Enterprise Strategy Group by TechTarget

SHOWCASE

OCTOBER 2024

The Undeniable Impact of GenAI on Infrastructure Decisions

Mike Leone, Practice Director and Principal Analyst

Abstract: Developing a strategy for generative AI (GenAI) is critical for organizations to turn IT investments into scalable and flexible solutions. To achieve and maintain a competitive position, organizations must make smart infrastructure decisions when building and optimizing a robust technical foundation.

GenAI: Transforming a Great Idea into a Successful Initiative

Adoption of GenAI is increasing due to a number of factors, including the growing visibility and support among technical and business decision-makers and organizations' interest in developing new, high-value use cases. According to TechTarget's Enterprise Strategy Group, a few unmistakable trends have emerged:[1]

- AI is now deeply ingrained into how organizations conduct business. Most organizations said AI is either fully embedded in their culture and operations (34%) or they are expanding their use of AI more broadly across the business (27%).
- Organizations are experiencing value from AI investments in relatively short timeframes. This trendline is rapidly accelerating: Four years ago, 39% of organizations said they saw value from AI investments within three months, and by 2023, the percentage surged to 72%.

Market Insight

72% of organizations have seen value from their AI investments within three months—nearly double the number of organizations that saw value in the same time period four years prior.

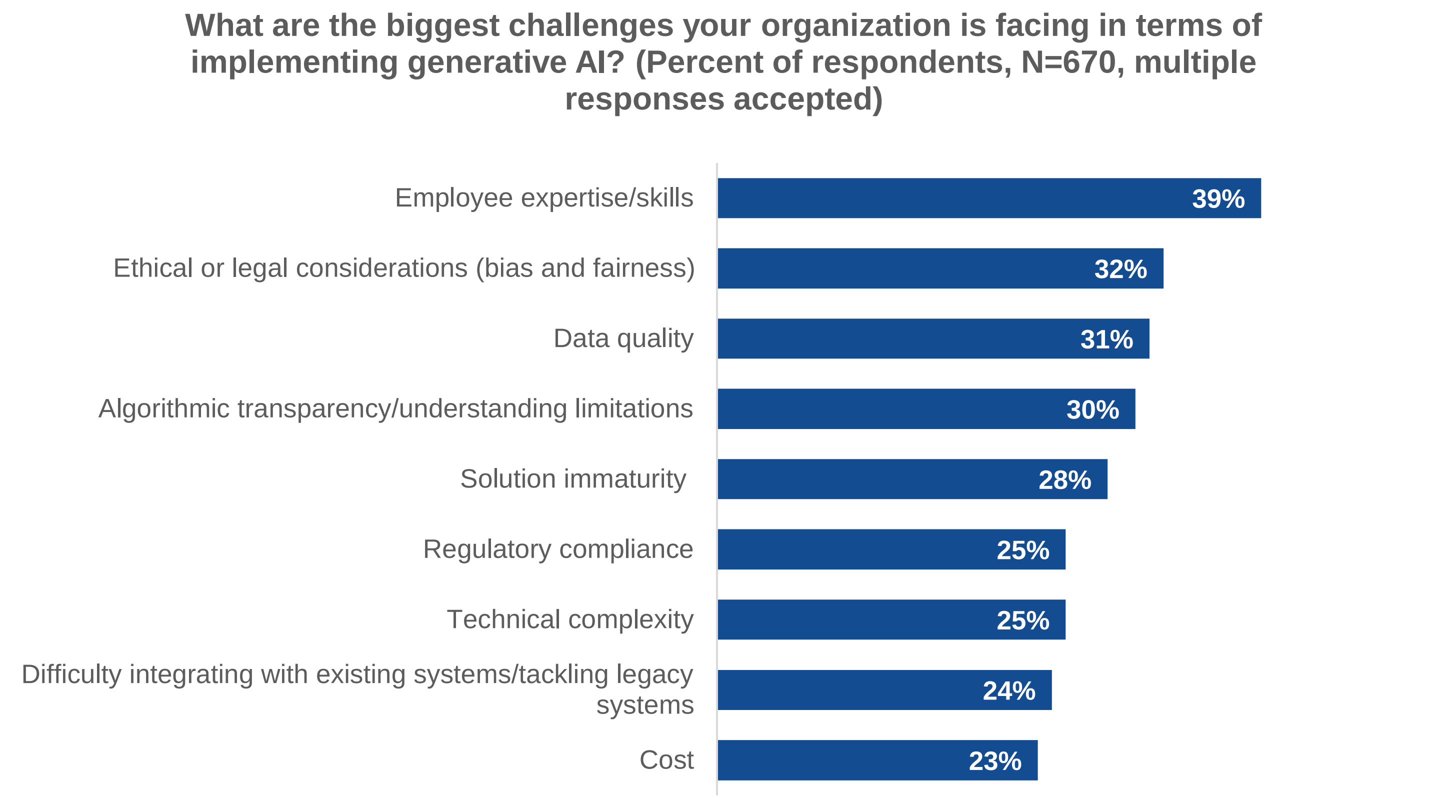
As a result, AI is at the top of most organizations' strategic business initiatives. This is especially true for GenAI, given the technology's proven ability to tackle critical use cases from data analytics and cybersecurity to facilitating decision-making and enhancing customer experiences.

