IDC

SPOTLIGHT

Sponsored by: AMD

Successful deployment of AI requires scalable infrastructure, robust data management, and alignment between business and technical teams. Businesses must focus on a robust tech stack that underpins a flexible, hybrid infrastructure strategy to support diverse AI use cases.

Playing the Long Game with Enterprise AI Insights on Overcoming Infrastructure Challenges to Scale AI

April 2025

Written by: Ashish Nadkarni, Group VP and General Manager, Worldwide Systems Infrastructure Research

AT A GLANCE

KEY TAKEAWAYS

Businesses face significant challenges in transitioning AI proof-of-concept projects to large-scale production, primarily because of the need for scalable infrastructure, data security, and AI-specific skills.

For future AI infrastructure investments to be successful and deliver a consistent and measurable ROI, business leaders must emphasize the importance of the following:

- Standardized, flexible, and hybrid infrastructure strategies, strong alignment between business and technical teams, and robust data governance
- Standardized platforms, improved data management, strategic provider partners, and cross-functional governance

Introduction

Businesses have been leveraging AI for nearly a decade, evolving from machine learning to deep neural networks to now generative AI (GenAI), which creates novel content. This shift has led to diverse applications in sectors such as fraud analysis, healthcare, and software development, enhancing productivity and efficiency. Future advancements, such as agentic AI and physical AI, promise further improvements in human productivity and sophisticated human-machine interactions.

A critical oversight by business leaders is the exclusion of IT decision-makers (ITDMs) early in AI initiatives, leading to misjudgments about necessary IT infrastructure. To avoid this, IT leaders should be involved from the start. Establishing a center for excellence, comprising business leaders and technical experts, can bridge this gap to ensure realistic outcomes and ROI calculations. The tech stack for AI initiatives – as they evolve from the proof-of-concept (POC) phase to production – requires considerations for computing power, storage, networking, and security. Compliance with regulatory frameworks such as GDPR is essential, and security must be integrated from the silicon level up.

Taking AI mainstream involves aligning models with the necessary technology, procuring appropriate systems, and ensuring that the business has the required skill sets. Partnering with vendors can help provide preconfigured infrastructure, minimizing obstacles. Effective communication and collaboration between business units and IT are crucial for a successful AI integration.

Rapid implementation of AI is vital for immediate business value, enhancing decision-making, streamlining operations, and gaining a competitive edge. Clearly defining use cases, calculating requirements, and collaborating with vendors are

essential steps for optimizing AI infrastructure for performance and scalability. Proper planning and vendor collaboration are key to overcoming the complexity of AI projects and ensuring successful implementation.

Playing the Long Game with AI

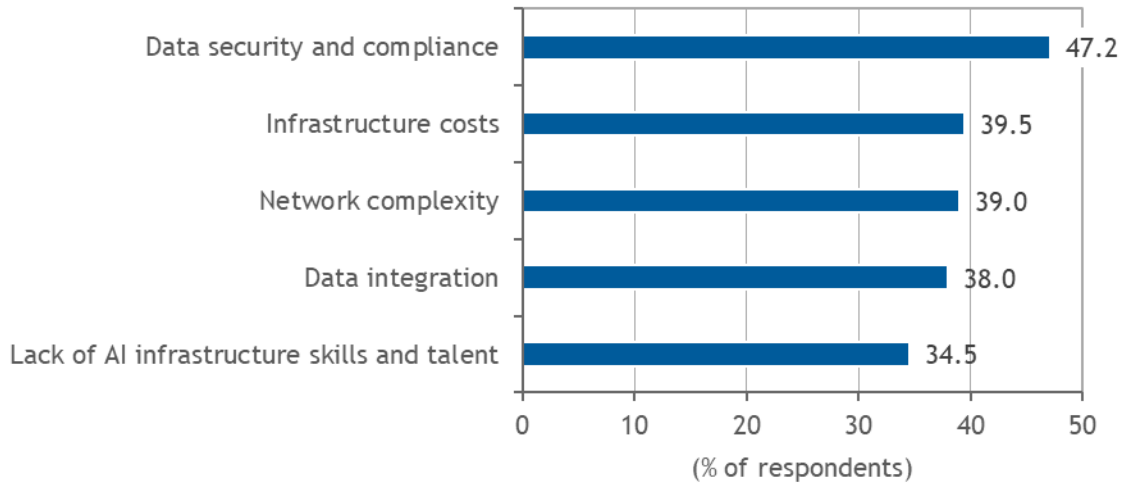
In 2024, IDC conducted a study to assess the experiences of IT decision-makers, including tech buyers, at 600 businesses that have successfully moved AI POC trials into large-scale production usage over the past year. Most have implemented a mix of different AI use cases representing generative AI, predictive AI, and interpretive AI. IDC estimates that these businesses are representative of the top 30% of businesses in terms of AI-ready infrastructure maturity.

