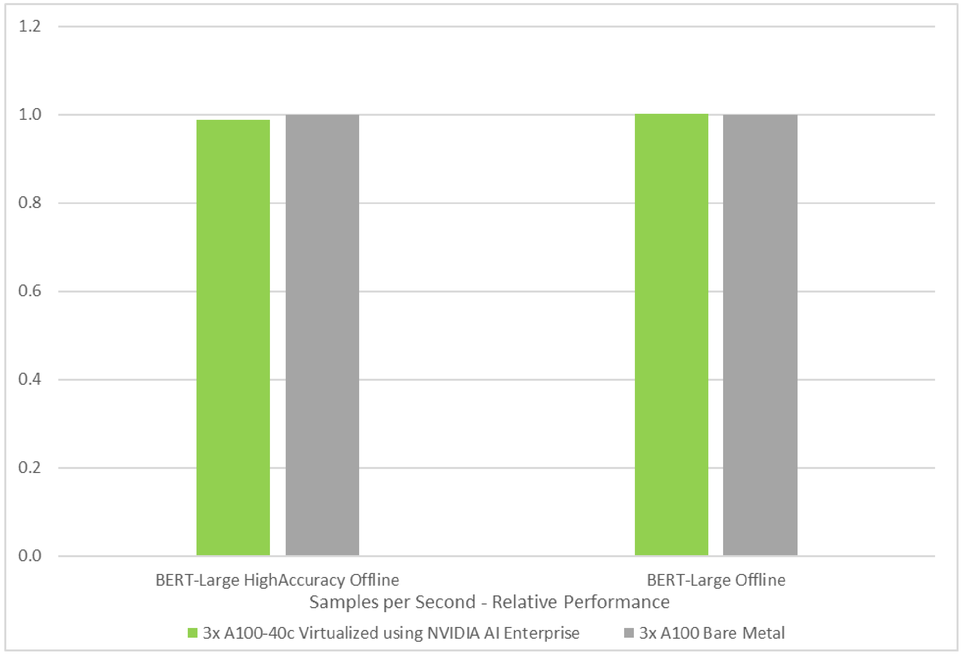


IT Administrator

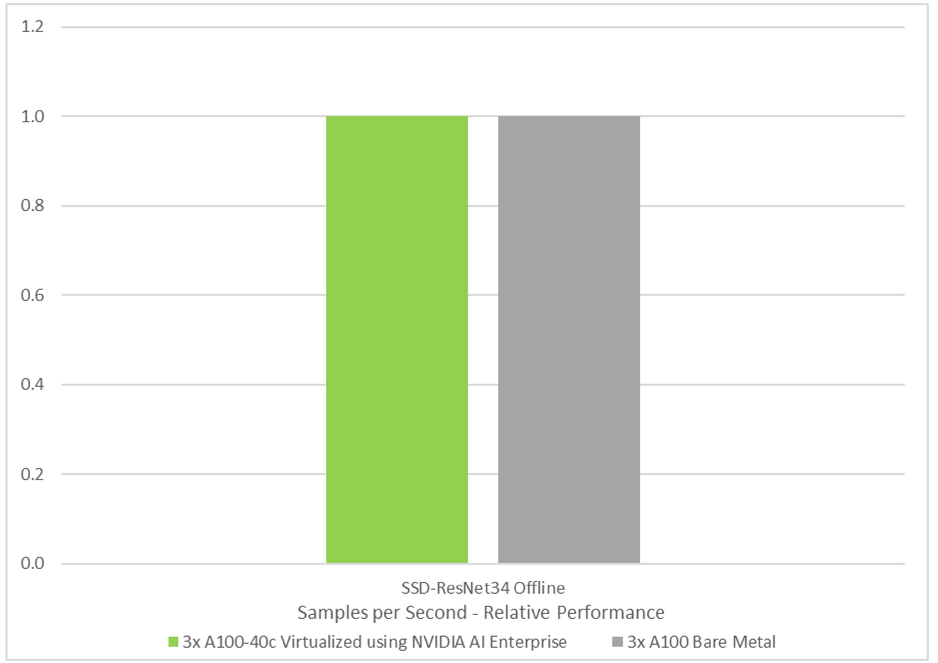
BARE METAL PERFORMANCE WITH NVIDIA AI ENTERPRISE ON VMWARE vSPHERE

MLPerf 1.1 Inference Benchmark: Datacenter – Virtualized

Natural Language Processing (BERT-Large) Inference on NVIDIA A100



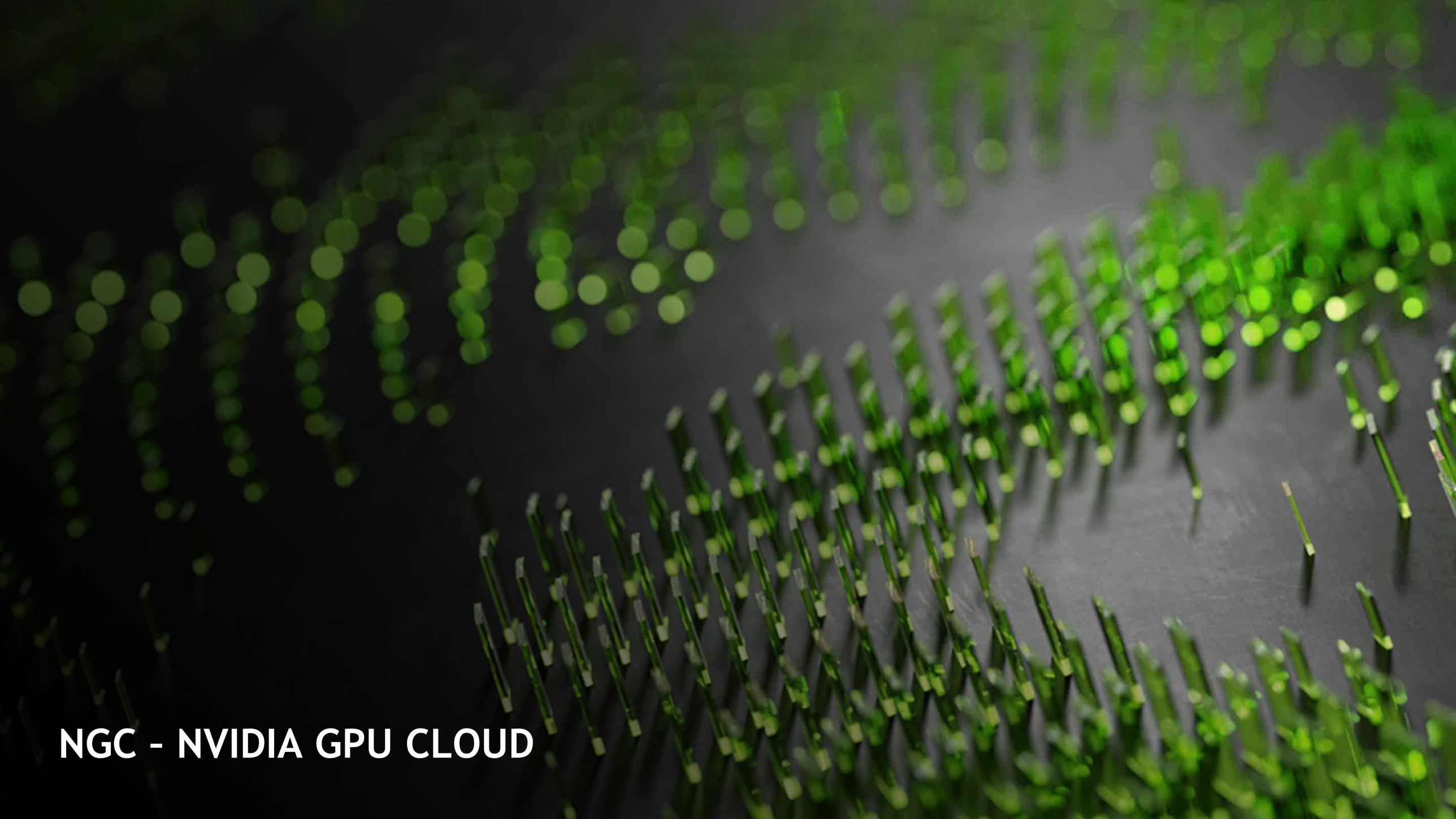
Object Detection (SSD-ResNet34) Inference on NVIDIA A100



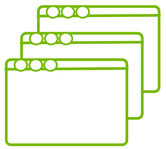
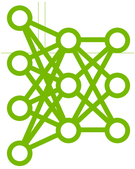
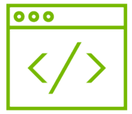
- Near Bare Metal performance
- Consistent performance across multiple MLPerf inference benchmarks
- NVIDIA AI Enterprise certified and optimized on VMware vSphere 7.0.2 with Dell EMC PowerEdge R7525 (NVIDIA-Certified)

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.
Server Config:: Dell EMC PowerEdge R7525, 3x A100-PCIE-40GB (Virtualized: 3x GRID A100-40C), TensorRT, AMD EPYC 7502, NVIDIA A100-PCIE-40GB, TensorRT 8.0.2, CUDA 11.3
