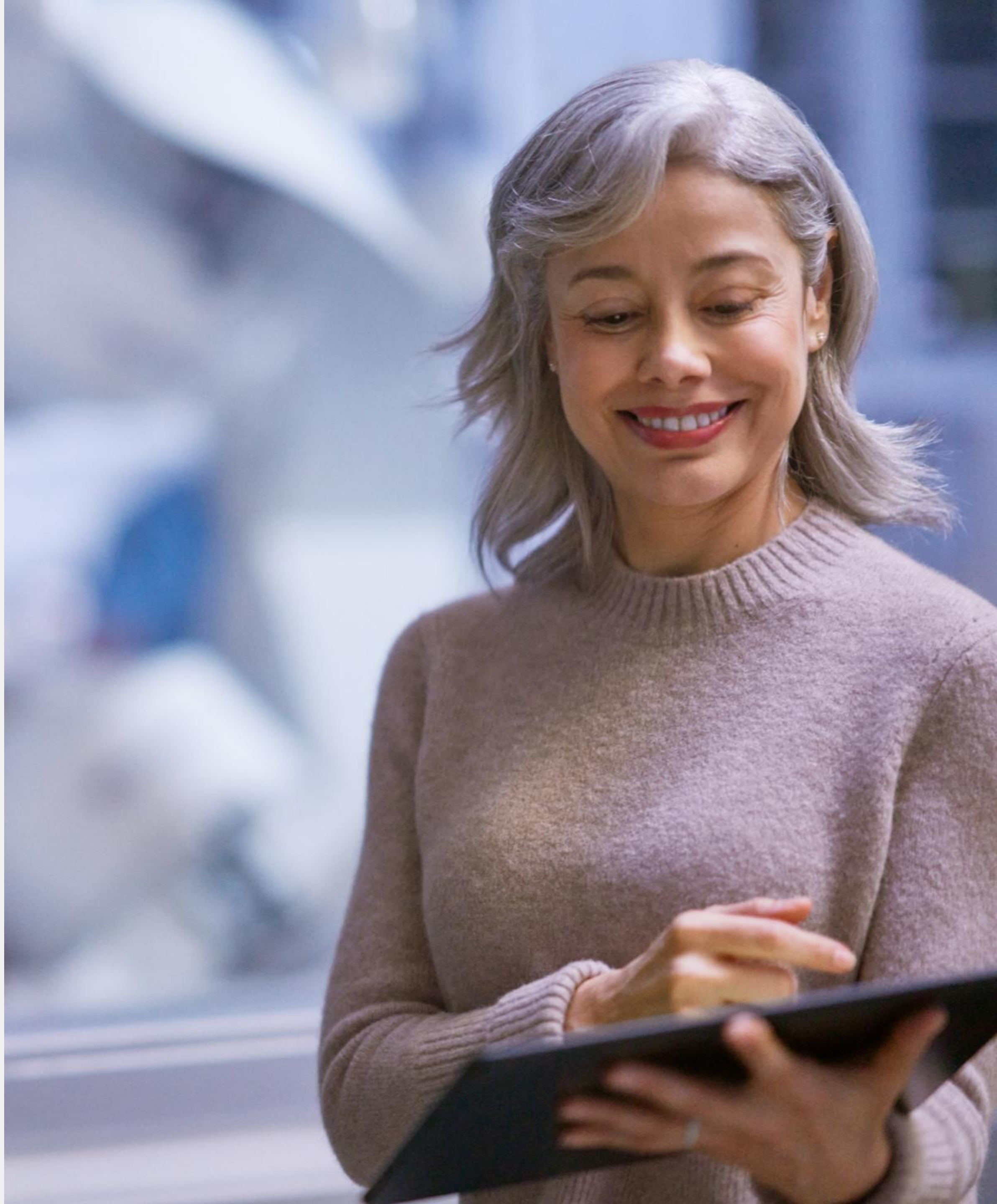
Microsoft Azure

NVIDIA

Meet AI Demands at Any Scale

Learn how Microsoft and NVIDIA are helping companies worldwide push the boundaries of AI innovation.



