White Paper

Unlocking AI and Generative AI Use Cases with AI-Ready Hybrid Cloud Infrastructure

Sponsored by: Nutanix

Dave Pearson
July 2024

IDC OPINION

Emerging AI and generative AI (GenAI) for enterprise use cases empower modern digital businesses by augmenting efficiency, optimizing operations, and unlocking insights from vast data sets. The ability to gain insight from disparate data sets, automate repetitive tasks, forecast trends, and personalize customer experiences can revolutionize workflows across industries. According to IDC's 2024 *Future Enterprise Resiliency and Spending (FERS) Survey, Wave 1,* 69% of respondents said that GenAI foundation models, platforms, and application technologies will significantly disrupt their competitive positioning and operating models within 18 months.

Enterprise AI and GenAI have presented organizations with new opportunities, but at the same time, the journey to AI transformation can prove difficult and expensive, exposing organizations to new risks when not planned, executed, and operated properly. In IDC's 2024 *FERS Survey, Wave 2,* the top 3 critical investment domains found to be immune to budget reductions were:

- Security, risk, and compliance (44% driven by AI/GenAI increases)
- Infrastructure and IT optimization (27% driven by AI/GenAI increases)
- AI and automation initiatives

While this shows just how quickly and deeply AI and GenAI have embedded themselves into our planning for digital business transformation, it should be noted that the top tech strategy risk factor identified in that survey was "managing AI model building and training costs" (34%), taking the lead from inflation and tech supply chain risks for the first time.

Finding appropriate use cases for enterprise AI and GenAI and creating the infrastructure foundation to leverage their capabilities in an efficient, productive, and cost-effective way must be a priority for businesses of all sizes in every industry.

For many organizations, the fastest route to kick-starting AI and GenAI initiatives has been with cloud-based AI model development built on existing IP. However, for many use cases, especially but not limited to industry-specific ones, internal data is essential to delivering value, and, for a variety of reasons, that data cannot be exposed to external applications or workflows. Privacy, security, data protection, regulatory compliance, and internal governance policies may prevent organizations from using public cloud resources for tasks such as retrieval-augmented generation, in which large language models (LLMs) are finely tuned using custom data to develop the most accurate, appropriate, and trusted insights. Inferencing at edge locations is another common AI activity that can require dedicated infrastructure for AI and GenAI initiatives. The amount of data being generated at the edge, the cost and challenges in moving that data to other deployments in order to act on it, and the often sensitive nature of that data make yet another compelling case for dedicated infrastructure.

SITUATION OVERVIEW

Overall Customer Challenges

AI is a top priority for many organizations, but uncertainty around the types of AI, use cases, and deployment methodologies often places organizations into a difficult decision matrix before they even start to consider data engineering, model development, and tuning.

In IDC's 2024 *FERS Survey, Wave 1,* we found that organizations are shifting their focus between the types of AI. In less than half a year, more enterprises are focusing on generative AI (27% in January 2024 versus 20% in August 2023) while shifting away from predictive AI (34% versus 35%) and interpretive AI (39% versus 45%). This shift has been primarily driven by advances in GenAI use case identification and implementation.

IDC defines a use case as a business-funded initiative enabled by technology that delivers a measurable outcome. There are three broad types of enterprise generative AI use cases:

- Productivity use cases (document search and analysis, identification of data monetization opportunities, knowledge discovery, code copiloting, etc.) are aligned to work tasks such as summarizing a report, creating a job description, and generating code. GenAI functionality for productivity improvement is also being infused into existing applications. For many of these use cases, business value can be delivered purely through the content and data that the underlying foundation models have been pretrained on.
- Business function use cases (customer service bots, improvement of supply chain efficiency, prediction of customer needs, etc.) tend to integrate a model (or multiple models) with corporate data for use by a specific department or function (marketing, sales, procurement, etc.). Many organizations are testing these types of use cases but are concerned about intellectual property (IP) leakage and data governance. These business function use cases require integration with established enterprise applications and platforms, and their capabilities will need to reference or be constrained by their clients' business data and privacy concerns (customer data, product data, knowledge bases, etc.).
- Industry-specific use cases (fraud and theft protection in banking, route planning, adaptive learning, healthcare imaging and identification, etc.) will generally require more custom work (and, in some cases, may even require building your own generative AI model). Examples such as generative drug discovery in life sciences and generative material design for manufacturing have accelerated real-world activities by several orders of magnitude. These are likely to be a source of real business value creation for larger enterprises that are able to put together a sufficiently large set of training data or work with other parties in their ecosystem to share data for training the model.