MLPerf name and logo are trademarks. See www.mlcommons.org for more information.

NVIDIA


NGC – NVIDIA GPU CLOUD

NGC CATALOG – GPU-OPTIMIZED SOFTWARE

AI FRAMEWORKS	AI TOOLKITS	MODELS	MODEL SCRIPTS	COLLECTIONS
				
PyTorch, TensorFlow…	TAO, TensorRT, Triton…	BERT, Transformer…	Jupyter Notebooks	Computer Vision, Speech…

Computer Vision

- Traffic Analysis
- Gesture Recognition
- Medical Imaging

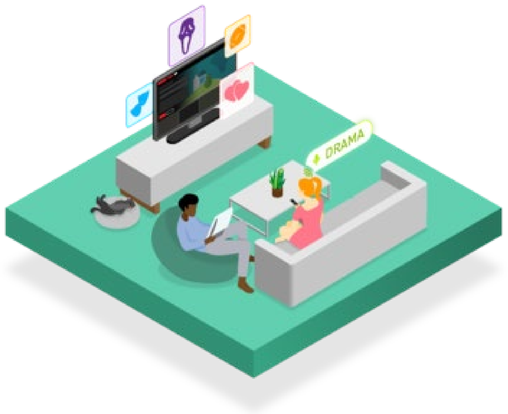
Conversational AI

- Chat bots
- Translation
- Music composition

Recommendation

- Page rankings
- Personalized shopping
- Music & movie suggestions

INDUSTRY SDKS FURTHER SIMPLIFY AI WORKFLOWS



Conversational AI

NVIDIA RIVA

Intelligent Video Analytics

NVIDIA DEEPSTREAM

Recommendation

NVIDIA MERLIN

Medical Imaging

NVIDIA CLARA

EVERYTHING YOU NEED TO BUILD AI IN ONE LOCATION

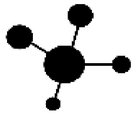
Search NGC for the app or use case

The Collection walks through all the required assets and how to use them together

