Accelerate your AI inference initiatives

Create a pathway to success for your enterprise with HPE, NVIDIA, and VMware

Hewlett Packard Enterprise | NVIDIA | vmware

AI — opportunity and challenge of the decade

Artificial intelligence (AI) is transforming every industry, from improving customer relationships in financial services, streamlining manufacturer supply chains, to helping doctors deliver better outcomes for patients. While most organizations know that it is critical to invest in AI to secure their future, they struggle with finding the right strategy and platform.

While growth in AI development and training remains robust, industry projections show that the AI inference market will grow rapidly to tens of billions of dollars this decade.

AI inferencing deploys trained AI models — at the point where data is created and can be acted upon quickly to generate business value.

Deploy AI at any scale

AI inference workloads span an array of use cases across many industry verticals such as healthcare, financial services, and manufacturing. These workloads are often compute- and data-intensive, so they require accelerator-optimized compute that is efficient, secure, and scalable.

Enterprise-grade AI platform

With more than 100 frameworks, pretrained models, and development tools, NVIDIA® AI Enterprise is designed to accelerate your enterprise to the leading edge of AI while also simplifying AI to make it accessible to every enterprise. When combined with HPE ProLiant that are NVIDIA-Certified, NVIDIA AI Enterprise ensures you get the right level of performance, scalability, and enterprise-grade support.

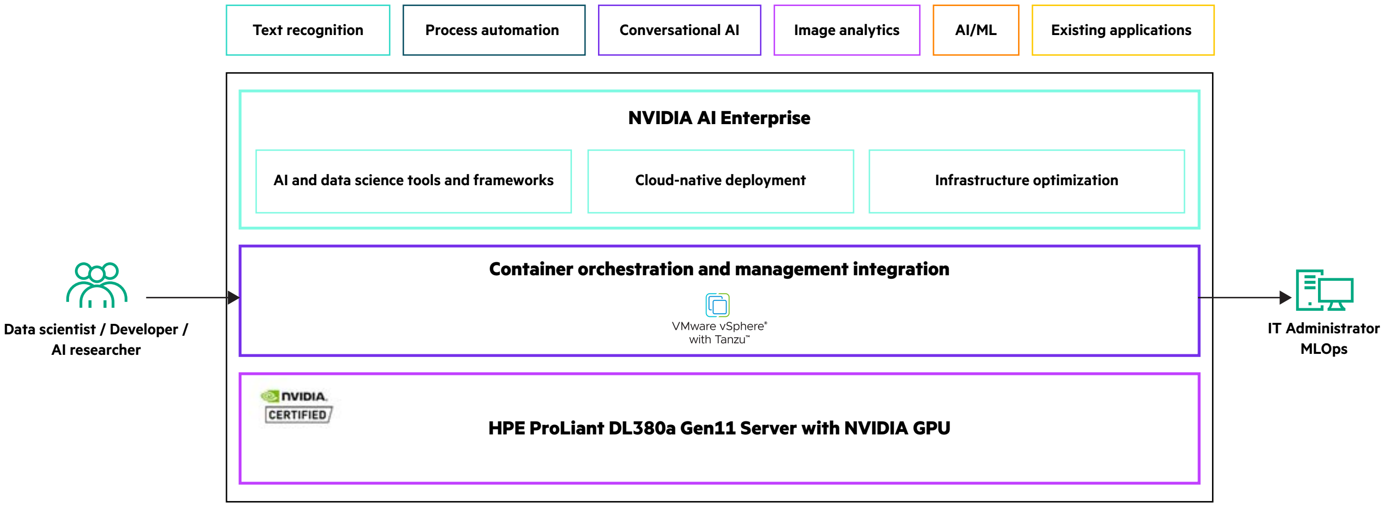

Text recognition
Process automation
Conversational AI
Image analytics
AI/ML
Existing applications
NVIDIA AI Enterprise
AI and data science tools and frameworks
Cloud-native deployment
Infrastructure optimization
Container orchestration and management integration
VMware vSphere® with Tanzu™
NVIDIA CERTIFIED
HPE ProLiant DL380a Gen11 Server with NVIDIA GPU
Data scientist / Developer / AI researcher
IT Administrator MLOps

Figure 1. AI-ready enterprise platform

Key enabling technologies

The solution consists of 4 key building blocks.

