

The complexity of enterprise AI/GenAI initiatives demands a full-stack solution, prioritizing use cases and optimizing infrastructure, making a full-stack vendor essential for a successful AI journey.

Unlocking AI's Potential: Navigating Infrastructure Challenges for Next-Gen Solutions

[September 2024]

Written by: Deepika Giri, Associate Vice President; Franco Chiam, Vice President

Introduction to AI and GenAI: Adoption Waves and Use Case Prioritization

Generative AI (GenAI) took the world by storm in 2023, with about 56.5% of WW organizations having said they are currently using GenAI (Source: *WW Generative AI Survey 2023*). By the end of 2024, 82.7% of WW organizations would have already deployed GenAI applications (Source: *GenAI Pricing Model Survey 2024*). It's important to note that traditional AI applications utilizing classic predictive or clustering models are even more deeply embedded in enterprise IT systems than GenAI. Here are a few additional noteworthy insights:

- » Organizations are augmenting their investment plans on advanced GenAI applications. An IDC survey shows that 33.3% of WW organizations plan to invest in business functional use cases and 31.3% in industry use cases in the next 18 months (Source: *FERS Wave 2*).
- » Organizations are leveraging GenAI not only to boost internal productivity but also to drive business model innovation, such as generating revenue, drug discovery, and insurance risk modeling (Source: *IDC Global GenAI Technology Trends Survey 2024*).

These shifts indicate that organizations are moving beyond experimenting with AI and GenAI in controlled, siloed environments. They are now embedding AI into core business models, adding complexity to systems integration across MLOps, DevOps, and multiple lines of business. The mantra "start small, scale fast" has evolved — now is the time to design scalable, enterprise-grade AI systems.

AT A GLANCE

KEY STATS

- IDC 2024 research indicates the no. 1 concern of 50% of WW organizations is the lack of data management and optimization, and skilled workforce.
- For 30% of respondents, excessive costs prevent them from meeting their ROI objectives.
- A strong AI infrastructure is crucial for businesses to efficiently implement AI.

WHAT'S IMPORTANT

- Modular systems provide flexibility and scalability for AI, allowing easy integration and performance optimization without overhauls.

Prioritizing Use Cases for Maximum Impact

Over the past two years, many organizations have gained experience in implementing point solutions enabled by AI/GenAI. However, they struggle when faced with the complexity of managing multiple AI/GenAI applications in production. Additionally, they now face challenges posed by expanding AI/GenAI/RAG pipelines to include core business data and systems, and optimization of cost and security issues arising from these expanding AI systems.

- » **Use case prioritization:** IDC's path to impact from GenAI provides a framework to realize value from GenAI through use cases. It classifies GenAI use cases into three types: productivity, business function, and industry use cases. In the realm of GenAI across business functions, IDC has catalogued at least 255 use cases with distinctive business impact, risk, and complexity levels. Each of these business function use cases can be mapped into AI development life cycle of ingest, train, tune, infer and run. Furthermore, the model development life cycle will need to accommodate data in diverse formats such as texts, images, videos, logs, and so forth. It is important for organizations to prioritize AI/GenAI use cases to maximize value delivered, bearing in mind the ROI. Use cases must be assessed for rich business outcomes and impact on KPIs such as worker productivity, revenue generation, or customer satisfaction. In the longer term, the use cases must help build organizational resilience and help drive adaptability, innovation and sustainable growth across the board.
- » **Model choice:** Many organizations are evaluating different options for AI models. While 59.5% of WW organizations want to use open-source models, the remaining 40.5% consider using commercial models. The approaches to using a model differ; 50.2% would use existing models as is, without any tuning, while 43.1% are fine-tuning an existing model, and 30.7% are even training their own GenAI models from scratch. Furthermore, 24.6% of WW organizations have already deployed small models for better model performance (Source: *IDC Global GenAI Technology Trends Survey 2024*). It is apparent that there is no preference to a specific approach; when evaluating the 'build or buy' decision for AI models, organizations are most likely to mix depending on the applications, ranging from commercial models to custom models fine-tuned from open-source options. The approaches are dictated primarily by the use cases, as the model specifications vary accordingly.