Contents

01

Why purpose-built infrastructure is critical for AI

AI-first approach to infrastructure

AI is one of the most inspiring areas of technology evolution and is emerging as a key differentiator for businesses worldwide. Performance requirements for AI, however, are significantly different from other enterprise applications.

Unlike conventional workloads, increasingly sophisticated AI models with billions of parameters require massive amounts of processing power plus lightning-fast networking and storage.

AI requires infrastructure built specifically for compute-intensive, large-scale AI workloads.

GPU-accelerated virtual machines, powerful host processors, and fast system memory interconnected together by a high-bandwidth, low-latency network running AI-optimized software are at the core of infrastructure purpose-built for AI.

Such a data center architecture allows tens, hundreds, or thousands of GPUs to work together on a single task, sharing the load and delivering the overall processing power needed to deliver the high performance, reliability and scale required to run complex AI algorithms.

An "AI-first" approach to infrastructure can accelerate model training and inference, increase performance and accuracy, and accelerate AI innovation.

Accelerating AI for growth: The key role of infrastructure

IDC illustrates growing importance of purpose-built AI Infrastructure

Do more with less in the cloud

The level of compute power and scalability needed for AI projects is difficult and costly to implement and maintain on-premises.

Even organizations with their own datacenters often do not have the right systems to handle complex model training or keep up with the rapid pace of new technologies. They find AI-related tasks can tie up existing system capacity for days to weeks or even months.

To meet performance-intensive computing demands required by AI and keep pace with technology advancements, IDC forecasts that by 2025 nearly 50 percent of all accelerated infrastructure will be cloud-based[1]. Cloud service providers can offer the latest and best technology available, ahead of even the largest system vendors.

With cloud-based infrastructure, businesses can immediately take advantage of the most advanced processors, accelerators, networks, storage, and software available for AI workloads.

The flexibility, power, and speed provided by cloud infrastructure can help companies quickly deploy and scale AI solutions and stay competitive without investing time and money in on-premises hardware and software.