Many businesses invested in POC projects to expand their use of AI in the enterprise, but IDC found that only a few were able to move significant numbers of these test cases into production deployments that deliver real value to their business. Moving from POC to production depends on scalable infrastructure — computing, storage, and networking — that can handle diverse AI workloads. Without a scalable and well-planned digital infrastructure tech stack, even the best AI use case will be unable to fully unlock the value of mission-critical data and consistently return high-quality results to customers and employees.

The learning curve related to matching different types of AI use cases to the most appropriate infrastructure resources can be steep. Requirements can change rapidly as more capable models and computational capabilities become available but are not necessarily matched to business needs. Businesses indicate that they expect many AI use cases will need to support thousands of concurrent users and will often rely on data coming from a dozen or more sources. Businesses that have implemented AI in production have learned that the ability to successfully move AI use cases from POC to production requires them to overcome a number of potential inhibitors, including data security and compliance, infrastructure costs, network complexity, data integration, and AI-related skills and talent (see Figure 1).

FIGURE 1: ***The Top 5 Factors Inhibiting a Full Rollout of Required AI Infrastructure***

Q What factors are most likely to slow down or inhibit the full rollout of required AI infrastructure between now and the end of 2026?



Source: AI-Ready Infrastructure Tech Buyer Adoption Trends Research Highlights Demand for Scalable Infrastructure Platforms and Alignment with Business ROI (IDC #US52101825, December 2024)

ITDMs indicate that as they develop a deeper understanding of the capabilities and business benefits provided by AI, especially as it becomes embedded within their business application landscape, they have come to appreciate how important it is to develop frameworks for optimizing infrastructure environments to flexibly support many different types of AI use cases and rapidly changing AI models. Part of this process calls for technical teams and business leaders to align on a shared understanding of the expected business benefits delivered by the specific AI use case and to weigh architectural and operational trade-offs related to speed, cost, security, and quality of AI results when making infrastructure investment decisions.

Specific to GenAI, these industry leaders expect that over the next two years, about 45% of their GenAI use cases will use small language models that have been customized to support targeted business needs, often using internal company data. They recognize that not every GenAI use case may require the highest performing GPUs available and need to be built by ingesting all possible content available in the public domain. They also understand that many use cases will rely on standard AI engines and patterns, such as chatbots or content summarization, and they are looking for ways to avoid duplication of these standard services.

Often, these decision-makers find that they do not need to custom build models from scratch. Rather, successful businesses are often investing in a range of options to customize and tune off-the-shelf commercial or open source models. They may use retrieval-augmented generation (RAG) techniques to bring internal data to a model to help refine inferencing outputs, or they may fine-tune the model using internal data. They are also evaluating ways to streamline model instruction and improve prompting. Many businesses are breaking AI use cases into multiple tasks and using orchestration to link different types of models to support different elements of more complex and automated workflows. The rise of agentic AI orchestration is expected to accelerate the use of these approaches.

The more internal data that is consumed by AI models, the more businesses recognize how critical it is to have clean, well-managed data workflows across their businesses. Many businesses struggle to overcome data silos and design AI infrastructure platforms that can support frequent model refreshes and replacements without reducing AI workload performance or compromising data security, confidentiality and privacy.