Fine-tune and deploy

Deploy container

Start Jupyter Notebook instance



Download model

COLLECTIONS

Compatible assets grouped together, removes guesswork

Curated software by use cases

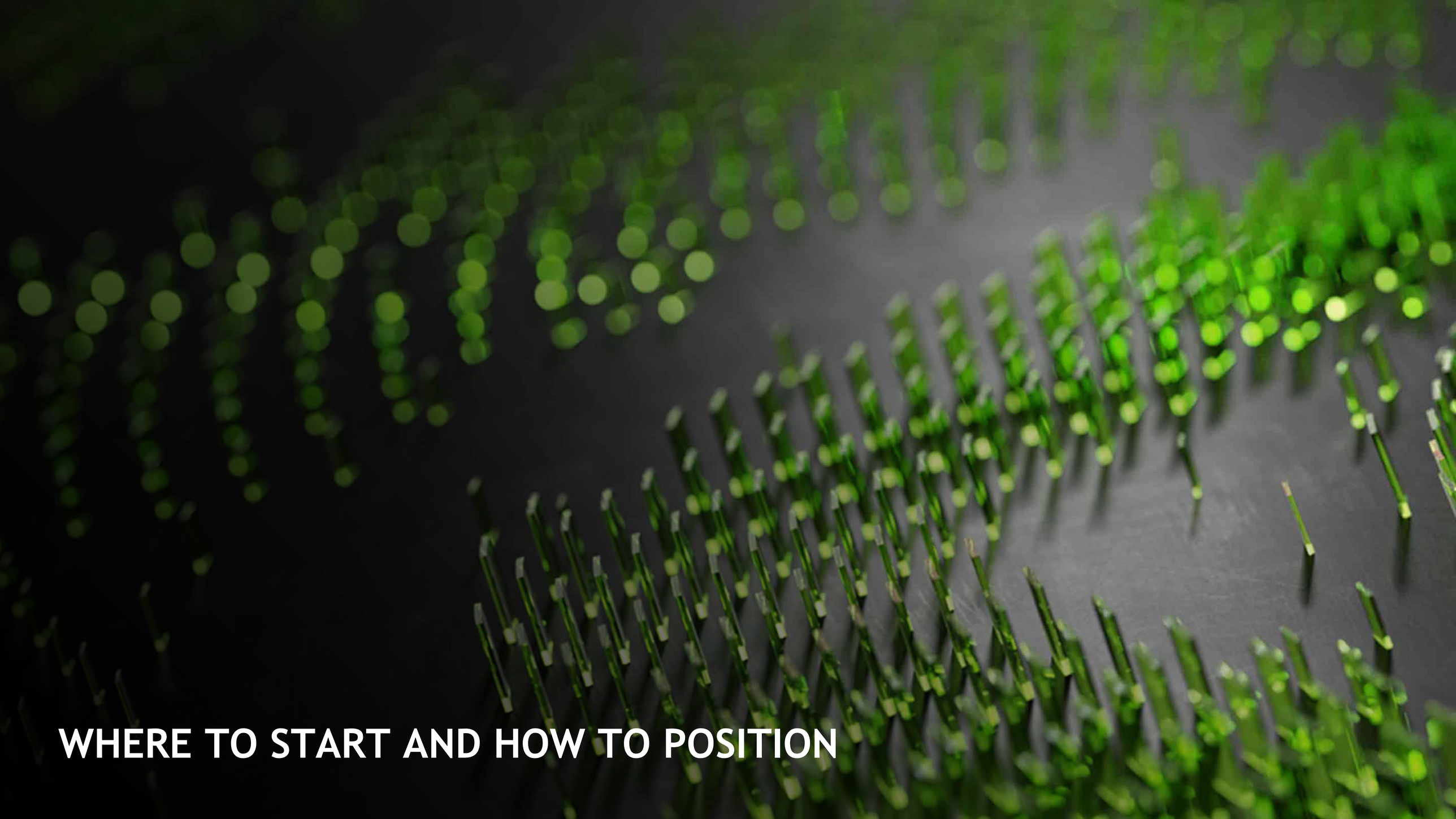
Detailed documentation further simplifies work for users

READY-TO-USE

Conversational AI

Computer Vision

NVIDIA AI App Frameworks



WHERE TO START AND HOW TO POSITION

NVIDIA AI ENTERPRISE - LICENSE & SUPPORT

NVIDIA AI Enterprise Offering	NVIDIA List Price	Business Critical Support (Optional add-on)
Subscription, 1-Year Term	$2,000 / CPU socket	$450 / CPU Socket / Year
Subscription, 3-Year Term	$6,000 / CPU socket	
Subscription, 5-Year Term	$8,000 / CPU socket	
Perpetual License + 1-Year Support	$4,494 / CPU socket	
Perpetual License + 3-Years Support	$6,292 / CPU socket	
Perpetual License + 5-Years Support	$8,090 / CPU socket	

Support Options (SUMS)	Services	Supported Channels	Service Level Agreement (SLA)
Business Standard (Included in price)	Issue resolution, bug fixes, software updates, maintenance	Phone, portal, email	9 x 5 Business Days (4hr initial response)
Business Critical (Optional upgrade/add-on)	Same as Business Standard Support	Same as Business Standard Support	24 x 7 (1hr initial response for S1 cases)

Note: "CPU Socket" means the number of physical processors in the computing environment on which NVIDIA AI Enterprise is installed, or number of virtual CPUs for the compute instance on which NVIDIA AI Enterprise runs. NVIDIA requires one license per CPU Socket.

Maximum 10 concurrent VMs per product license

CUSTOMER SCENARIOS

Virtual GPU 13.0 and later for **VMware vSphere** Customers

	Software Product
I need to deploy virtual PCs (VDI) and virtual workstations	NVIDIA vPC & NVIDIA RTX vWS
I need to deploy virtualized compute	NVIDIA AI Enterprise
I need to deploy virtual PCs and virtualized compute	NVIDIA vPC & NVIDIA AI Enterprise
I need to deploy virtual workstations and virtualized compute	NVIDIA RTX vWS & NVIDIA AI Enterprise

TARGET USE CASES

Support AI in the Core Enterprise Data Center

Complement to Bare Metal AI/ML

NVIDIA AI Enterprise benefits

Operational fit with existing server IT standards, processes, and tools

Flexible infrastructure - run AI alongside enterprise applications

Scale out for multi-node workloads on VMware vSphere

Alternative or Complement to Cloud

NVIDIA AI Enterprise benefits

Data proximity, security & governance

Like the cloud, build out as you scale with moderate incremental investment (< $20K/server)

Can create self-service solutions for efficient and easy AI/ML access

Inference at Scale

NVIDIA AI Enterprise benefits

Maximize GPU utilization with fractional/sharing of GPU

Provision right-sized VMs to data scientists in minutes

Develop in the environment you will deploy

NVIDIA AI ENTERPRISE

Enabling AI and Data Analytics on VMware vSphere

Optimized for Performance

Achieve near bare-metal performance across multiple nodes to power large, complex training and machine learning workloads.