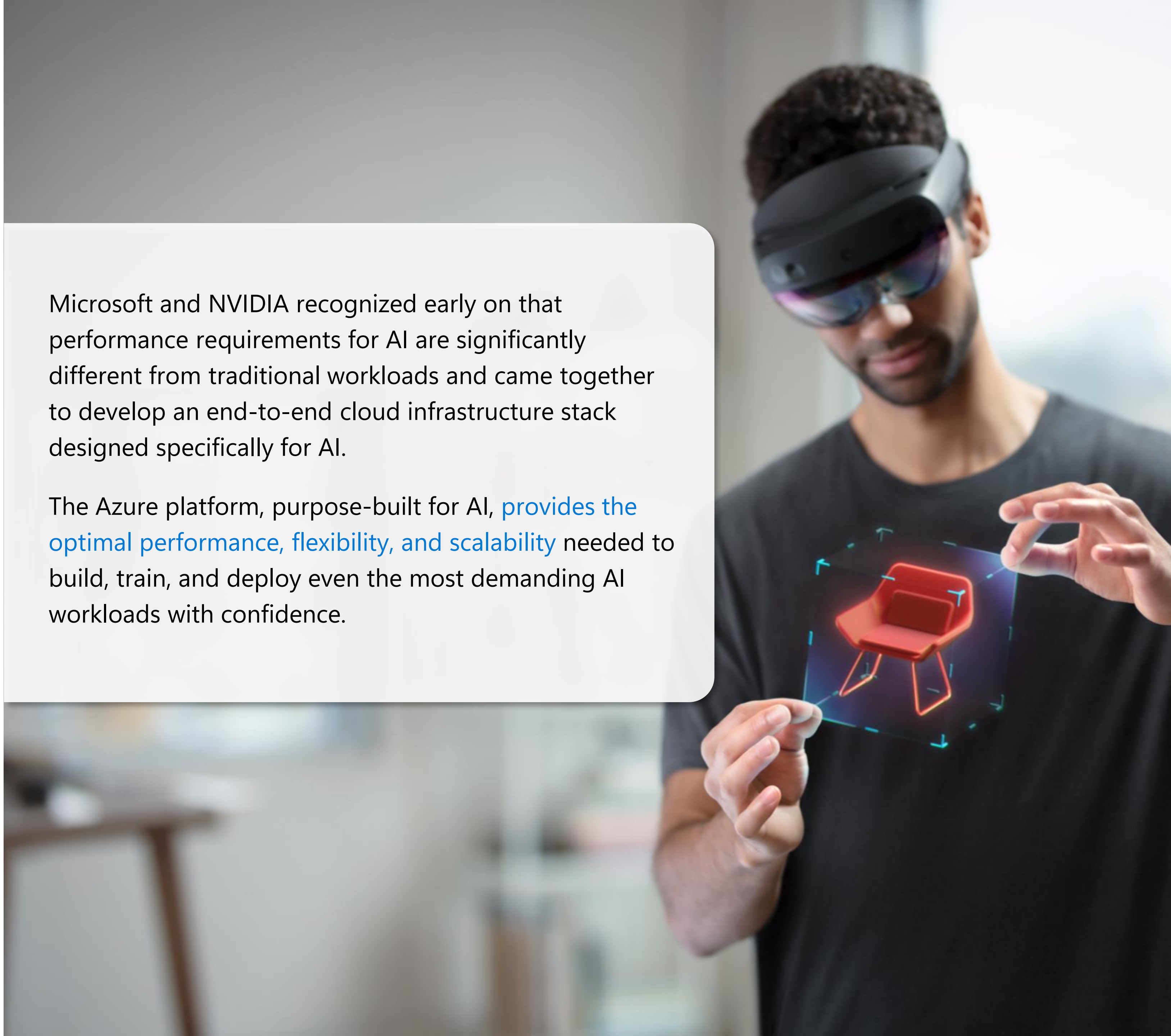
But not all cloud infrastructures are equal.

[1] IDC FutureScape: Worldwide Cloud 2022 Predictions

02

A powerful platform for AI at any scale

AI-first infrastructure from AI leaders



Microsoft and NVIDIA recognized early on that performance requirements for AI are significantly different from traditional workloads and came together to develop an end-to-end cloud infrastructure stack designed specifically for AI.

The Azure platform, purpose-built for AI, provides the optimal performance, flexibility, and scalability needed to build, train, and deploy even the most demanding AI workloads with confidence.

Microsoft

“AI is fueling the next wave of automation across enterprises and industrial computing, enabling organizations to do more with less as they navigate economic uncertainties.

Our collaboration with NVIDIA unlocks the world’s most scalable supercomputer platform, which delivers state-of-the-art AI capabilities for every enterprise on Microsoft Azure.”

Scott Guthrie
Executive Vice President Cloud and AI

NVIDIA

“AI technology advances as well as industry adoption are accelerating. The breakthrough of foundation models has triggered a tidal wave of research, fostered new startups, and enabled new enterprise applications.

Our collaboration with Microsoft will provide researchers and companies with state-of-the-art AI infrastructure and software to capitalize on the transformative power of AI.”