An effective AI strategy must prioritize use cases, establish AI governance as a prerequisite, and streamline the model selection process to maximize business impact in a cost-efficient manner.

Infrastructure Reliance for Enterprise GenAI/AI Adoption

The disruptive shift to AI infrastructure reliance is a critical factor that organizations must navigate.

The adoption of AI use cases in the enterprise landscape revolves around the choices of technologies that must be guided by the needs of the end workload. The considerations must involve several critical elements inclusive of scalability, performance, integration, and data management. These considerations must be applied and adapted according to needs, whether this is on private, public, on-premises, core or edge.

IDC's recent survey (Source: *IDC Syndicated Survey 2024: AP State of Datacenter Infrastructure Survey*) indicates that 30% of organizations that own and/or operate datacenters(co-locations) are in the initial planning stage with regard to AI readiness.

Infrastructure plays a crucial role in enabling GenAI. This is the technological foundation that supports the development, deployment, and foundation of advanced AI systems. However, with the constant changes, the question everyone has in mind is: build or buy?

The answer is not so simple; it comes down to the organization strategies. For example, larger organizations that leverage AI for game-changing strategies will build a private AI infrastructure, retaining more control while leveraging open integration for scaling.

Small and medium-sized enterprises (SMEs) or digital-native businesses (DNBs) that do not have a broader AI strategy or the necessary skills and means will leverage existing models and tune them to their own data to create specific models for their needs. The complexity of the spectrum of needs and existing tech and skills is broad.

The successful adoption of generative AI and AI technologies within enterprises demands a strategic approach to infrastructure investment and management. Organizations must ensure their data management systems, computational resources, network capabilities, security protocols, and integration strategies are aligned to support the effective deployment and utilization of AI technologies.

Challenges in AI Infrastructure

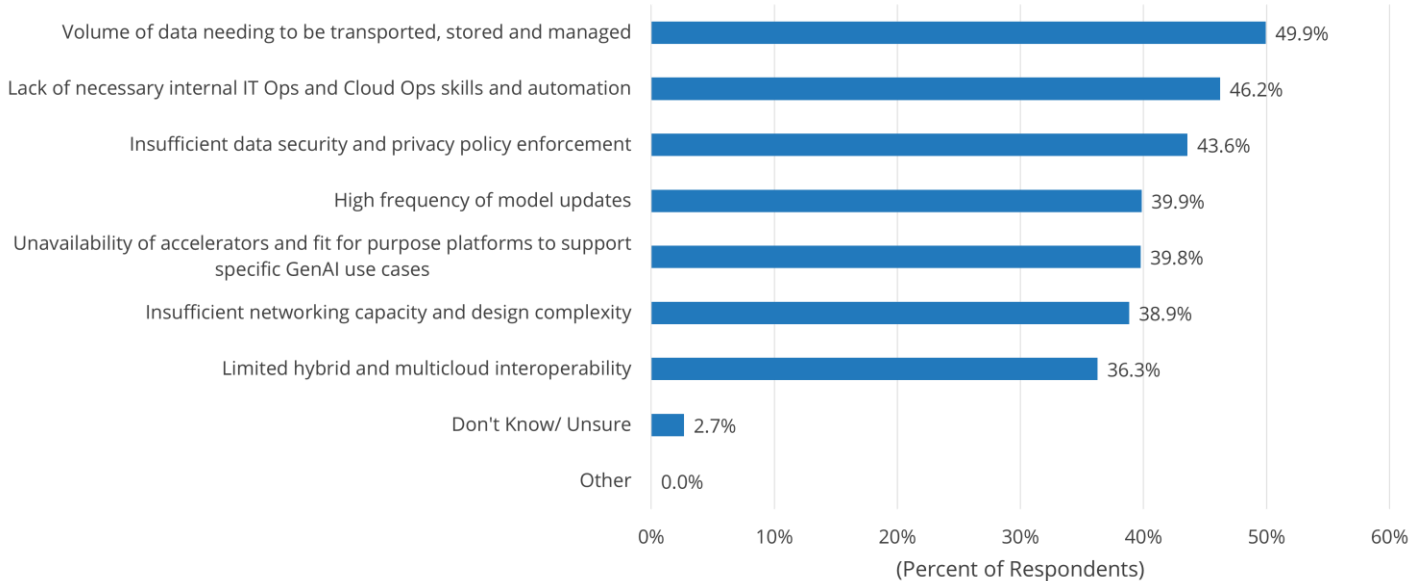
When deploying AI Infrastructure, enterprises should be guided by the required balance among compute, memory, storage, and networking technologies necessitated by their major workloads.