NVIDIA AI Enterprise Suite

This end-to-end, cloud native suite of AI and data science applications and frameworks are optimized and exclusively certified by NVIDIA to run on VMware vSphere® with NVIDIA-Certified Systems. It includes key enabling technologies and software from NVIDIA for rapid deployment, management, and scaling of AI workloads in the modern hybrid cloud. NVIDIA AI Enterprise is licensed and supported by NVIDIA.

VMware vSphere with Tanzu

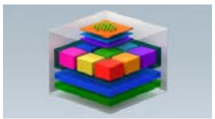
VMware vSphere®
with Tanzu™

This industry-leading virtualization platform runs more than 70 million workloads across hundreds of thousands of organizations worldwide. It transforms bare-metal servers (including CPU- and GPU-based resources) into centrally managed AI and ML infrastructure pools.

- Optimize performance
- Increase availability
- Tighten security
- Streamline maintenance
- Reduce costs
- Create an agile, efficient, resilient, and intrinsically secure infrastructure platform that supports existing workloads and next-gen applications such as AI

VMware vSphere has been enhanced to optimally run the NVIDIA AI Enterprise software suite leveraging NVIDIA's latest NVIDIA Ampere architecture-based GPUs.

NVIDIA GPUs: accelerated computing technologies



NVIDIA accelerated computing technologies tackle computational challenges far beyond the capabilities of ordinary computers. Accelerated computing requires more than just powerful GPUs. The combination of NVIDIA® CUDA® general purpose programmable GPUs and numerous GPU-accelerated SDKs, APIs, and algorithms provides full-stack computing solutions to deliver incredible application speed-ups across multiple domains. Distributed GPU computing systems and software scale processing across an entire data center. Cloud data centers worldwide are increasingly scaling up and scaling out with NVIDIA GPU-accelerated systems and architectures, running a diverse set of AI, HPC, and data analytics applications.

NVIDIA now has full stack solutions for different industries, fields of science, and applications. Over 450 NVIDIA SDKs, toolkits, libraries, and models serve industries and applications from gaming and design, to life and earth sciences, robotics, self-driving cars, quantum computing, supply-chain, logistics, cybersecurity, 5G, climate science, digital biology, and more. Over 25,000 companies use NVIDIA AI technologies today.

HPE ProLiant DL380a Gen11 Server

The HPE ProLiant DL380a was designed from the ground up as an ultra-scalable platform for accelerator-optimized AI workloads. Combining the density of a 2U dual socket rack server with expandability of up to four double-wide NVIDIA GPUs making it ideal for modern AI inference workloads.

The innovative HPE ProLiant DL380a Gen11 Server delivers advanced engineered solutions to resolve today's hybrid cloud infrastructure challenges. They combine the best of on-premises and cloud computing with:

- An intuitive cloud operating experience
- HPE trusted security by design
- Optimized performance for large, complex AI workloads

To learn more about HPE ProLiant Gen11 Servers visit hpe.com/proliant

HPE GreenLake cloud services

In addition, HPE GreenLake platform delivers a trusted, enterprise-grade cloud experience to accelerate data modernization initiatives, on-premises, in your data center, or at a colocation. Take control of your data while gaining insights and optimizing operations by scaling as needed with continuous monitoring and a flexible architecture that enables right-sized capacity bursting on demand.

To learn more about HPE GreenLake, visit hpe.com/greenlake

HPE ProLiant and NVIDIA AI Enterprise solutions

Generative visual AI

- Generate lifelike and dynamic 3D animations including character movements, physics simulations, and environmental effects.
- Curate new visual content, such as images or videos, that mimic real-world data.