Manuvir Das
Vice President
Enterprise Computing at NVIDIA

Purpose-built, full-stack infrastructure for AI

Transformative AI services

Trusted machine learning platform

Optimized, scalable compute

End-to-end workload orchestration

Fast, secure networking

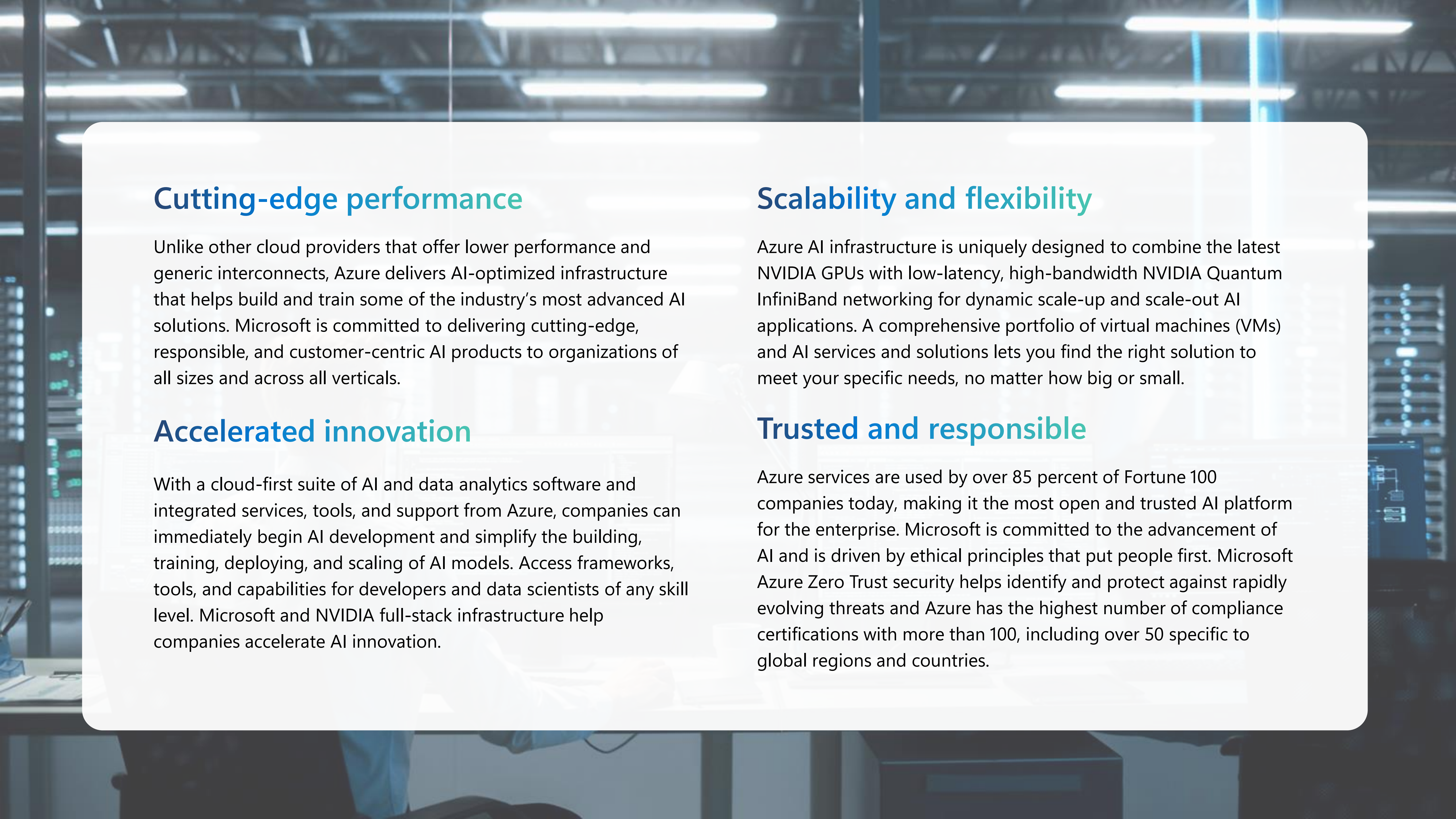
High-performing storage

Azure AI infrastructure, featuring the latest NVIDIA GPUs, combines hardware and software optimized for compute-intensive AI workloads, empowering businesses to develop AI-enabled products and services at any scale.