New Technology Spawns Important Challenges

In order for organizations to experience the full potential of GenAI, decision-makers must recognize and overcome key challenges, such as the large and widening GenAI skills gap, the rapid changes in responsible GenAI use and governance, the need for improved data quality, and the growing web of compliance mandates for handling sensitive data.[2] The complete list of GenAI challenges is shown in Figure 1.

[1] Source: Enterprise Strategy Group Research Report, *Navigating the Evolving AI Infrastructure Landscape*, September 2023. All Enterprise Strategy Group research references and charts in this showcase are from this research report unless otherwise noted.
[2] Source: Enterprise Strategy Group Research Report, *Beyond the GenAI Hype: Real-world Investments, Use Cases, and Concerns*, August 2023.

Figure 1. Top 10 Challenges When Implementing GenAI



Source: Enterprise Strategy Group, a division of TechTarget, Inc.

As more organizations move from experimentation and proofs of concept (POCs) to training and inference, there emerges a growing need for AI infrastructure to deliver high performance, broad scalability, strong security frameworks, and increased energy efficiency. Without the right infrastructure choices, organizations often struggle to turn pilots and POCs into production systems and to gain demonstrable economic and operational value from GenAI.

Rethinking IT Infrastructure

Market Insight

98% of organizations said GenAI has expanded their infrastructure needs—35% of them indicated it has had “significant” impact on those infrastructure needs.

As organizations are experiencing the demand for GenAI adoption, they are also having to consider the underlying hardware and software infrastructure needed to drive successful projects. Organizations acknowledge that GenAI adoption necessitates the need to purchase more AI infrastructure to support training and inferencing of compute and memory-intensive workloads like large language models. Nearly all organizations (98%) said GenAI has expanded their infrastructure needs, with 35% of them indicating it has had “significant” impact on those infrastructure needs.

When evaluating AI infrastructure choices, organizations should consider the following questions:

- Does my GenAI infrastructure need to be implemented across the full spectrum of environments: cloud, data center, edge, and PCs/workstations? While this is a best practice, not all infrastructure vendors are positioned to deliver this.
- How are my infrastructure decisions addressing power consumption, sustainability, and security, as well as governance, risk, and compliance requirements?
- Will my AI infrastructure choices include a performant software stack with an open, broad, and well-resourced ecosystem?

This best practice will ensure that the selected GenAI infrastructure is versatile and capable of operating effectively across various environments, which is crucial for scalability and flexibility. By addressing power consumption, sustainability, and security, organizations can demonstrate a commitment to responsible use and compliance with regulations. And when considering a performant software stack within a well-resourced ecosystem, organizations are enabled to maximize the efficiency and effectiveness of AI initiatives, ultimately leading to better outcomes and innovation.

Diverse GenAI Use Cases Drive Unique Enterprise and Infrastructure Needs

Many popular transformative GenAI use cases such as medical imaging analytics and personalized marketing campaigns require massive amounts of compute power. These use cases are typically marked by distributed workloads, very large models that must be frequently updated, increasingly complex training and inference needs, massive data sets, the growing need to aggressively manage energy consumption, and the requirement for much higher levels of data storage and network bandwidth. GenAI also impacts a wide range of organizations and functional areas within the broader organization, including sales and marketing, human resources, workforce management, legal/compliance, and finance.