Key enterprise GenAI use cases for early adopters include private GPT (on premises, based on custom data), copilots or virtual assistants for code development and content creation, and AI assistants to help with knowledge discovery and document insights. As organizations become more mature and confident in their approach to AI, these use cases will expand, and access to data resources across the company will demand a reduction in technological and operational silos. Greater collaboration between line of business (LOB) and IT as well as between business units addresses one origin of innovation silos; flexible, performant infrastructure that provides the capabilities, capacity, and flexibility to operationalize new AI and GenAI initiatives can help ameliorate the other.

The lack of available skills at all stages of AI transformation is a major concern to organizations — and not just data scientists and modelers. From business-focused areas such as AI strategies and operating models to policy development and business process reengineering and technical aspects such as security, privacy and trust in AI systems, and infrastructure modernization and implementation, organizations are seeking trusted partners to accelerate their journey. According to IDC's *FERS Survey, Wave 1,* the top strategic partnerships will come from IT consulting, systems integration, cloud providers, application developers, and digital infrastructure providers.

Determining the goals of AI transformation, the key use cases for your organization, the infrastructure required to operate it, and the stakeholders, skills, and partnerships required are all critical for positive business outcomes. Deploying AI infrastructure cost effectively is a key early goal in enterprise AI transformation, balancing the capabilities required for new initiatives and the economic pressures on IT to deliver insight efficiently.

Addressing Those Challenges

One of the key determinations that organizations must make early in their AI transformation is the "build versus buy" question. There are multiple approaches available to organizations today, each with its own pros and cons, and each AI workload and use case may have a unique set of requirements related to performance, security, data management, and costs. Organizations need to deploy AI-ready infrastructure in such a way that additional use cases can be added and integrated over time and wide ranges of data formats and sources can be supported, all the while understanding that AI and GenAI are just a part of their application portfolios. Ensuring that AI transformation doesn't increase point solutions and technology silos at the expense of other mission-critical initiatives is vital for effective and efficient innovation.

Collaborative governance across IT and cloud infrastructure teams as well as DevOps, data science, and LOB decision-makers is required to coordinate infrastructure buy versus build decisions. Getting alignment on critical areas such as cybersecurity and resiliency, data volumes and data integration, infrastructure costs, vendor support, and hybrid cloud and multicloud interoperability is vital to making optimal AI-ready infrastructure investment decisions. In IDC's 2023 *Worldwide Generative AI Study,* it was found that 52% of GenAI spending would go to infrastructure (29% dedicated, 23% public cloud) with the remainder relatively evenly distributed between external AI platforms, internal employees, and services firms, demonstrating how large and critical the investment in infrastructure is from a decision-making perspective.

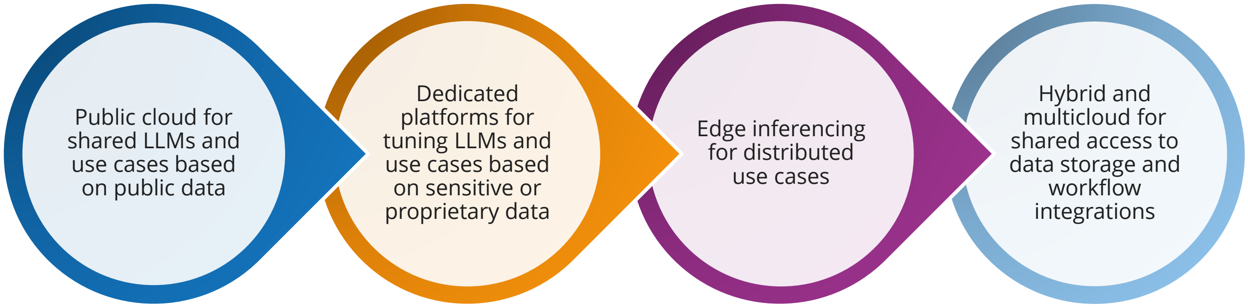
There are pros and cons to each option in the build versus buy discussion:

- **Building your own infrastructure from the ground up:** Selecting, deploying, integrating, and customizing individual infrastructure components in a best-of-breed or fit-for-purpose manner may be an option for companies with extensive in-house expertise, unconstrained budgets, or overwhelming reasons to protect their own IP by avoiding any external access to their custom data and processes. The many dependencies on different parts of the hardware stack, however, can make this highly technical exercise difficult, time-consuming, and expensive. Surprisingly, 44% of respondents surveyed in IDC's 2023 *AI View Study* indicated that networking spend was their largest line item when it came to performing GenAI model training from scratch in their own datacenter.