Build, deploy, and manage high-quality models faster using Azure Machine Learning. Access high-quality vision, speech, language, and decision-making AI models through simple API calls with Azure AI services.

Create your own machine learning models using an AI supercomputing infrastructure, familiar tools like Jupyter Notebooks and Visual Studio Code, and open-source frameworks like TensorFlow and PyTorch. Seamlessly orchestrate your AI workloads on the cloud with Azure Batch and Azure CycleCloud.

Inside Microsoft's AI supercomputer



Cutting-edge performance

Unlike other cloud providers that offer lower performance and generic interconnects, Azure delivers AI-optimized infrastructure that helps build and train some of the industry’s most advanced AI solutions. Microsoft is committed to delivering cutting-edge, responsible, and customer-centric AI products to organizations of all sizes and across all verticals.

Accelerated innovation

With a cloud-first suite of AI and data analytics software and integrated services, tools, and support from Azure, companies can immediately begin AI development and simplify the building, training, deploying, and scaling of AI models. Access frameworks, tools, and capabilities for developers and data scientists of any skill level. Microsoft and NVIDIA full-stack infrastructure help companies accelerate AI innovation.

Scalability and flexibility

Azure AI infrastructure is uniquely designed to combine the latest NVIDIA GPUs with low-latency, high-bandwidth NVIDIA Quantum InfiniBand networking for dynamic scale-up and scale-out AI applications. A comprehensive portfolio of virtual machines (VMs) and AI services and solutions lets you find the right solution to meet your specific needs, no matter how big or small.

Trusted and responsible

Azure services are used by over 85 percent of Fortune 100 companies today, making it the most open and trusted AI platform for the enterprise. Microsoft is committed to the advancement of AI and is driven by ethical principles that put people first. Microsoft Azure Zero Trust security helps identify and protect against rapidly evolving threats and Azure has the highest number of compliance certifications with more than 100, including over 50 specific to global regions and countries.

Proven performance and scale

#3

Supercomputer Top500 List 2023 [1]

#1

Cloud provider Top500 List 2023 [1]

30%

Faster training for LLMs [2]

2X

Faster throughput per GPU [3]

3X

Estimated ROI for machine learning projects [4]

Scale record

in LLM training MLPerf 3.1 2023 [5]