GenAI language learning models and software require much higher processing compute capability, which depends on specially designed hardware. The availability of accelerators to support the desired performance is one of the biggest concerns for enterprises prioritizing their digital infrastructure.

IDC's *FERS Wave 1 2024* indicates that the no. 1 top concern of close to 50% of organizations is the lack of data management and optimization, as well as skilled workforce to support the changes.

FIGURE 1: **Greatest Challenges to Successful Implementation of GenAI Use Cases**

Q. What operational concerns related to digital infrastructure will pose the greatest challenges to successful implementation of your organization's highest priority GenAI use cases in the next 18 months?



Source: IDC's *Future of Enterprise Resiliency and Spending Survey, 2024, Wave 1 (n=881)*

Data management and model updates are big tasks that enterprises need to manage. 46% of organizations found this to be one of the biggest challenges other than skillsets. This requires substantial effort to mobilize large data sets to and from language models, which can lead to issues on data security and management. When it comes to maximizing the value of AI/ML initiatives, IDC's research identifies "issues of data availability and data quality" as one of the most significant challenges cited by respondents worldwide. A strong data infrastructure underpins AI maturity by providing the foundation necessary for effective data management,

which directly impacts the success of AI projects and the realization of significant business benefits. Without it, organizations may struggle with inefficiencies, inaccuracies, and limitations in their AI efforts.

Although most enterprises adopt a hybrid multicloud approach, there is still limited interoperability between different environments. Moving data from one cloud to large language models (LLMs) hosted on another will have cost, performance, and migration challenges.

Implementing GenAI: What Does It Take to Do This?

Importance of Data Platform for GenAI Implementation: Getting Your Foundation Right

Despite the widespread adoption of GenAI by enterprises, a significant number of GenAI projects do not advance to the production stage. Globally, 44.5% of organizations experience a success rate of less than 50% in transitioning their GenAI projects to production.

TABLE 1: **Why Organizations Struggle to Implement GenAI Projects in Production**

Reasons for Failing to Bring GenAI to Production	Responses
Excessive costs prevented meeting ROI objectives	39.80%
Lack of developers with required skills/tools	38.20%
Inadequate infrastructure performance/availability	36.90%
Inability to access required data sets	33.70%
Tech/Services partners didn't meet project specifications	31.80%
Misalignment of use case scope/requirements	31.10%
Poor coordination between IT and line of business teams	28.20%
Poor quality/poorly labeled data sets	27.60%
Unacceptable bias and/or confabulation	26.80%

Source: IDC Global GenAI Technology Trends Survey 2024 n=624

The increasing number of GenAI applications brought an unexpected cost increase from using clouds and cloud-based AI models. While the token prices of famous foundation models have been dramatically decreasing, the cost complexity of implementing GenAI workloads over production infrastructure is increasing, as it is not just about the token prices but also involves calculation of all kinds of infrastructure options to cover — from data storage and network to computing resources.