Certified for VMware vSphere

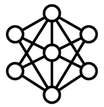
Reduce deployment risks with a complete suite of NVIDIA AI software certified for the VMware data center

NVIDIA Enterprise Support

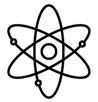
Ensure mission-critical AI projects stay on track with access to NVIDIA experts.

RECOMMENDED ACCELERATORS FOR NVIDIA AI ENTERPRISE

GPU COMPUTE ACCELERATION



DL Training



Scientific Research

Data Analytics

Highest Perf Compute
AI, HPC, Data Processing

Fastest Compute, FP64
Up to 7 MIG instances

A100
300W | 80G
2-slot FHFL | NVLINK

Language Processing

Conversational AI

Recommender Systems

AI Inference &
Mainstream Compute

Versatile Mainstream Compute
FP64, Up to 4 MIG instances

A30
165W | 24GB
2-slot FHFL | NVLINK

NETWORK ACCELERATION

Key Management &
Secure Boot

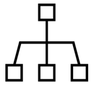
TLS & IPsec Acceleration

High-Performance
SmartNIC for Data Centers

Unmatched Scale and
Functionality

ConnectX-6 Dx
Dual-Port 25/50/100GbE/
Single-Port 200GbE

IDS/IPS



Software-Defined
Networking

Fully programmable DPU for
Next-Gen Infrastructure

Offloads, Isolates, and
Accelerates

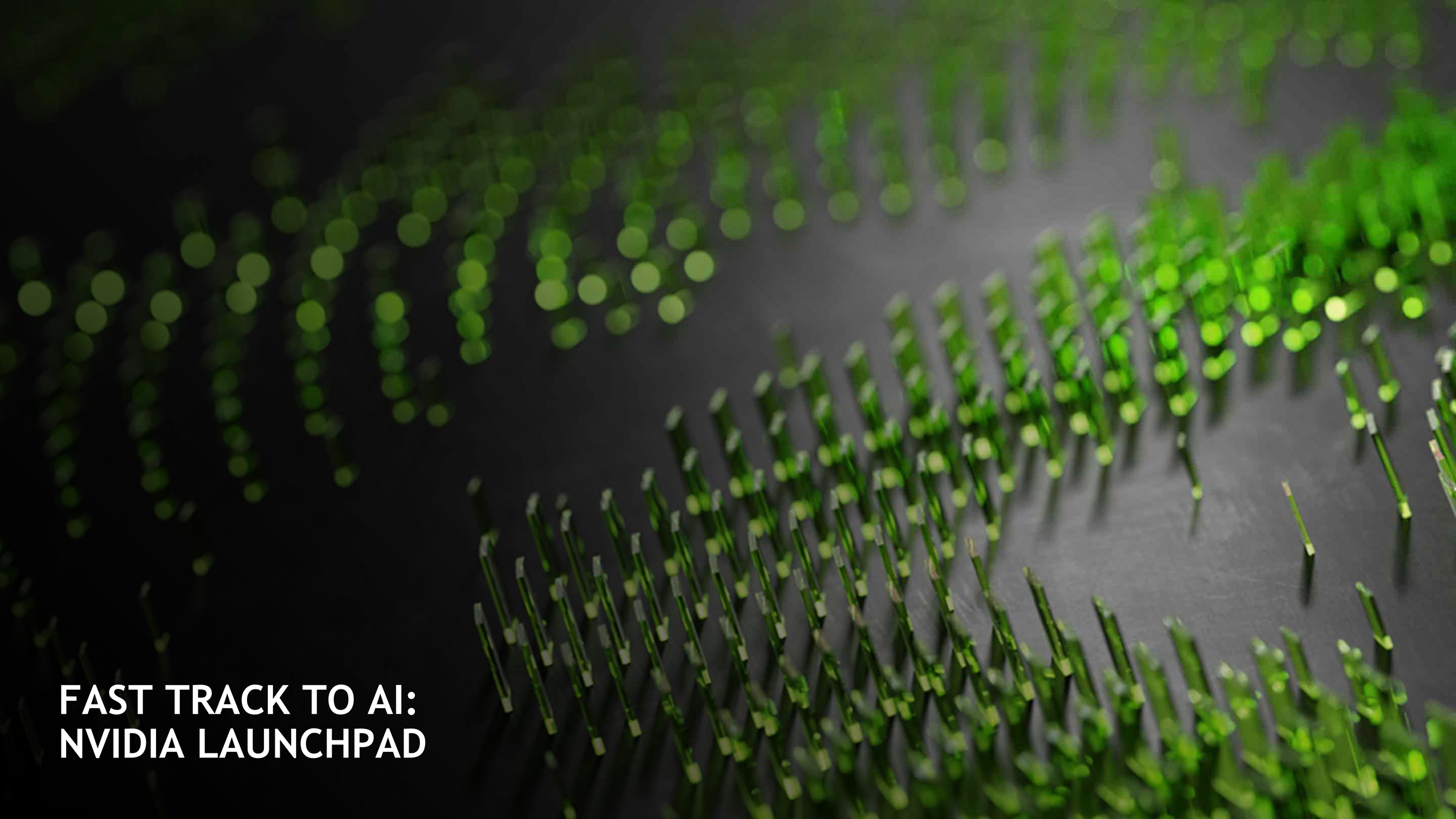
BlueField-2 DPU
Combines 8x A72 Arm Cores and
ConnectX-6 Dx SmartNIC