03

Proving it through real-world solutions

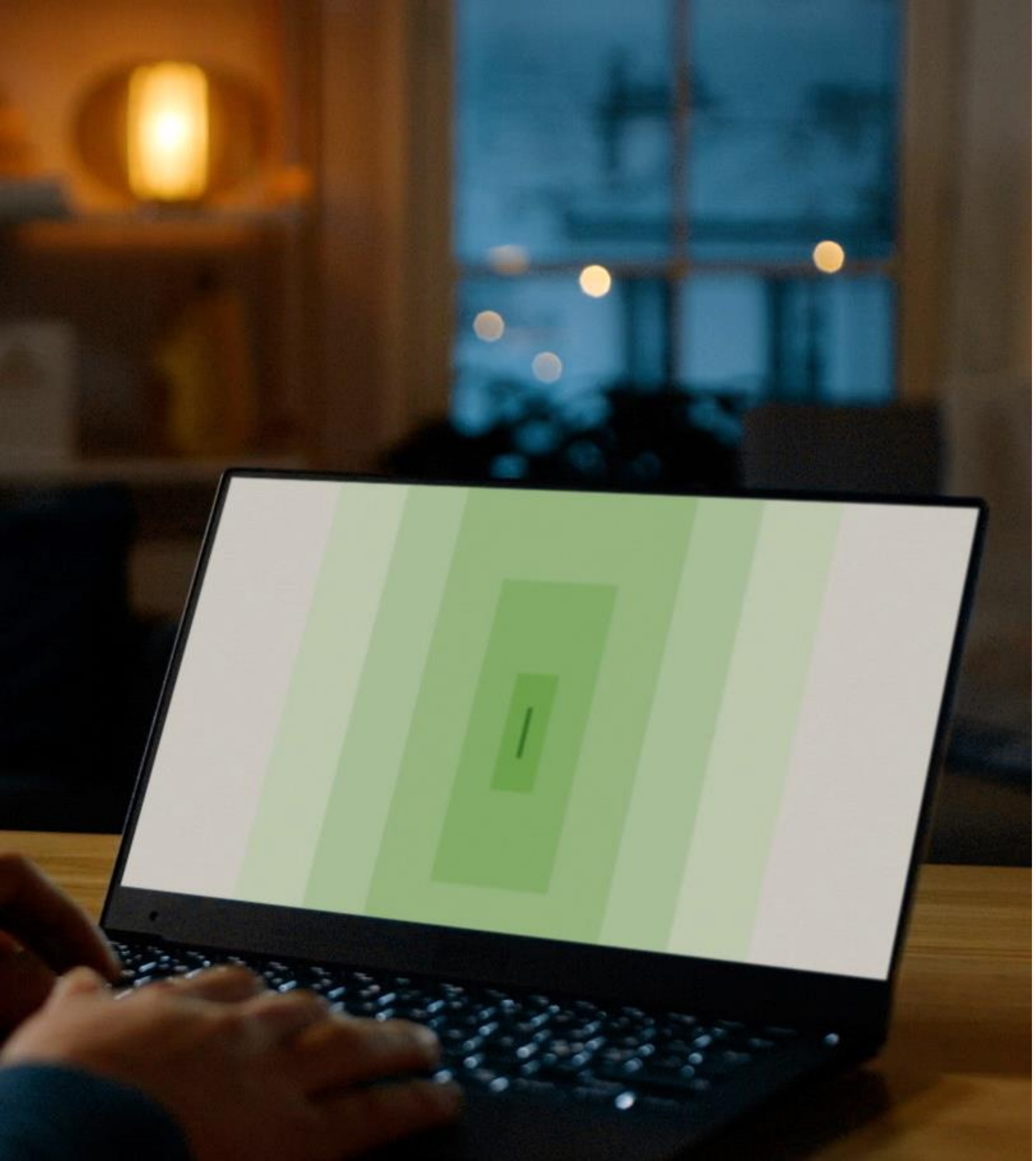
WAYVE

Wayve is developing the next generation of automated vehicles, AV2.0, by using machine learning in complex urban environments, overturning traditional hand-coding and robotics approaches. Its deep learning approach requires vast amounts of data and a platform capable of managing it.

By using Azure AI infrastructure, Azure Machine Learning and PyTorch, an open-source machine learning framework, Wayve can gather, manage, and process petabytes of images, GPS, and other data to build and iterate driving models for complex urban environments.

Wayve can iterate and prototype faster and adjust machine learning models more nimbly. It has increased data throughput by 50X and reduced its AV2.0 model training time by 90% even while it trains massive computer vision systems with petabytes of data at an unprecedented scale.

Watch the video
Read the full story



Inflection

Inflection AI sought to accelerate time to AI development and reduce downtime, better positioning it to be a leader in empathetic personal intelligence. Inflection AI turned to Microsoft Azure.

Underpinned by reliable and stable Microsoft Azure AI technology, Inflection's Pi chatbot works with large language models to help people organize their days and offer them advice while matching their knowledge and interests.

Accessing powerful AI-optimized Azure virtual machines with InfiniBand networking through Azure AI infrastructure has helped Inflection AI gain superfast connections between GPUs and accelerate training of the largest language models available.

Read the story
Watch the video

WILDLIFE PROTECTION SOLUTIONS

Wildlife Protection Solutions (WPS) uses remote cameras to gather image data about the status of the species they protect, but the number of images that must be analyzed before action can be taken is overwhelming.

Wildlife Protection Solutions (WPS) overcomes this barrier with MegaDetector, an AI model developed by Microsoft AI for Earth, running on Microsoft Azure virtual machines powered by NVIDIA GPUS to accelerate the processing of camera trap images.