The ability to cost-effectively support thousands of simultaneous human and machine users requires businesses to rethink many traditional assumptions about digital infrastructure architectures and operating models. Specifically, ITDMs indicate that they are considering how to better optimize the use of expensive accelerated computing resources (when needed) and optimized general-purpose computing resources (most commonly used), along with approaches to networking, data security, and storage. They are beginning to explore opportunities related to other types of computing platforms, accelerated computing platforms, virtualization and containerization, high bandwidth and low-latency networking, and intelligent client devices (e.g., AI PCs, AI workstations, and edge devices) to cost-effectively extend access across wider ranges of locations, devices, and user groups.

How CIOs and ITDMs Can Make AI Investments Work in Their Favor

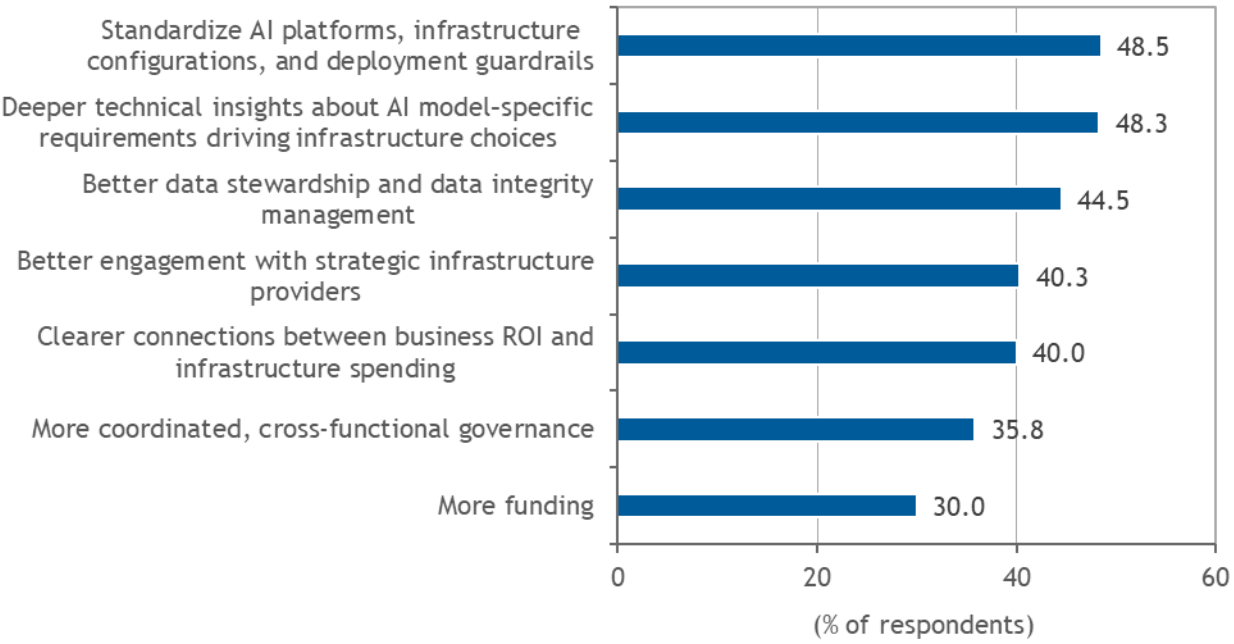
Regardless of the specific type of AI being deployed, successful businesses emphasize the need to embrace an infrastructure tech stack-oriented approach that allows them to standardize and automate infrastructure configuration and operations and share reusable AI capabilities across use cases. Recognizing that the consumption of any one model may fluctuate and that models will be updated, interconnected, and modified frequently, these businesses are implementing standardized approaches to preparing and integrating data, optimizing computing resource consumption, and minimizing networking cost and complexity.

In parallel, successful businesses are planning for flexible, hybrid infrastructure strategies that allow them to develop or tune models in one location and deploy them in others while consistently protecting confidential data and optimizing cost, quality, and performance. Most participants in the research indicate that they expect to rely on both public cloud and dedicated infrastructure resources, depending on the specific use case and usage profile being evaluated.

Standardized infrastructure platforms, configurations, and guardrails are the primary areas these businesses believe they need to invest in to support their AI-fueled business vision for 2026 and beyond (see Figure 2). Infrastructure platforms will allow businesses to create reusable templates and blueprints for automating and consistently deploying models, APIs, and security and access controls across workloads, regardless of whether they are deployed in public clouds, datacenters, colocation facilities, or edge locations.