AI Training *AI Inference*

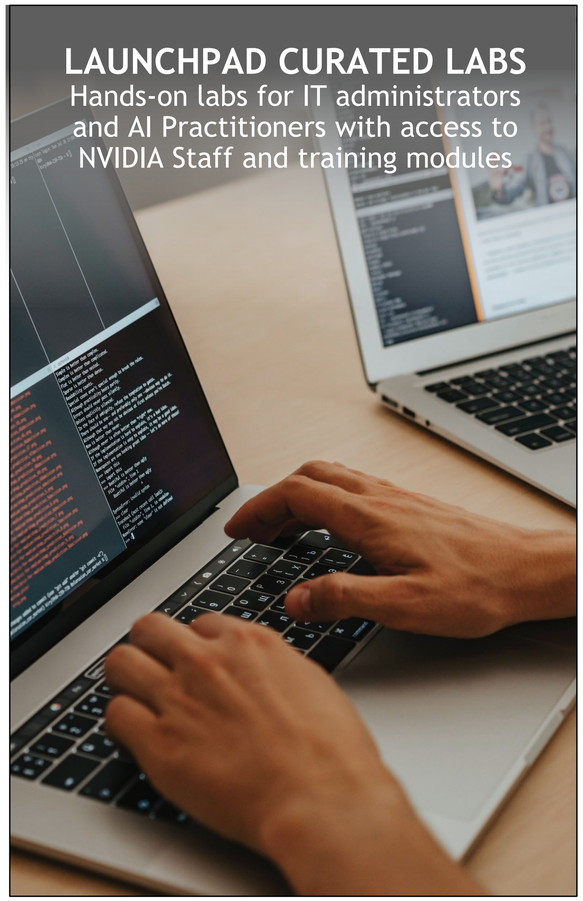
Hardware Accelerated, Software-Defined

NVIDIA A40, A2 GPUs are also supported for compute acceleration and graphics workloads.

NVIDIA


FAST TRACK TO AI:
NVIDIA LAUNCHPAD

NVIDIA AI ENTERPRISE ON LAUNCHPAD

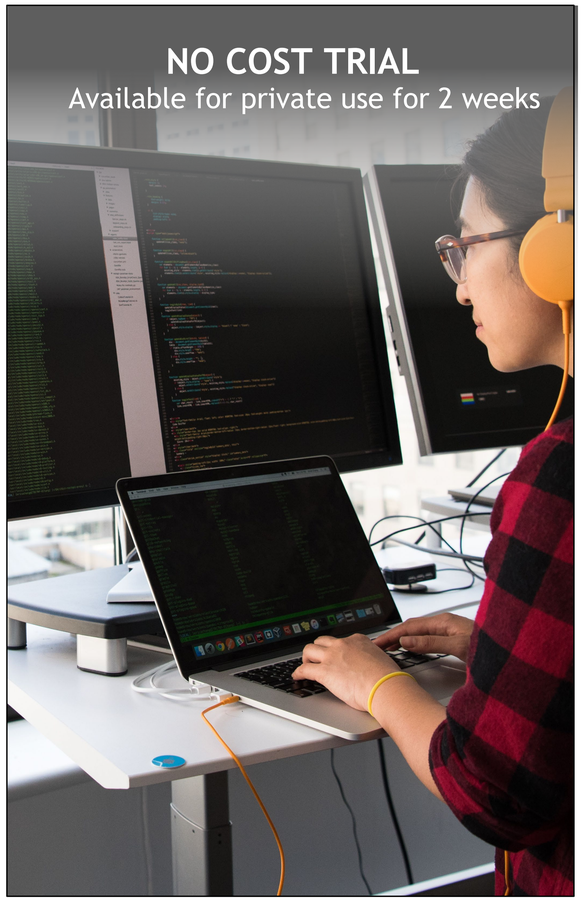


LAUNCHPAD CURATED LABS

Hands-on labs for IT administrators and AI Practitioners with access to NVIDIA Staff and training modules

NVIDIA CERTIFIED

Fully provisioned NVIDIA-Certified systems



NO COST TRIAL

Available for private use for 2 weeks

NVIDIA AI ENTERPRISE

Access to NVIDIA AI Enterprise running on VMware vSphere with Tanzu

NOW AVAILABLE IN 9 REGIONS WORLDWIDE

EQUINIX

NVIDIA

FOR RAPID TESTING AND PROTOTYPING

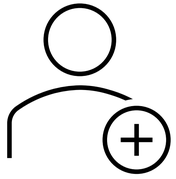
2-4 week program to accelerate the AI journey

Rapid access to AI infrastructure available in 9 regions for free to qualified customers

Fully provisioned with what you need to get started

Self-paced learning and hands-on labs for both IT admins and AI practitioners / data scientists

Terms of Use:
For limited time testing only—not for production use



IT ADMIN - BENEFITS

- Detailed labs for deployment
- Gain deep understanding of AI components best practices and how to support AI teams
- Access to resources to facilitate conversations with AI practitioners

AI PRACTIONER- BENEFITS

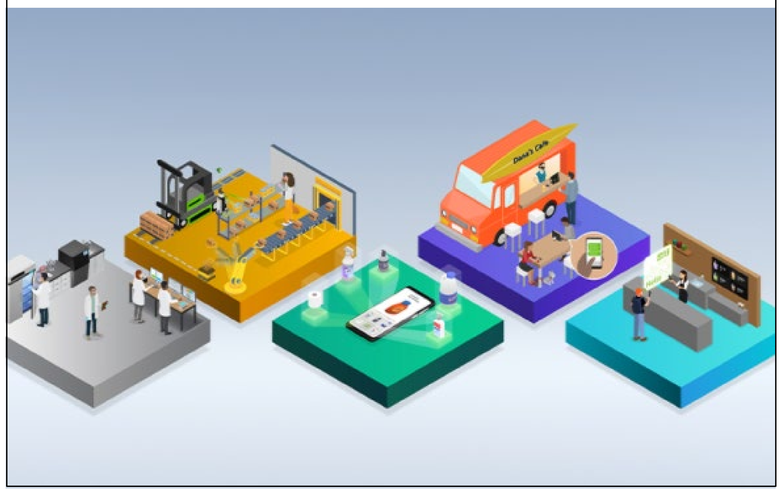
- Detailed labs for developing AI projects
- Personalized learning Journeys
- DLI credits
- Expedite ETL, AI training workflows
- Deploy models quickly to Triton

GETTING STARTED WITH NVIDIA AI

NVIDIA AI Enterprise Trial Programs

Test Drive Demo

- Self-directed, remote access demo
  - Predicting NYC Taxi Fares with RAPIDS
  - BERT Question Answer in TensorFlow
- Requires ~1 hr / Access for 48 hrs