Natural language processing AI

Develop and deploy end-to-end AI to power natural language processing for:

- Speech AI
- Fraud detection
- Predictive maintenance and more

HPE ProLiant DL380a Gen11

Up to four double-wide NVIDIA GPUs

NVIDIA AI Enterprise

Optimized for visual apps

Powering large language models

Optimized AI workloads on HPE ProLiant Gen11

HPE ProLiant delivers accelerator-optimized solutions that help you respond quickly to business needs, and scale with growth.

- Ultra-scalable architecture with up to 33% more GPU density
- DDR5 Memory for more performance, less power, and greater data consistency per server
- Trusted edge-to-cloud security posture built on an HPE compute core hardened through a proven, zero trust approach to security
- Intuitive cloud operating experience through a unified console to automate and simplify operations
- Enhanced chip-to-chip communications for breakthrough performance, throughput, and efficiency in accelerated solutions

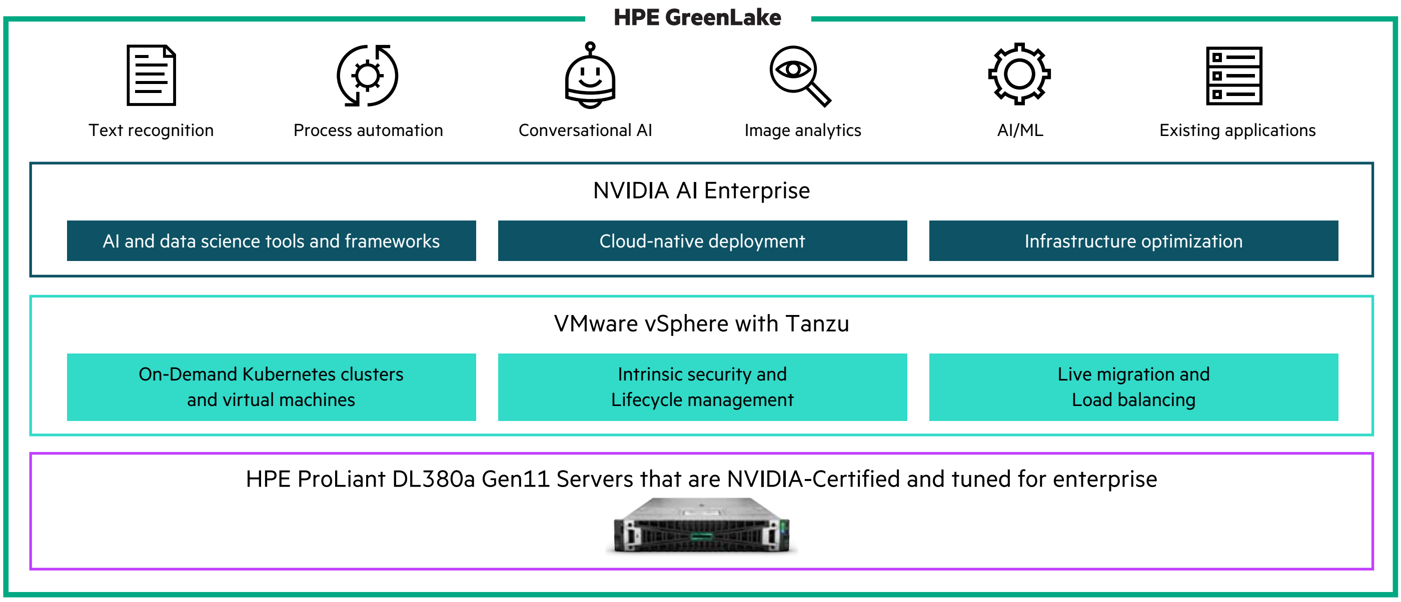

HPE GreenLake
Text recognition
Process automation
Conversational AI
Image analytics
AI/ML
Existing applications
NVIDIA AI Enterprise
AI and data science tools and frameworks
Cloud-native deployment
Infrastructure optimization
VMware vSphere with Tanzu
On-Demand Kubernetes clusters and virtual machines
Intrinsic security and Lifecycle management
Live migration and Load balancing
HPE ProLiant DL380a Gen11 Servers that are NVIDIA-Certified and tuned for enterprise

Figure 2. HPE ProLiant and NVIDIA AI Enterprise: End users can access the software they need to build successful AI projects, and IT admins can support the projects using familiar tools.

