

Azure AI Infrastructure

Why AI Infrastructure Matters

AI is changing the way the world works, and advancements in AI infrastructure are at the center of that revolution. AI infrastructure powers your AI innovations—from optimizing hardware, processors, and accelerators, to networking and virtualization.

Requirements for AI infrastructure are advancing in two fundamental ways. First, the ever-increasing complexity of AI workloads requires having an infrastructure that’s optimized to provide massive scale with parallel computation. Second, the increasing availability of AI infrastructure in the cloud is making AI innovation more accessible to organizations of all sizes.

Performance at scale

Azure AI infrastructure provides access to large-scale AI models, GPU-based virtual machines (VMs), AI supercomputing, and AI services, allowing customers to build, train, test, and deploy their AI applications with speed and efficiency.

Accelerate your **AI innovations** with **AI infrastructure**

2x faster throughput¹
Per GPU compared to competitors

#1 for cloud performance²
among Top 500 supercomputing rankings

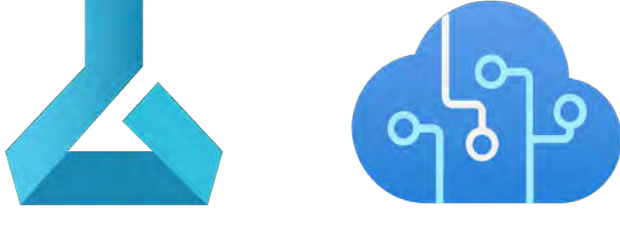
Microsoft: optimized for AI

Microsoft is on a continuous journey to revolutionize computing advancements by leading the development of AI-optimized infrastructure.

Customers that want to lead in the AI era need cloud computing infrastructure optimized specifically for AI workloads. With highly scalable interconnected GPUs powering virtual machines (VMs), Azure AI infrastructure helps remove bottlenecks and deliver rapid data processing throughput and low latency—a must for AI.

The infrastructure is engineered for Azure AI services, like Azure Machine Learning, to operationalize models with MLOps seamlessly, and Azure AI services for pre-trained foundation models. It also natively integrates with open-source frameworks and optimization technologies such as PyTorch and ONNX Runtime. This integration allows developers and data scientists to build and deploy high-performing AI models quickly and easily.

Azure AI



Open source



Proven workload performance

Azure AI infrastructure offers proven supercomputing performance for the most advanced AI initiatives. It supports cross-platform interoperability and collaboration, letting you use their preferred tools and languages across different environments and devices.

Microsoft runs on Azure AI infrastructure



7.5 trillion
characters translated
per month in
Microsoft Edge



45 million
meeting hours
transcribed per month
in Microsoft Teams



100 million
monthly active users
of AI text prediction
in Microsoft 365

Partnerships that power ingenuity



Azure is the exclusive cloud provider for OpenAI, powering workloads across research, products, and API services.



Azure supercomputing is enabling **AI-based autonomous driving** technology.



Azure HPC provides the agility to **scale storage and compute resources for cancer research**.

Purpose-built for innovation

Your considering AI as part of your innovation strategy, and that means you’ll need secure, high-performance infrastructure to support those workloads. Azure AI offers a range of GPU-based VMs that can handle different workloads and scenarios, including training large AI models and running high-performance computing (HPC) simulations.

From current and future-generation transformer-based models for natural language processing (NLP), to scalable implementations for midrange training projects and inference, Azure provides the flexibility required to strike the optimal balance between cost and performance.

A track record you can trust

Azure has been scaling supercomputing capabilities since 2020, deploying new infrastructure and services. Here’s a look at some of the most important milestones from the last few years:

2020

2021

2022

2023

Built for OpenAI

Microsoft ranks **Top 5** in supercomputing³

Microsoft announces AI supercomputing VMs available to customers

First in cloud supercomputing performance

Microsoft ranks **#1** for supercomputing⁴

Azure delivered **#1** performance by a cloud provider in MLPerf results⁵

Reaching new supercomputing milestones

Microsoft partners with NVIDIA on latest GPUs

Microsoft runs a 530B-parameter NeMo Megatron benchmark⁶ on **175** VMs

Extending AI capabilities

Microsoft announces Azure **OpenAI** Service and the new Bing trained with Azure AI infrastructure

Take the next steps

Connect with Azure Sales to learn more ›

¹ Microsoft performance benchmark for 1-Trillion parameter model with DeepSpeed, compared to another published cloud vendor running 1-Trillion parameter model with DeepSpeed optimized for throughput. As of July 2022.
² Microsoft Azure ranks #10 on supercomputing performance and #1 for cloud-based performance—Top500 List. Published November 2022.
³ Microsoft supercomputer for OpenAI, completed in May 2020, was a single system with more than 285,000 CPU cores, 10,000 GPUs and 400 gigabits per second of network connectivity for each GPU server. Compared with other machines listed on the Top500 supercomputers in May 2020, ranks in the top five.
⁴ Microsoft Azure ranks #10 on supercomputing performance and #1 for cloud-based performance—Top500 List. November 2021.
⁵ MLPerf results show a debut performance by Azure delivering the #2 performance overall and the #1 performance by a cloud provider.
⁶ Microsoft Azure published NVIDIA NeMo Megatron benchmarking demonstrating Scale to 530B Parameter GPT-3 Model