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, vSphere 7 u2 (or later)
- Free evaluation licenses for on premises POC
- 90 days to test and experience

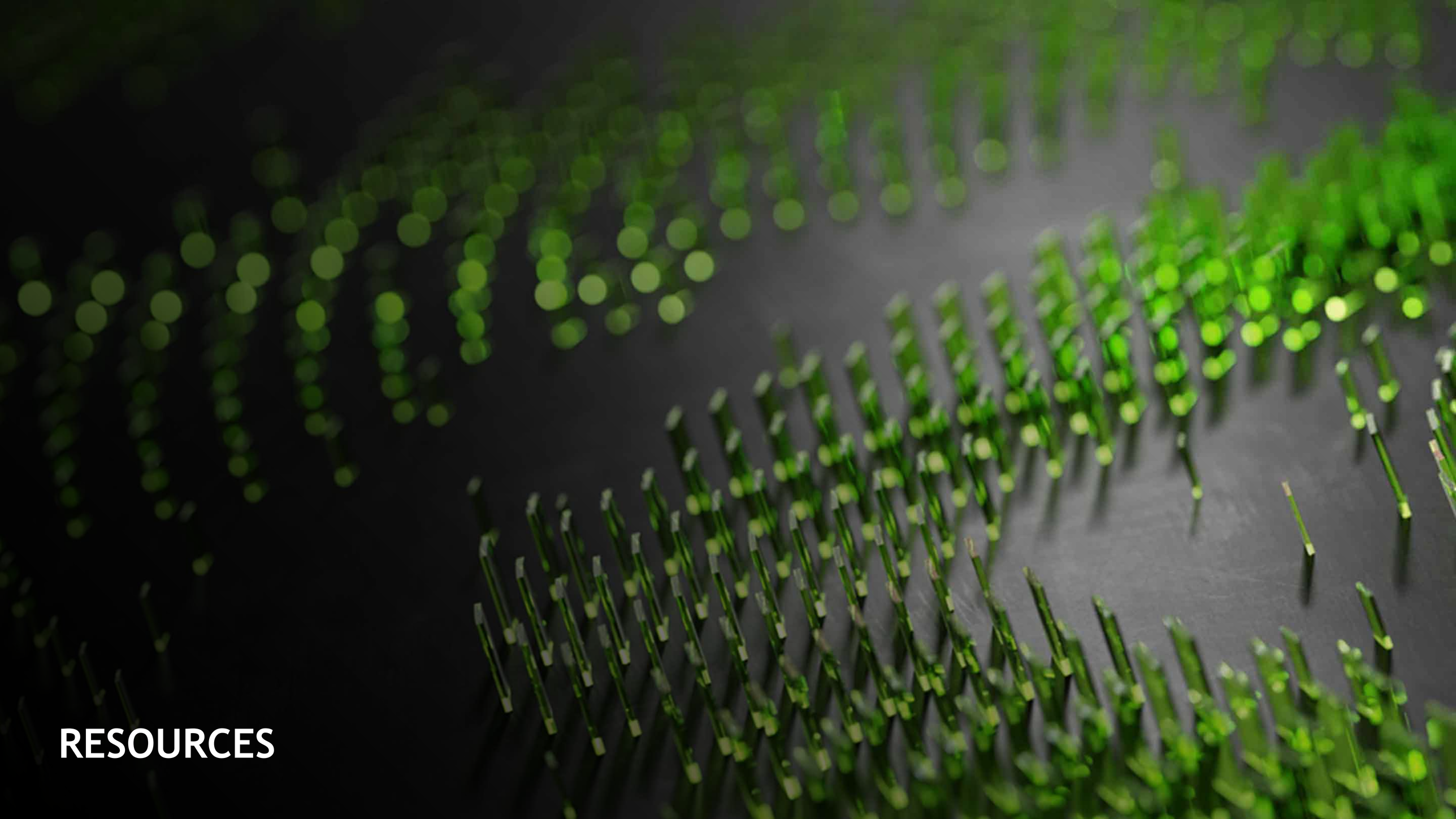
FAQ´S

NVIDIA AI ENTERPRISE FAQ

1. Which hypervisors are supported with NVAIE?
2. Can I use NVIDIA V100 with NVAIE?
3. What do customers who are using vCompute (vCS) within VMware? Do they have to buy NVAIE?
4. For a VDI-by-day, Compute-by-night concept, do I have to buy NVAIE?
5. Does NVAIE support graphics workloads (vPC, vWS, vApps)?
6. Is VMware reselling NVAIE?
7. Is VMware Tanzu a pre-requisite for NVAIE?

NVIDIA AI ENTERPRISE FAQ

8. Does NVAIE include vSphere licences?
9. Does NVAIE works on DGX?
10. What happens to vCS customers on vSphere?
11. Is vCS end of life?



RESOURCES

NEXT STEPS IN THE AI-READY JOURNEY

Customer Resources

Webpages:

- NVIDIA AI Enterprise Product Page
- NVIDIA AI Enterprise Resources

Papers:

- IDC Whitepaper: Scaling Artificial Intelligence and Machine Learning Workloads
- Whitepaper: AI-Ready Enterprise Platform with NVIDIA AI Enterprise and VMware vSphere 7
- Solution Brief: AI-Ready Enterprise Platform

Videos:

- NVIDIA AI Enterprise on VMware vSphere
- Natural Language Processing with NVIDIA AI Enterprise

Partnership:

- Video: NVIDIA & VMware Partnership
- VMware's NVIDIA Partnership Page
- NVIDIA's VMware Partnership Page

NVIDIA AI ENTERPRISE – CUSTOMER TESTIMONIALS

UNIVERSITÀ DI PISA

"NVIDIA AI Enterprise allowed us to expand our support for our researchers and students who utilize data analytics and AI deep learning and machine learning, while making these applications easier to deploy and manage. Our testing has shown that these latest collaborations between NVIDIA and VMware deliver the full potential of our GPU-accelerated virtualized infrastructure at near bare-metal speed."

