

# Unlock the Power of Enterprise AI

Full-Stack Solutions Accelerate AI Deployment

## Are You Ready for Enterprise AI?

AI and emerging GenAI technologies are disrupting the status quo, prompting enterprises to accelerate their AI initiatives. However, many organizations report being unprepared to address the infrastructure and other requirements to support diverse AI use cases. To navigate these challenges, most enterprises are adopting a hybrid approach, utilizing diverse infrastructures across datacenters, the edge, and the cloud.

Public cloud shared LLMs and use cases based on public data

Dedicated platforms for turning LLMs and use cases based on sensitive or proprietary data

Edge inference for distributed use cases

Hybrid and multicloud for shared access to data storage and workflow integrations

69%

said that GenAI will significantly disrupt their competitive positioning and operating models within 18 months

91%

believe they will need to modernize infrastructure for AI

## Popular Enterprise AI Use Cases

There are numerous AI use cases to consider:

### AI Assistant

Increase efficiency, improve customer service and deliver a more personalized experience.

### Content or Code Copilot

Accelerate content creation and code development. Fine-tune using proprietary data or leverage Retrieval Augmented Generation (RAG).

### Industry-Specific

Build or license AI models to help with fraud detection, medical imaging or other industry-specific use cases.

### Private GPT

Prevent confidential company data from "leaking" onto public AI services.

## Choosing AI Infrastructure

Infrastructure is the largest GenAI expense, with options ranging from build-your-own solutions to public cloud and full-stack offerings. Informed decisions can help reduce upfront CapEx and manage development costs.

Guidelines for avoiding common pitfalls:

Avoid point solutions

Choose infrastructure that supports multiple AI use cases to prevent technology silos.

52%

of Gen AI spending will be for infrastructure

Prioritize flexibility

Select infrastructure that can adapt to evolving business and technical needs.

29%

Dedicated

Maintain control over data and IP

Choose solutions with robust security and data protection, ensuring data governance is not compromised.

23%

Public Cloud

## Do I Need GPUs?

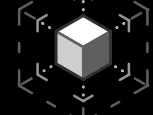
While GPUs are often considered essential for AI, many models can function effectively with CPUs alone.

Intel® Advanced Matrix Extensions (AMX), built into Intel® Xeon® Processors 4th Generation and later, supports AI inference for smaller GenAI models without GPUs.



### Inference

LLMs up to 10 billion parameters



### AI Pipeline

Any stage of the AI pipeline, except for large training workloads



### RAG

Accelerate retrieval augmented generation (RAG) in which a model uses data retrieved from a database

Intel® AMX is ideal for



### Cost Control

Save capital, operational, and licensing costs and decrease power and cooling demand

## A Full-Stack AI Solution

Nutanix, Cisco, and Intel have joined forces to provide a fully integrated, validated AI solution that simplifies AI integration.

### Nutanix

#### Nutanix Cloud Platform

A complete infrastructure platform for VM and container operations across datacenters, edge, and public cloud.

#### GPT-in-a-Box

A full-stack AI platform for building and deploying AI applications with full control over data and infrastructure.

### Cisco

#### Cisco UCS

A unified computing system that brings together compute, networking, and storage to power your applications.

#### Intersight

IT operations platform that can see, control, and automate HCI at a global scale.

### Intel

#### Intel® Xeon® Processors

The latest Xeon Processors with Intel® AMX are featured in Cisco UCS and optimized for Nutanix GPT-in-a-box, enhancing AI performance.

#### Broad Portfolio

Intel® Xeon® Processors address AI workload needs, boosting performance, energy efficiency, and security.

## Solution Benefits

One unified platform for AI and other workloads

Collaborative support strengthens teamwork

Integrates AI into mainstream IT operations

Simplifies end-to-end operations with Prism and Intersight

Ideal for datacenter and edge environments

Optimizes performance by storing valuable data on clusters

## Ready to Find Out More?

See why IDC believes that a full-stack AI solution with CPU acceleration is key to overcoming Enterprise AI challenges.

[Read the IDC White Paper](#)