- **Existing public cloud AI services:** For most organizations, this is the fastest way to explore AI and GenAI initiatives, and it can be extremely flexible and responsive to changing needs through the early stages of transformation. However, the risks to custom data and company IP can make it a nonstarter for some production workloads and use cases, and buyers risk vendor lock-in, which can both be costly and stifle innovation. Getting data into and out of cloud service providers' infrastructure can also introduce additional costs into the transformation process.
- **Full-stack solution:** Deploying a validated full-stack solution on premises, in colocation or at the edge, can be a shortcut to many of the design decisions in the early stages of AI transformation. If deployed in an on-demand or as-a-service manner, enterprises can achieve similar financial flexibility to cloud deployments, with the security, compliance, privacy, and governance profile of traditional infrastructure.
- **Hybrid approaches:** Varying use cases and different stages of the AI and GenAI data pipeline can lend themselves to a hybrid approach in which a variety of deployment modalities are used with unified management and observability tools across the organization to simplify management and operations within the AI workstreams. This mix of on-premises, edge, and cloud resources in hybrid cloud and hybrid multicloud approaches is expected to be the dominant mode of deploying AI and GenAI workloads across the majority of mature enterprises (see Figure 1).

FIGURE 1

Hybrid Use Cases


Public cloud for shared LLMs and use cases based on public data
Dedicated platforms for tuning LLMs and use cases based on sensitive or proprietary data
Edge inferencing for distributed use cases
Hybrid and multicloud for shared access to data storage and workflow integrations

Source: IDC, 2024

As approaches to AI and GenAI modeling, tuning, and inferencing continue to become better understood, smaller data sets and models are being found to be capable – and sometimes even optimal – in a variety of situations. Small footprint-integrated infrastructure deployments that can deliver business value using traditional CPUs and smaller sets of parameters and data are garnering consideration by 85% of organizations worldwide. The purchase price of graphics processing unit (GPU)-intensive systems, not to mention the operating costs, power and cooling, and idle costs of these power-hungry and expensive systems, can be daunting.

Lightweight enterprise AI and GenAI models, known as small language models (SLMs), are trained on relatively small datasets (often in the millions to low billions in parameters, as opposed to the hundreds of billions to trillion plus parameters of current LLMs) require less infrastructure to support, and provide less complexity in terms of design and deployment. In extreme cases, they can be operated on mobile devices, but they can provide a powerful complement to LLMs for use cases with limited knowledge requirements, especially in small footprint deployments, including edge and IoT. Narrowly focused vertical and horizontal use cases are ideal for SLM deployments, where industry-specific or functional needs can be met by effective, efficient infrastructure.

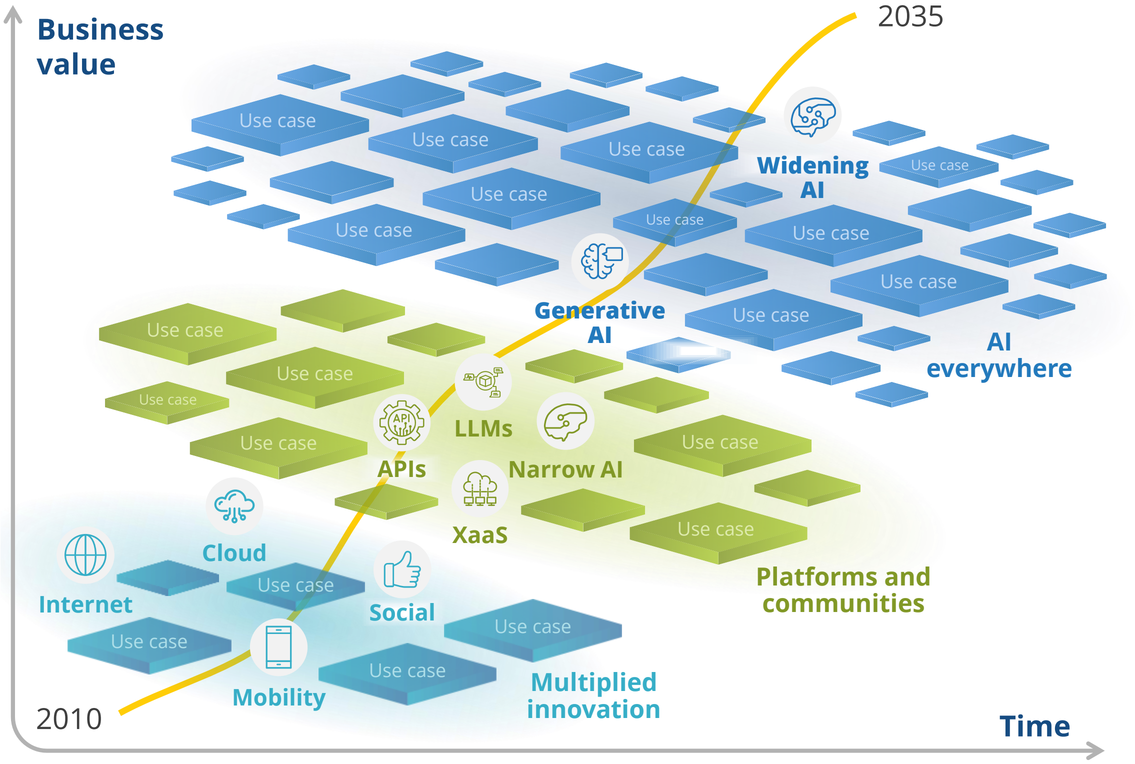
Organizations in the many stages of AI maturity or inappropriate for use cases that can be satisfied by more efficient and cost-effective hardware, and, as such, full-stack "easy button" infrastructure with CPU acceleration rather than GPU and smaller datacenter space, power, and cooling impacts can be extremely attractive to organizations at various points on the AI maturity curve, on a use case-by-use case basis.