Maurizio Davini, Chief Technology Officer

Mass General Brigham

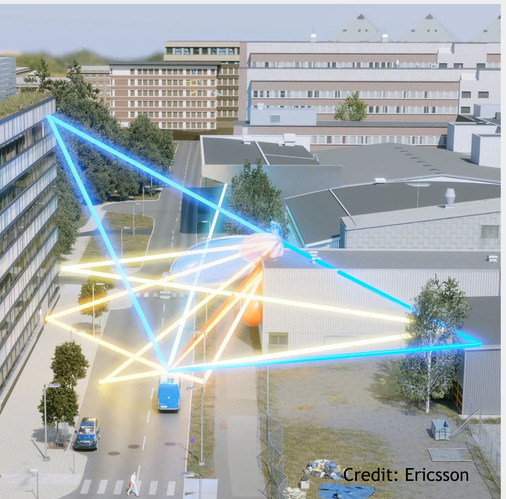
“Virtualization is enabling healthcare systems to deliver services to clinicians and patients at scale, across radiology departments and facilities. It has the potential to significantly increase the adoption of GPU-based AI applications. This allows for better utilization of technology infrastructure and minimizes the need for dedicated GPU systems for each project, which means AI can be applied more broadly to improve patient services.”

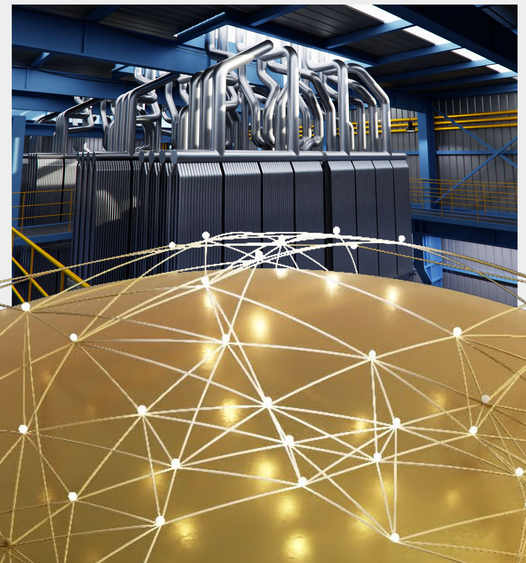
Tom Schultz, Director of Information Systems, Enterprise Medical Imaging & Clinical Data Science

NVIDIA GTC

ONLINE CONFERENCE

March 21-24, 2022





NVIDIA GTC

WHAT TO EXPECT AT GTC 2022

500+ SESSIONS

Live sessions, on-demand presentations, interactive panels, beginners content, and more.

AMAZING SPEAKERS

Jensen Huang (GTC Keynote) and thought leaders from every industry.

CONNECT WITH THE EXPERTS

Opportunities to connect with subject-matter experts from NVIDIA to get your pressing questions answered.

TRAINING

Workshops from the NVIDIA Deep Learning Institute (DLI) and NVIDIA Academy with courses for AI, accelerated computing, data science, and more.

STARTUPS

NVIDIA's startup ecosystem and industry execs sharing what it takes to succeed as a startup working in AI, data science, or HPC.