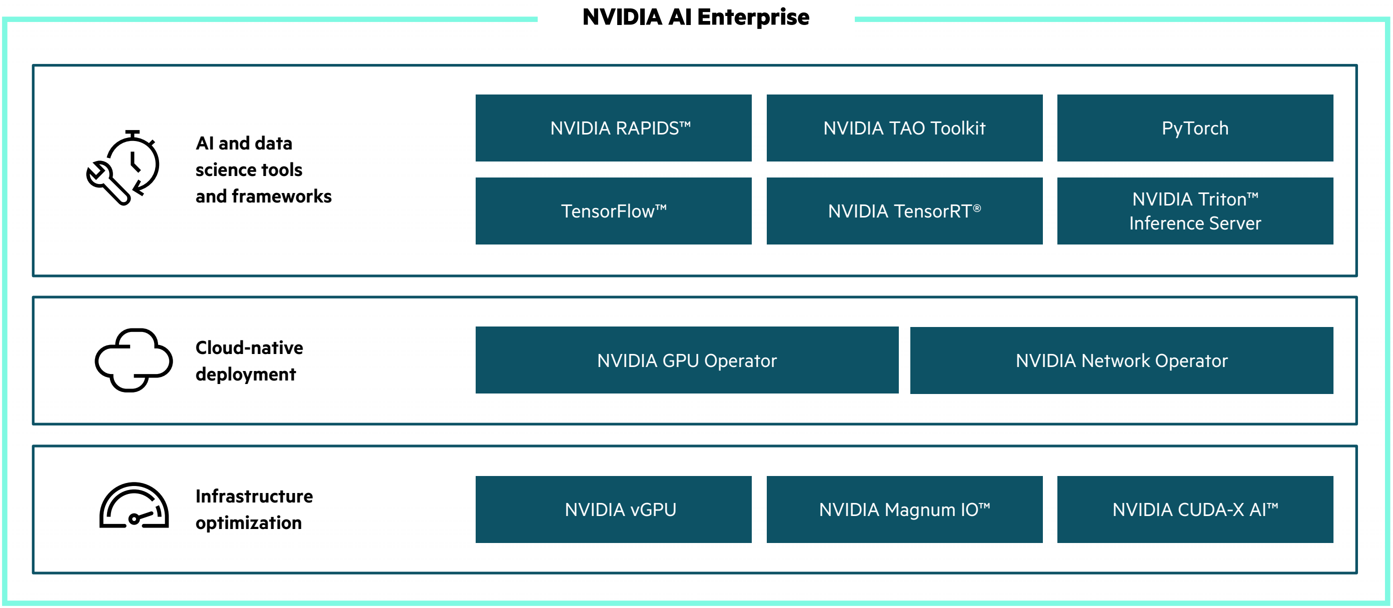


Figure 3. The NVIDIA AI Enterprise software suite includes applications, frameworks, and tools used by AI researchers, data scientists, and developers, as well as tools for infrastructure optimization.