More GenAI applications mean more data integration within the organization, and this need for data integration reveals even more issues in enterprise data management. If GenAI applications cannot consume enterprise data due to data accessibility, data quality, and data label issues, they cannot transform core business functions of any enterprise or industry.

Human factors, such as skills gap and communications, to name a few, are also critical considerations in preventing GenAI from being implemented into production. As the recruitment of AI talent is getting more difficult, organizations need to think of better project management approaches. However, they also need to find a way to convert human issues associated with GenAI deployment into technical solutions.

Building a Robust Infrastructure

An AI infrastructure comprises not just the capabilities of GPUs or CPUs, but also the key components of networking and software, which are just as essential for developing, deploying, and managing AI.

Building a robust AI infrastructure ecosystem needs to start by replacing the rigid, siloed approach of traditional infrastructure design. Composable Infrastructure with a dynamic pool of shared resources from compute power, GPUs, storage, and networking capabilities will help with seamless resource sharing, and microservices can be utilized to enhance flexibility. Prioritizing a centralized data management system will also help ensure real-time insights and improve collaborations across departments.

An IDC Survey (Source: *IDC's Future Enterprise Resiliency and Spending Survey, 2024, Wave 4*) found that inadequate infrastructure and high costs are the top hindrances in GenAI implementations. Over 30% of respondents said these excessive costs are preventing them from meeting ROI objectives.

Data Management and AI

Data is crucial for AI and GenAI projects, especially for organizations developing or fine-tuning their own models. GenAI models handle various types of data, so organizations must integrate structured, semistructured, and unstructured data into their systems. To support this, companies need to build data pipelines as part of the retrieval-augmented generation (RAG) architecture. According to an IDC survey, 40% of organizations using RAG relied on unstructured data, 34.8% used structured data, and 25.3% used semistructured data. Even those using commercial language models need to enhance their data pipelines to manage different data types. For those creating their own models, managing unstructured data adds complexity to existing MLOps requirements.

Role of Infrastructure in AI Implementation

The disruptive shift of AI infrastructure will help enterprises accelerate deployment with new AI-ready hardware and software infrastructures. Organizations are gearing up to drive meaningful gains in business and productivity and reimagine customer digital experiences. Effective implementations of AI adoptions will rely and hinge on key infrastructure components.

Take for example, in terms of pure computing, the computation requirements of core systems differ markedly from those of edge systems. While both core and edge are essential for AI applications, their functions and environments necessitate distinct approaches. On one hand, core systems are engineered for high-performance tasks, such as complex training models, therefore they rely on more powerful CPUs and GPUs. On the other hand, edge systems prioritize efficiency due to limited resources, favoring low-power CPUs or specialized AI silicon for inferencing operations.

A strong AI infrastructure is crucial for organizations to efficiently implement AI.

The growing availability of cloud AI platforms presents a compelling alternative for organizations that are hesitant to invest heavily in AI infrastructure. These infrastructure platforms offer scalable solutions without the need for substantial upfront capital expenditure, making them particularly a definite choice for businesses looking to leverage AI technology without significant financial commitment.

Furthermore, cloud AI platforms provide flexibility and most hyperscalers are now providing access to cutting-edge specialized CPU and GPU cores for AI workloads. This can be advantageous for organizations that lack the internal infrastructure to support large-scale

AI projects. This approach will allow companies to integrate advanced AI capabilities into their operations while managing costs most effectively.

Infrastructure Requirements for AI

To date, most infrastructure setups have been directed toward integrating AI within current systems architectures, which include distributed computing, the deployment of data processing, movement, and storage technologies accordingly to where the data is residing. Alongside this is a similar approach for heterogeneous computing, the mix and match of technologies according to the needs of the intended workloads, which aims to enhance AI capabilities within an existing infrastructure architecture.