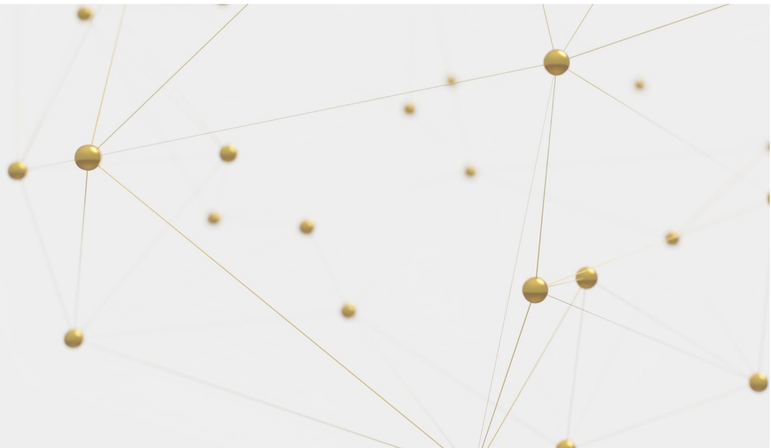
DEMOS

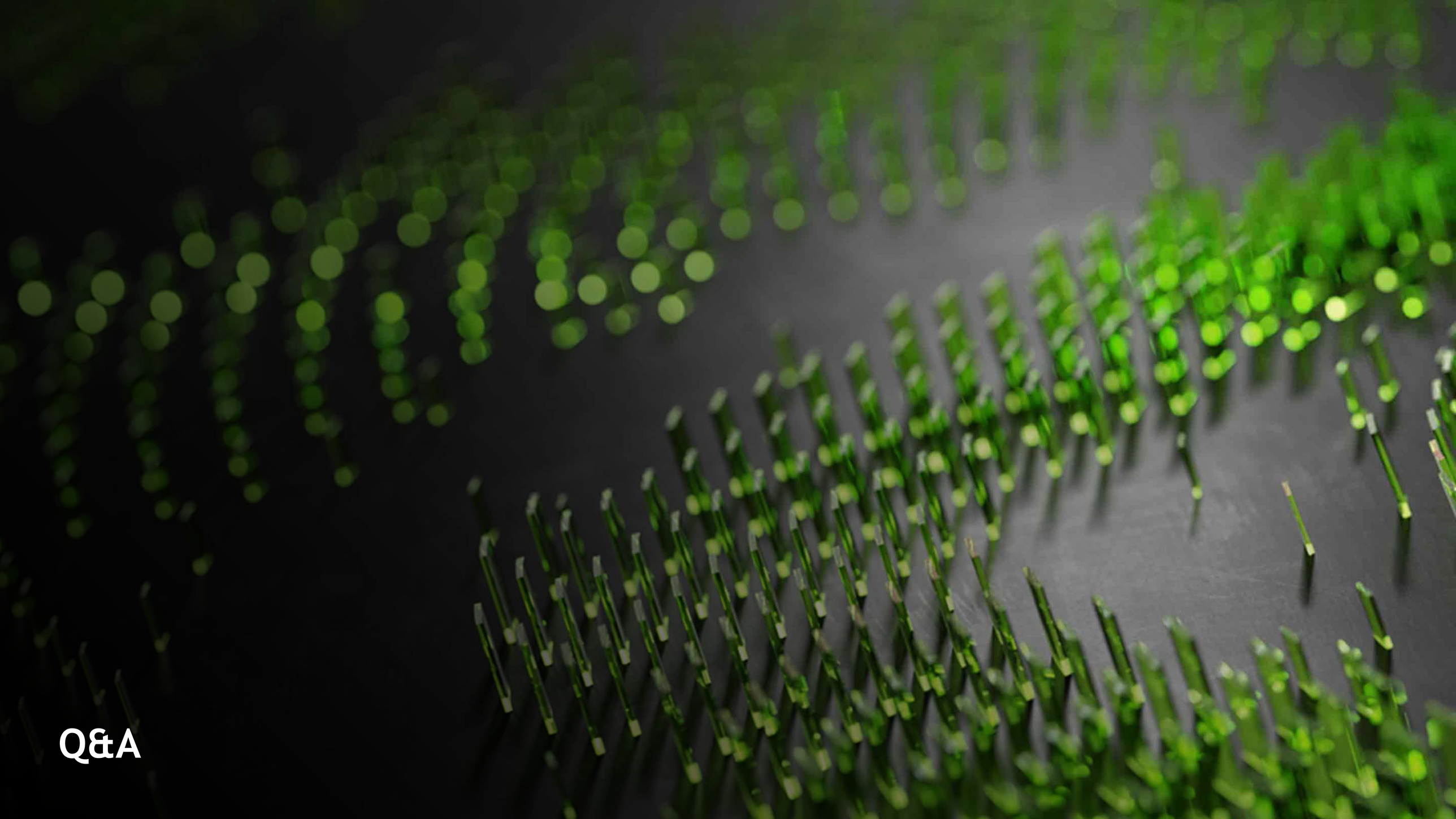
Solutions for workloads by subject-matter experts on the latest NVIDIA innovations and partner applications.

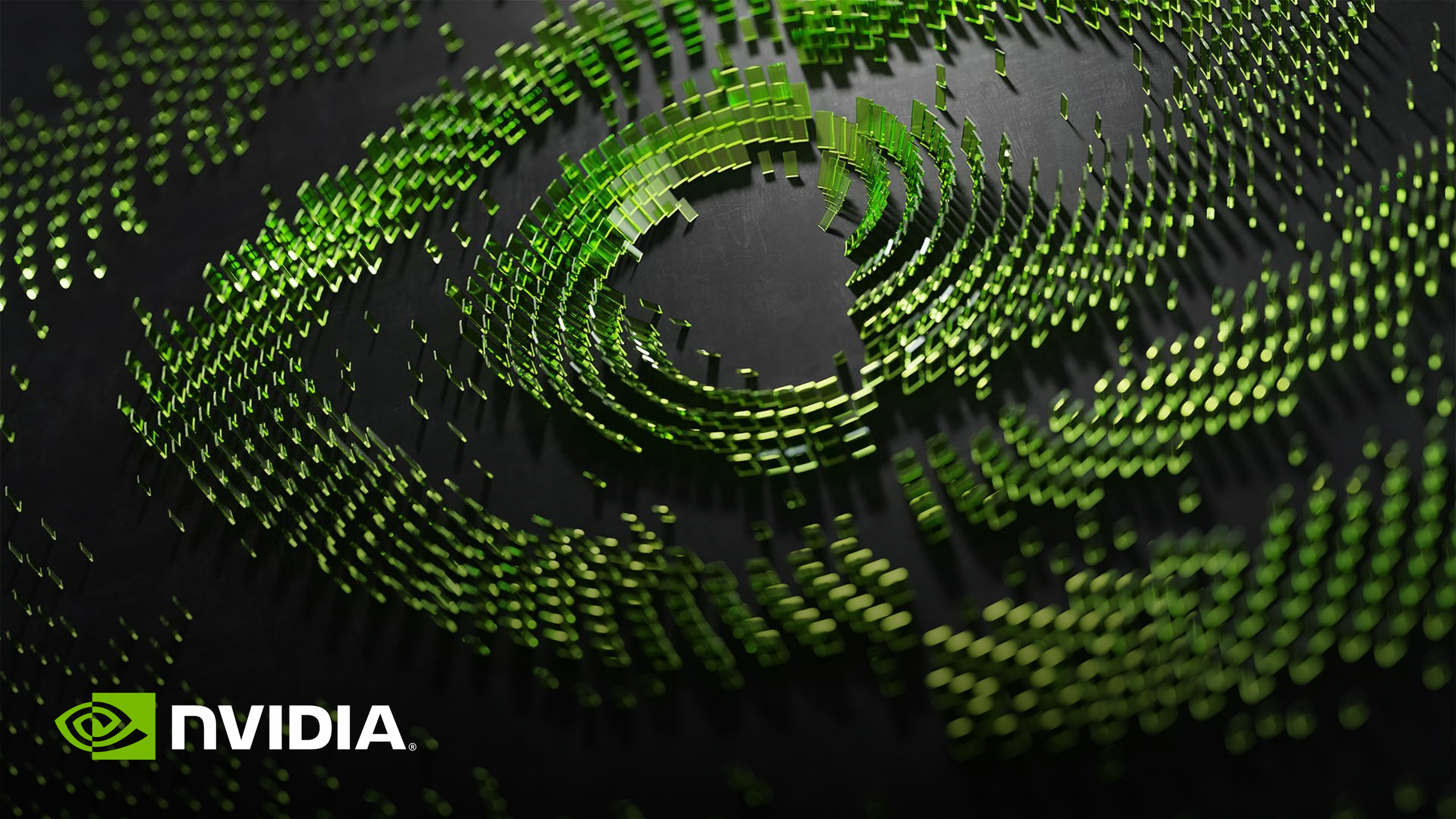
Conference and Training: March 21-24, 2022

Keynote: March 22, 2022

GTC sessions will run live in local time zones across NALA, EMEA, and APAC




Q&A


NVIDIA