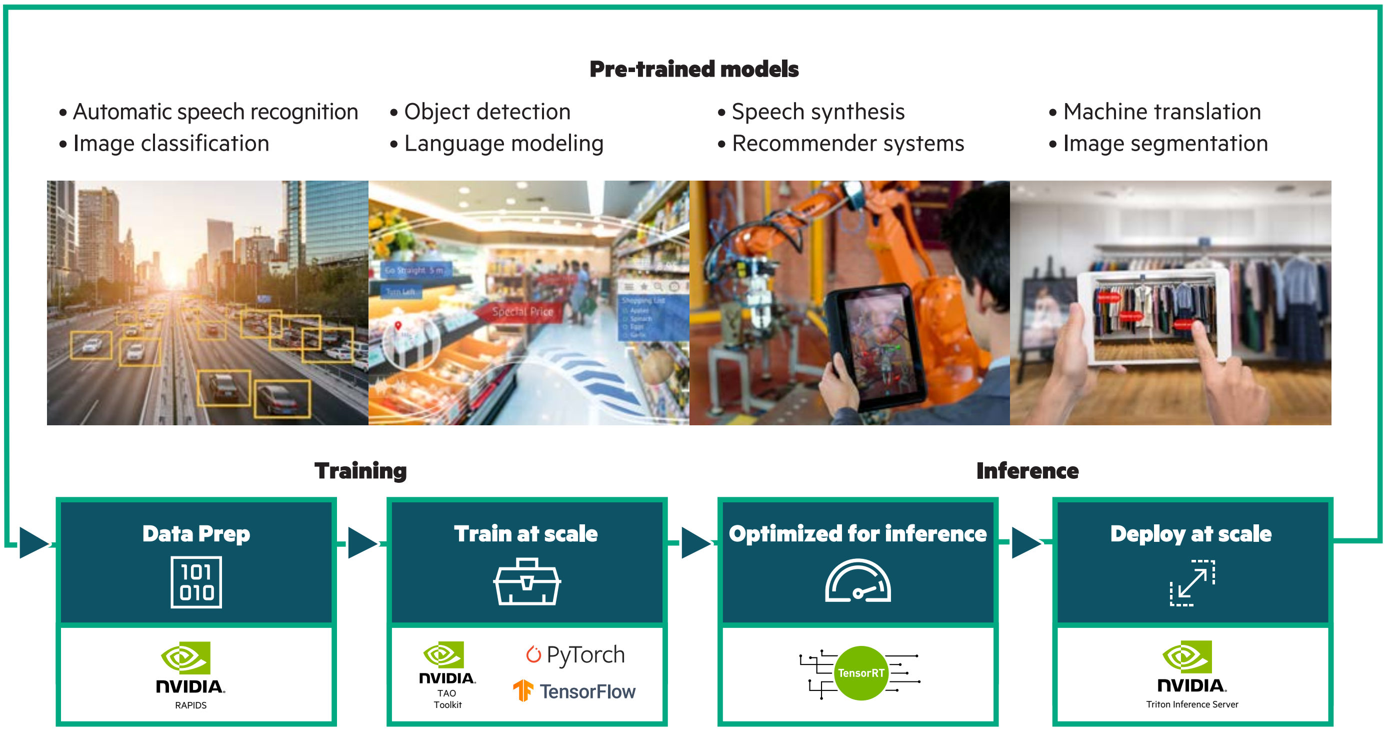
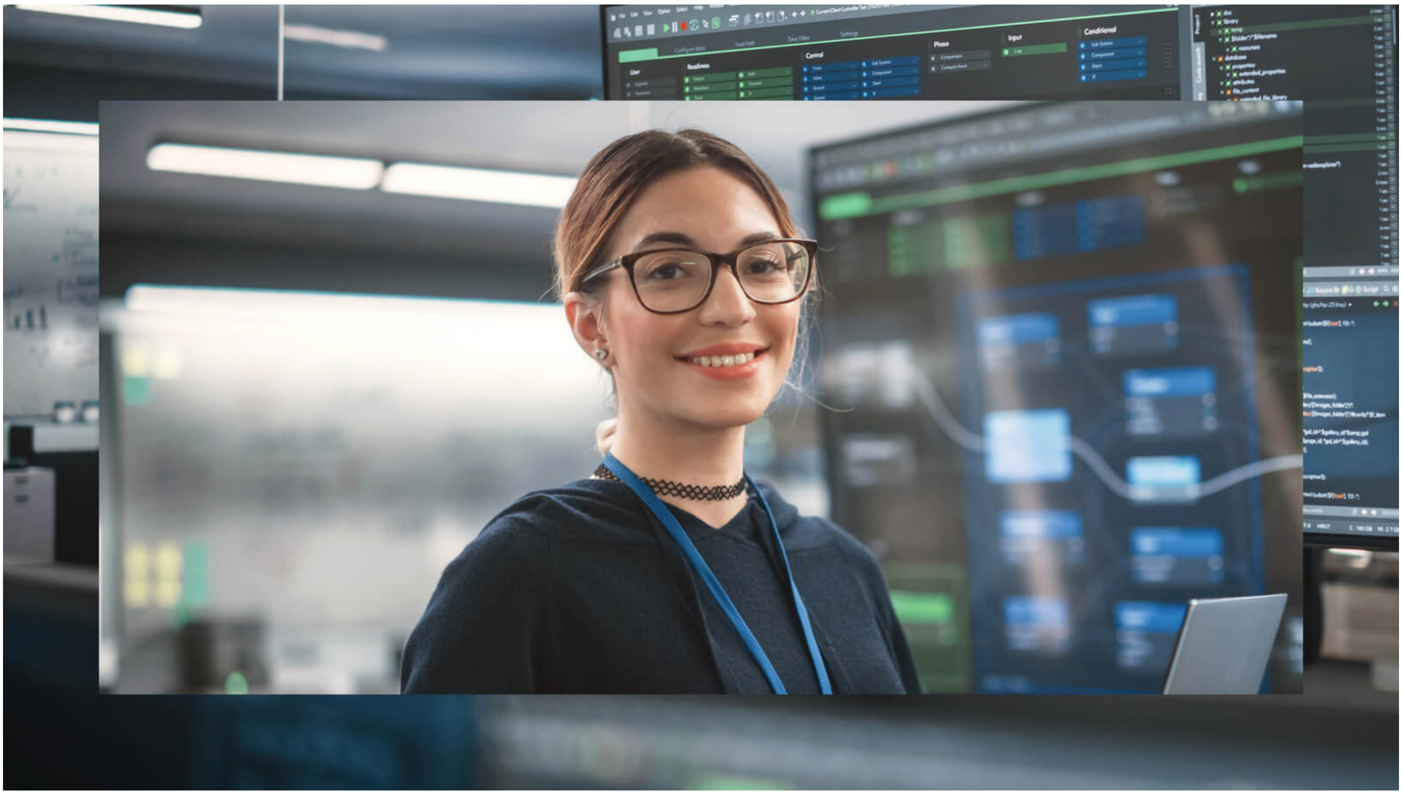


Figure 4. Common use cases deployed on NVIDIA AI Enterprise include automatic speech recognition, image classification, object detection, language modeling, speech synthesis, recommender systems, machine translation, and image segmentation.

Better together: HPE, NVIDIA, and VMware®

HPE, NVIDIA, and VMware have come together to help businesses unlock the power of AI — by delivering an end-to-end enterprise platform optimized for AI workloads for NVIDIA AI Enterprise. The platform is deployed on industry-leading HPE ProLiant servers that are NVIDIA-Certified to accelerate the speed at which developers can build AI and high-performance data analytics.

Enabling organizations on AI and deep learning through online and instructor-led workshops, reference architectures, and benchmarks on NVIDIA GPU-accelerated applications to enhance time to value.



Schedule an AI workshop today

Contact us today to learn more about our AI workshops designed to help you accelerate your AI journey.

HPE and NVIDIA Alliance info

Find out more

HPE and VMware Alliance info

hpe.com/us/en/alliance/vmware.html

Learn more at

HPE ProLiant AI inference solutions

Chat now (sales)

Hewlett Packard
Enterprise