Examples in industries that have become major adopters of GenAI use cases include:

- **Quality control in manufacturing.** Organizations are transforming quality control in manufacturing by improving operational efficiency, reducing waste, and minimizing downtime. Models analyze massive data sets to identify even the smallest of defects that may be missed by humans to better predict potential equipment failures and provide proactive maintenance. Additionally, production bottlenecks can be identified faster, while virtual prototyping enables organizations to optimize their ability to identify potential quality issues before they arise.
- **Drug discovery in life sciences.** Life sciences organizations are leveraging GenAI to speed the process of finding and optimizing drug candidates. This includes creating new molecular structures with the right properties, modifying molecules to improve effectiveness, and designing new drugs that work best based on the biological ecosystem. Once an approved drug candidate has emerged, GenAI can help streamline the clinical trial process by helping to create and execute recruitment strategies for clinical trials.
- **Credit risk analysis in financial services.** GenAI is enabling financial services organizations to improve prediction accuracy, decision-making, and efficiency as they seek to optimize credit risk analysis. By analyzing large customer data sets, GenAI can create advanced credit scoring models and simulate multiple scenarios to improve risk forecasting. Through synthetic data generation, models can introduce training data sets to help test and improve fraud detection capabilities.
- **Product visualization in retail.** GenAI is transforming product visualization by delivering higher levels of customization and personalization. Customers are now empowered to see how clothing or accessories look through virtual try-ons. 3D product models can be created from 2D images to deliver more detailed viewing. Based on shopper preferences, GenAI can deliver personalized recommendations, while augmented reality (AR) lets customers visualize products from their phone cameras.

These and other use cases increase the need for diverse computing options to meet both training and inferencing workload requirements across diverse infrastructure environments.

Building and Optimizing IT Infrastructure for GenAI

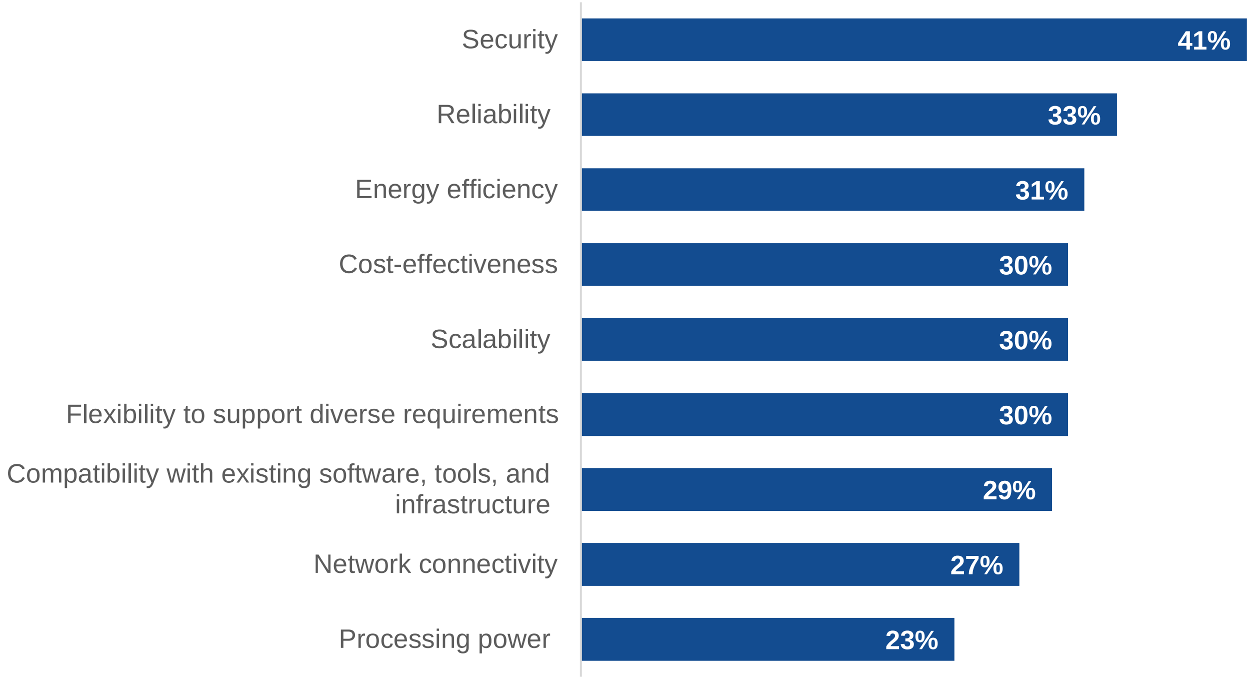
According to Enterprise Strategy Group research, the need for high-performance computing capabilities tops the list of capabilities organizations said they need when selecting AI infrastructure (42%). Organizations also prioritize such requirements as ease of deployment and integration with existing infrastructure (39%); data governance, security, and privacy features (39%); and scalability and flexibility for handling large data sets and models (39%).