FUTURE OUTLOOK

AI is not an entirely new discipline – the earliest research and initiatives in the field occurred nearly three-quarters of a century ago. However, the explosion of capabilities in machine learning, high-performance computing, and predictive, interpretive, and generative AI in the past few years has been groundbreaking in its growth and future potential (see Figure 2). IDC expects AI to reach into every aspect of our connected life, not simply through AI "apps" and LLMs but through AI augmentation of existing applications and services, often undetectable to end users except by improving user experiences and quality of services.

FIGURE 2

Widening AI Growth



Source: IDC, 2024

AI spending growth will reach $521 billion by 2027 at a compound annual growth rate (CAGR) of 31.4%. AI-ready infrastructure is roughly one-third of that number. Further, spending on GenAI infrastructure will see a five-year CAGR of 71.1% through 2027, leading to a total spend of $48.5 billion. As with many transformational technologies, there is a progression from early experimentation to aggressive buildout with targeted implementation to widespread adoption with extensions to the edge of all business activities.

CONSIDERING NUTANIX, CISCO, AND INTEL FOR FULL-STACK AI-READY INFRASTRUCTURE

A joint solution from Cisco and Nutanix, powered by Intel, delivers a turnkey platform designed to simplify AI deployments and secure and manage AI workloads without costly and high-power GPU acceleration and licensing. Nutanix Cloud Platform enables GenAI workloads and data services with security built in. Cisco UCS provides infrastructure and SaaS management through Intersight for core to edge to cloud management across small and large enterprises. Intel AMX is a built-in AI accelerator in every Generation 4 and later Intel Xeon processor, supporting inference on generative AI, as well as machine learning and other predictive and interpretive AI workloads without the need for GPUs. This approach saves capital and operational costs as well as power and cooling for a range of AI workloads.

Nutanix, Cisco, and Intel Partnership Value Proposition

Nutanix Cloud Platform provides a virtualized infrastructure platform designed to allow customers to run applications and access data anywhere (including AI and GenAI applications and data) from the core to the edge to public cloud using an identical operating model, creating an operational standard fit to the demands of a large proportion of businesses considering hybrid cloud and hybrid multicloud deployments.

Cisco UCS brings together compute, networking, and storage in a single system to power your applications, including AI workloads. The full portfolio of Cisco servers, SaaS management, and Nutanix Cloud Platform capabilities plus the latest processor, accelerator, network, and drive technologies offer customers a flexible, integrated hyperconverged platform to modernize their IT environments.

Cisco Compute Hyperconverged with Nutanix, powered by Intel, simplifies and accelerates the delivery of infrastructure and applications at a global scale through flexible cloud operating models and enhanced support and resiliency capabilities. Cisco, Intel, and Nutanix have partnered to deliver a hyperconverged solution for AI through expanded engineering, support, and go-to-market collaboration to provide a more seamless experience, foster innovation, and accelerate customers' hybrid multicloud journeys.

The Cisco Compute Hyperconverged with Nutanix GPT-in-a-Box solution aims to remove complexity from the adoption of generative AI by providing prescriptive steps for deploying the underlying infrastructure for Nutanix GPT-in-a-Box. This solution combines Cisco servers and SaaS operations with Nutanix software testing a number of the most popular LLMs to produce a fully validated AI-ready platform to speed up AI initiatives from the datacenter to the edge.