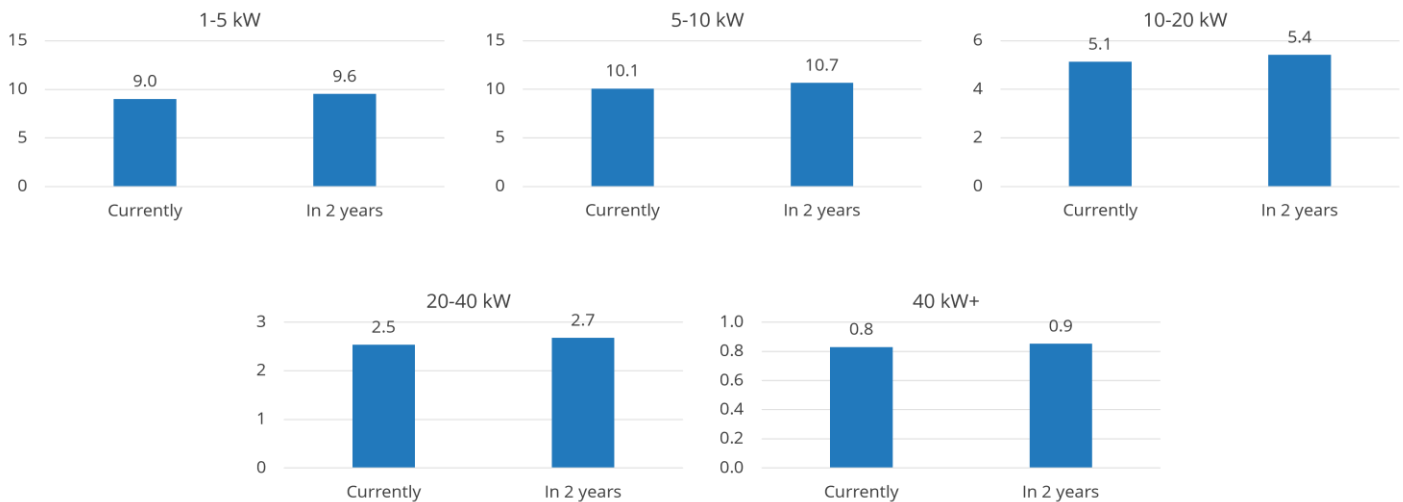
Processing technologies have been aimed at enhancing data throughput directly at the source, supported by specialized memory technologies designed to minimize data latency for processors. Traditionally, processors, memory, storage systems, and networking technologies' have been function-specific, with their resources typically not shared or synchronized.

Upcoming components will refocus and re-architect infrastructure systems to support synchronized and pooled data processing, storage and movement. These changes will allow and enable more application-specific system architectures such as AI. Data can be processed at core, edge, or hybrid cloud environment in an integrated, modular ecosystem.

Sequentially, findings from IDC's *Datacenter and Sustainability Survey* indicate that datacenter operators of all types (enterprise and service providers) expect to increase the power densities by 2026.

FIGURE 2: **Rack Power Density Currently and in Two Years**

Q. What are and will be the rack densities of the Currently deployed and the planned in 2 years?



Source: *Datacenter Operations and Sustainable Survey*, IDC, March 2024 (n=766)

This will also shift datacenters to adopt advanced cooling techniques such as liquid cooling to handle higher power density.

What Is Needed?

Organizations must scrap siloed technical designs and work with various stakeholders from business leaders to IT teams to address both technical and organizational challenges effectively. An established and well-defined AI infrastructure needs constructive shared platforms and tools that allow cross-functional teams to contribute to each layer. Business stakeholders need to establish clear

communication on strategic requirements that are converted to measurable desired outcomes for AI projects. They also play key roles in prioritizing initiatives to ensure resources are pooled toward the most impactful projects.