FIGURE 2: ***Most Important Investments to Ensure That AI Infrastructure Is Ready for Business***

Q ***What are the most important things your organization needs to do to ensure that its AI infrastructure is ready for business in 2026 and beyond?***



Source: AI-Ready Infrastructure Tech Buyer Adoption Trends Research Highlights Demand for Scalable Infrastructure Platforms and Alignment with Business ROI (IDC #US52101825, December 2024)

To cost-effectively design, implement, and operate large-scale AI-ready infrastructure, businesses must:

- Standardize AI infrastructure platform architectures and operational guardrails.
- Match AI models with appropriate digital infrastructure resources.
- Implement data stewardship and data management.
- Engage more effectively with strategic infrastructure providers.
- Establish a connection between AI infrastructure investment and business ROI.

Lessons learned from businesses that have successfully transitioned multiple AI POCs to large-scale production environments show that success depends on strong alignment across business and technical teams. It also requires businesses to support seamless workload and data portability and to deploy flexible infrastructure architectures while ensuring consistent enforcement of corporate guardrails for data compliance, security, audit, and sharing.

Essential Guidance

IDC finds that businesses anticipate that AI will be the primary driver of their infrastructure tech stack investments and hybrid cloud and datacenter deployment strategies over the next several years. For the most part, many of them are still working to develop consistent, repeatable decision frameworks, as just 25% report they have a standard framework to guide AI infrastructure decisions. These businesses recognize they will need help from strategic vendors and partners as they explore infrastructure options and work to keep pace with the rapid evolution of the state of AI technology.

The experiences of these more mature AI infrastructure implementers highlight the infrastructure priorities and challenges many businesses will need to address as the mainstream market moves from testing AI in POCs to using AI in production across many business activities. As businesses embrace scalable, infrastructure platform–oriented approaches to automate and optimize AI infrastructure, decisions will need to increasingly focus on overall data capacity, the optimization and sharing of computational resources, and strategies to optimize trade-offs across the quality of outputs versus the cost. The move from POC to production requires businesses to recognize use-case patterns and invest in repeatable infrastructure platforms and services that provide multiple use cases access to standardized AI engines for services, such as chatbots or content summarization, while allowing the models to act on data and automate activities specific to the individual use case at hand.

Conclusion

Lessons learned from businesses that have successfully transitioned multiple AI POCs to large-scale production environments show that success depends on strong alignment across business and technical teams. Businesses must support seamless workload and data portability and deploy flexible infrastructure architectures while ensuring consistent enforcement of corporate guardrails for data compliance, security, audit, and sharing.

Businesses must support seamless workload and data portability and deploy flexible infrastructure architectures while ensuring consistent enforcement of corporate guardrails for data compliance, security, audit, and sharing.

About the Analyst

Ashish Nadkarni, *Group VP and General Manager, Worldwide Systems Infrastructure Research*

Ashish Nadkarni leads IDC's worldwide research on compute and storage infrastructure systems, platforms and technologies, enterprise, emerging and performance-intensive workloads, cloud and edge infrastructure and infrastructure services, and infrastructure software platforms.

MESSAGE FROM THE SPONSOR

AI is radically changing how businesses operate, compete, and innovate, but success hinges on having a high-performance datacenter. As AI workloads continue to grow in scale and complexity, integrating specialized storage and compute and efficient networking capabilities is paramount for ensuring long-term sustainability and growth.

Working with the right AI infrastructure partner grants you access to the expertise, hardware, and software ecosystem needed to accelerate AI adoption while ensuring cost-efficiency, performance, and scalability. Being at the forefront of AI datacenter innovation, AMD delivers end-to-end portfolio of AI solutions — from CPUs and GPUs like AMD EPYC™ and AMD Instinct™ to advanced networking solutions and even AI PCs — you need to build the AI-ready datacenter of the future.

Learn more about AMD's cutting-edge AI solutions ↗

IDC Custom Solutions

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
X: @IDC
blogs.idc.com
www.idc.com