For appropriate use cases, AI can be simplified and be made more efficient with a turnkey platform to enable an identical operating model for AI workloads anywhere. Nutanix's GPT-in-a-Box on Nutanix Cloud Platform was built to address this need — to build, deploy, and run AI workloads with the same considerations, requirements, and management capabilities in any deployment location.

CPUs Versus GPUs and Intel AMX

Currently, GPUs are seen as the primary method to help crunch data from LLMs to an AI app (e.g., interfacing with ChatGPT or other bots or assistants). However, LLMs vary in size and capabilities, and many edge or distributed systems can function using CPUs alone through special AI extensions from Intel called Advanced Matrix Extensions (AMX). Intel AMX are new, built-in CPU accelerators designed to improve performance for AI workloads including natural language processing, recommendation systems, and image recognition as well as inference activities on smaller, industry- or domain-specific LLMs and models trained on private data.

Nutanix Cloud Platform software leverages secure Cisco UCS servers with Intel Xeon processors and built-in AMX accelerators and/or GPUs. Together, these technologies enable systems that can run AI as just another application anywhere, even across public and private cloud resources, in a consistent way. For use cases that don't necessarily require massive amounts of computational ability provided by way of GPUs, this can reduce both capital expenditures — on infrastructure — and operational costs — on proprietary software licensing, power, and cooling costs.

CONSIDERATIONS

The top 3 inhibitors to GenAI adoption, according to IDC's 2024 *FERS Survey, Wave 1,* are:

- Loss of control over data and IP (30%)
- Excessive GenAI application development and add-on costs (26%)
- Excessive infrastructure costs for models and training (25%)

On-premises as-a-service infrastructure from trusted partners can alleviate these top concerns, both by establishing the control needed over data to satisfy security, data protection, privacy, regulatory, and governance requirements and by tying costs to utilization — effectively making that infrastructure a component of cost of goods sold rather than a fixed capex that doesn't necessarily reflect value-added activities. While there may be a premium paid for on-demand infrastructure, the ability to match costs to demand and utilization can mitigate that discrepancy.

In addition, enterprises need to deal with the three legs of the tripod that support any transformative undertaking, namely:

- **People — Both new and emerging skills and available resources.** Trusted partners and infrastructure that simplify implementation and operations, rather than adding complexity and silos, can help alleviate strain on overtaxed resources. Validated designs lend themselves to shared learnings and IP, unlike one-off, build-your-own implementations that external partners and service providers will have little or no experience with.
- **Process — Operational complexity within business units that can stifle innovation.** Collaboration between various AI stakeholders and personas within the organization based on a common understanding of business goals is critical — buy-in from the C-suite downward can ensure that teams are on the same page.

- **Technology — Appropriate deployment choices for AI workloads and an understanding that implementations are rarely "one and done."** Technology and business requirements are fluid, and an infrastructure that not only provides the ability to manage and monitor resources and workflows across the enterprise but enables application and data mobility with the same operational characteristics and management can provide the flexibility and ability to pivot based on business requirements that can provide real ROI on AI transformation initiatives.

CONCLUSION

Organizations trying to navigate the uncertain waters of AI transformation need to prioritize GenAI use cases according to business value, cost, and potential business risk and think through individual/team productivity and functional and industry-specific use cases as part of a holistic strategy. Choosing which use cases are best served by different deployment methodologies is daunting, which is why simple, powerful, flexible infrastructure that enables applications to run anywhere and data to provide value in any context can be an enabler of business value in AI and GenAI initiatives.

Most organizations expect to deploy AI workloads across a hybrid mix of cloud, edge, and dedicated infrastructure based on workload-specific requirements, including security, performance, and cost considerations. Deploying AI and GenAI workloads in an efficient manner is key, especially for organizations that cannot simply procure and utilize expensive-to-buy, expensive-to-operate GPU-accelerated compute resources. Appropriate technology, including full-stack solutions and smaller footprint models, appeals to users for use cases that don't require GPU acceleration.

In addition, IDC suggests that users ensure that they lay the AI and GenAI technology foundations with:

- A data-centric platform underpinning the enterprise
- Cost-effective and appropriate digital infrastructure for AI workloads
- An approach toward different deployment modalities on a case-by-case basis to maximize return on AI investments and get access to the best technologies, capabilities, and operating paradigms for your organization
- Infrastructure decisions that allow for, and a willingness to adapt to, the changing business and technical requirements for AI and GenAI workloads

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology, IT benchmarking and sourcing, and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives. Founded in 1964, IDC is a wholly owned subsidiary of International Data Group (IDG, Inc.).

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com