This collaboration will help IT teams to ensure that infrastructure is planned and optimized across resources; not just from CPUs, GPUs, or TPUs but also across data repositories, distribution, and access within relevant compute systems. These are essential for supporting AI pipelines and generating actionable results for downstream applications. The data infrastructure itself must possess a certain level of intelligence to effectively handle management tasks, diverse formats, capacity issues, security, and compliance challenges. The right tools and platforms will optimize high-bandwidth, providing low-latency networking infrastructure for auto-scaling, prioritizing AI workloads to ensure resources can scale dynamically based on model training or inference needs.

Organizations should opt for modular software and hardware solutions to address their diverse AI needs. Unlike monolithic systems, which are hard to scale and adapt, modular systems use well-defined input and output interfaces, making them more flexible. Modular components allow organizations to build and adjust AI systems more easily, integrate new technologies and optimize performance without needing to overhaul entire infrastructures.

Conclusion

As organizations seek to enhance their AI/GenAI capabilities, they must navigate the entire enterprise AI technological stack, from applications to infrastructure. However, the complexity of integrating all AI technical layers with business functions often hinders the development of enterprise-grade AI systems. In particular, strong competency in handling AI infrastructure is crucial for efficiently implementing AI and GenAI workloads cost-effectively. Also, the AI skills gap is most pronounced in AI infrastructure operations. To prepare for future AI readiness, organizations should consider working with a full-stack AI solution provider that can manage not only machine learning application life cycles but also the associated infrastructure complexities concerning private, public, on-premises, core or edge AI.

Navigating enterprise AI stacks requires addressing infrastructure and application

About the Analyst

	<p><i>Deepika Giri, Associate Vice President</i></p>
	<p><i>Franco Chiam, Vice President</i></p>

Deepika manages and leads the research programs in big data and analytics (BDA), artificial intelligence (AI), blockchain, and Web3 across Asia/Pacific. She has extensive experience in software delivery as well as sales leadership and management. She also has over 20 years of experience in IT services, including leadership roles, at Capgemini, Infosys, and Accenture, and has strong industry expertise in the telecommunications and retail industries.

Franco Chiam is the vice president for IDC's Asia/Pacific (excluding Japan) Cloud, Datacenter, Telecommunication, and Infrastructure Research Group. He manages and shapes the aforementioned domains' offerings to IDC clients, which include cloud and infrastructure surveys, market analysis and perspective, speaking engagements, and executive briefings.

MESSAGE FROM THE SPONSOR

HCLTech and NetApp have partnered for over a decade, delivering innovative solutions that empower enterprises to tackle today's digital challenges. Together, they optimize hybrid and multicloud environments with cutting-edge data services and AI-ready infrastructure, ensuring seamless integration across platforms such as Azure, Google Cloud, and AWS. Their joint solutions drive continuous availability, robust AI workloads, and accelerated innovation to fuel business growth. With a shared vision for the future, HCLTech and NetApp empower organizations to meet the evolving demands of the digital age and stay ahead in an AI-driven world.

- » Visit HCLTech to learn more: <https://www.hcltech.com/about-us/alliances/netapp>
- » Visit NetApp to learn more: <https://www.netapp.com/partners/partner-connect/hcl-tech/>



The content in this paper was adapted from existing IDC research published on www.idc.com.

IDC Research, Inc.
140 Kendrick Street
Building B
Needham, MA 02494, USA
T 508.872.8200
F 508.935.4015
Twitter @IDC
blogs.idc.com
www.idc.com

This publication was produced by IDC Custom Solutions. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis independently conducted and published by IDC, unless specific vendor sponsorship is noted. IDC Custom Solutions makes IDC content available in a wide range of formats for distribution by various companies. A license to distribute IDC content does not imply endorsement of or opinion about the licensee.

External Publication of IDC Information and Data — Any IDC information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the appropriate IDC Vice President or Country Manager. A draft of the proposed document should accompany any such request. IDC reserves the right to deny approval of external usage for any reason.

Copyright 2024 IDC. Reproduction without written permission is completely forbidden.