MegaDetector improved threat detection accuracy and processes images faster than other AI models tried—in some cases, by 50 percent. The processing power of Azure AI infrastructure makes a critical difference and combines price and performance gains.

Read the full story
Watch the video

NUANCE®

Doctors spend a significant portion of their days doing medical administration and entering patient data into electronic health records (EHRs). Nuance wanted to provide a solution that would reduce data entry time and improve patient experience.

Nuance created a conversational AI solution, Dragon Ambient eXperience (DAX), using Azure Machine Learning and PyTorch on Azure infrastructure, that records the doctor-patient conversation and automatically documents patient encounters at the point of care.

By building DAX on Azure and PyTorch, Nuance created a fast and highly scalable solution and can now train models 2.5X faster. Patients experience a more engaging, personal visit because doctors can focus on the patient rather than entering notes into a computer and doctors report significant time savings.

Read the story

04

AI infrastructure from Microsoft and NVIDIA

Azure Virtual Machines powered by NVIDIA GPUs

From fractional GPUs to multiple GPUs across multiple nodes for distributed computing, Microsoft and NVIDIA provide the right-sized GPU acceleration for your AI workload.

ND H100 v5-series

This Azure virtual machine series features eight NVIDIA® H100 80GB Tensor Core and can scale up to thousands of GPUs with 3.2Tb/s of interconnect bandwidth per VM. Each GPU within the VM is provided with its own dedicated, topology-agnostic 400 Gb/s NVIDIA Quantum-2 CX7 InfiniBand connection with NVLINK 4.0 connectivity.

This series is best suited for high-end deep learning training and tightly coupled scale-up and scale-out Generative AI workloads.

Learn more

ND A100 v4-series

This Azure virtual machine series features eight NVIDIA® A100 40GB Tensor Core GPUs, NVIDIA NVLink® 3.0, and a dedicated NVIDIA Quantum 200 gigabits per second (Gb/s) InfiniBand connection per virtual machine (VM) for scale out, multi-node, multi-GPU distributed computing.

This series is best suited for AI training, deep learning inference, machine learning, industrial HPC, and data analytics workloads.

Learn more

NC A100 v4-series

This series has the flexibility to select one, two, or four NVIDIA® A100 80GB Tensor Core GPUs per VM to provide the right-sized GPU acceleration for your workload. NVIDIA NVLink 3.0 is supported for GPU-to-GPU communication within the VM.

This series is best suited for single-node deep learning training, batch inference, interactive machine learning development and exploration, modeling, simulation, and data analytics.

Learn more

Azure AI portfolio

Azure AI services and machine learning platform help developers and organizations rapidly create intelligent, cutting-edge, market-ready, and responsible applications.

Azure AI services

Microsoft offers a portfolio of artificial intelligence (AI) services designed for developers and data scientists, helping you do more with less. Take advantage of the decades of breakthrough research, responsible AI practices, and flexibility that Azure AI offers to build and deploy your own AI solutions.

Access high-quality vision, speech, language, and decision-making AI models through simple API calls, and create your own machine learning models with tools like Jupyter Notebooks, Visual Studio Code, and open-source frameworks like TensorFlow and PyTorch.

Azure Applied AI Services help you deploy AI solutions quickly—no machine learning expertise required. Azure Cognitive Services lets developers and data scientists of all skill levels to easily add AI capabilities to their apps.

Learn more

Azure Machine Learning

Azure Machine Learning is an enterprise-grade service that provides business-critical machine learning models at scale, enabling developers to build, deploy, and manage models faster.

Automate machine learning to identify suitable algorithms and tune hyperparameters faster. Improve productivity and reduce costs with autoscaling GPU clusters and built-in machine learning operations, and seamlessly deploy to the cloud and the edge.

Access all these capabilities from any Python environment using open-source frameworks such as PyTorch, TensorFlow, and scikit-learn. Azure Machine Learning also integrates with NVIDIA Triton Inference Server and NVIDIA RAPIDS™ to provide more performance gains.

Learn more

Azure workload and storage services

Easily build, manage and optimize your AI workloads with industry-leading tools for end-to-end AI workflow agility and high-performance, secure storage.

Azure Batch

Azure Batch runs large-scale applications efficiently in the cloud. Schedule compute-intensive tasks and dynamically adjust resources for your solution without managing infrastructure.

Choose the operating system and development tools you need. Scale to tens, hundreds, or thousands of virtual machines. Stage data and execute compute pipelines. Pay only for what you use.

Learn more

Azure CycleCloud

Azure CycleCloud helps enterprise IT organizations provide secure and flexible cloud HPC and big-compute environments to their end users. With dynamic scaling of clusters, you get the resources needed at the right time and the right price.

Automated configuration from Azure CycleCloud allows IT to focus on providing service to business users.

Learn more

Azure Virtual Machine Scale Sets

Azure Virtual Machine Scale Sets lets you create and manage a group of heterogeneous load-balanced virtual machines (VMs).

Build on your terms, large-scale services for batch, big data, and container workloads Increase or decrease the number of VMs automatically in response to demand or based on a schedule you define.

Learn more

Azure Managed Lustre

Azure Managed Lustre is designed for data-intensive workload and is an open-source parallel file system that can scale to massive storage sizes while also providing high performance throughput and Azure multi-level security.

Azure Managed Lustre is used by the world's fastest supercomputers and in data-centric workflows for many types of industries.

Learn more

NVIDIA AI solutions

NVIDIA AI Enterprise

GPU-accelerated instances on Microsoft Azure are certified and supported with NVIDIA AI Enterprise, a fully managed and secure, cloud-first suite of AI and data analytics software.

NVIDIA AI Enterprise streamlines each step of the AI workflow, from data processing and AI model training to simulation and large-scale deployment, reducing the time to move from pilot to production of AI solutions.

NVIDIA AI Enterprise is certified on Azure Virtual Machines.

Learn more

NVIDIA Modulus

NVIDIA Modulus is a neural network framework blending the power of physics in the form of governing partial differential equations with data to build high-fidelity, parameterized surrogate models with near-real-time latency.

Learn more

NVIDIA NeMo

NVIDIA NeMo offers an easy, efficient, and cost-effective containerized framework to build and deploy large language models.

Learn more

NVIDIA NGC

NVIDIA NGC is a collection of fully managed cloud services, including natural language understanding and speech AI solutions. NGC hosts a catalog of GPU-optimized AI software, SDKs, and Jupyter Notebooks to accelerate AI workflows.

Learn more

NVIDIA Riva

NVIDIA Riva is a GPU-accelerated speech AI SDK for building and deploying fully customizable, real-time speech processing AI pipelines with high accuracy.

Learn more

Make AI your reality

Get the purpose-built cloud infrastructure your AI projects demand

Choosing infrastructure that's highly performant, versatile, and scalable is key to building and deploying AI-enabled products and services at scale. With purpose-built, full-stack cloud infrastructure designed to simplify and accelerate end-to-end AI workflows, Microsoft and NVIDIA make AI-powered applications and services a reality.

Whether your project is big or small, local, or global, Microsoft Azure and NVIDIA are empowering companies worldwide to push the boundaries of AI innovation.

- Azure AI infrastructure
- NVIDIA accelerated computing on Azure
- Azure AI portfolio
- Rethinking Cloud Strategies for Advanced AI

Microsoft Azure

Limitless innovation

This document is provided "as-is." Information and views expressed in this document, including URL and other internet website references, may change without notice. You bear the risk of using it. Some examples are for illustration only and are fictitious. No real association is intended or inferred. This document does not provide you with any legal rights to any intellectual property in any Microsoft product. You may copy and use this document for your internal, reference purposes.