A fundamental issue many organizations grapple with is whether to build their AI infrastructure around specialized GPUs or to leverage increasingly powerful, energy-efficient, and cost-effective general-purpose compute engines. In fact, 66% of organizations said they use or plan to use general-purpose compute infrastructure for high-performance computing requirements such as AI, compared to 62% that said they use or plan to use specialized compute engines (i.e., GPUs). Clearly, both types of compute engines have an important position on the AI infrastructure decision spectrum: CPUs offer the benefits of lower cost, wider availability, and multitasking capabilities, while GPUs are excellent for large-scale performance acceleration and parallel processing.

What else matters to organizations when making decisions on high-performance computing infrastructure for use cases such as AI? Figure 2 highlights recent Enterprise Strategy Group research that revealed the top factors organizations look for in high-performance computing infrastructure.

Figure 2. Most Important Factors in High-performance Computing Capabilities

When it comes to high performance computing (e.g., GPUs) capabilities, what factors matter the most to your organization? (Percent of respondents, N=339, three responses accepted)



Source: Enterprise Strategy Group, a division of TechTarget, Inc.

Unlocking the Full Potential of GenAI With AMD-based Infrastructure

As a long-standing market leader in high-performance compute infrastructure and related software, AMD is well positioned to meet the rapidly growing and diversifying infrastructure requirements to support GenAI workloads. Not only does AMD offer a wide array of compute engines, accelerators, and other hardware products optimized for GenAI, but its software portfolio also represents an essential element of the overarching strategy to meet organizations' GenAI needs.

AMD has built its AI strategy upon four priorities:

- A broad portfolio of AI training and inference compute engines.
- Open, proven, and developer-friendly software capabilities.
- An open AI ecosystem with deep "co-innovation" leveraging relationships with a wide range of hardware, software, and services partners.
- Leadership training and inferencing solutions at the rack, cluster, and data center level that can be rapidly deployed at scale across cloud and enterprise customers.

The AMD GenAI-centric product portfolio centers on its compute engines, starting with AMD EPYC™ processors. AMD EPYC™-based servers enable organizations to benefit from high-performance AI inference, leveraging the performance, scalability, compatibility, and energy efficiency of these highly optimized processors. AMD CPUs also enable workload consolidation to help organizations reduce requirements for physical space and power that often surge in GenAI use cases.

For use cases requiring greater computing performance, AMD Instinct™ MI300X accelerators help extend functionality and speed as GenAI workloads evolve and grow. AMD Instinct™ solutions offer greater memory capacity and the bandwidth necessary for compute- and memory-intensive workloads like LLMs, helping to effectively reduce the number of GPUs required for certain workloads. AMD Instinct™ accelerators enable leadership performance for the data center, at any scale—from single-server solutions up to the world's largest, exascale-class supercomputers.

AMD ROCm™ is the first optimized open AI software stack of its kind to enable organizations to run popular LLMs out of the box on AMD Instinct™ accelerators. AMD ROCm is engineered to align with performance and functionality requirements for GenAI and other high-performance computing workloads. AMD ROCm supports a wide range of popular models like Llama and open frameworks, such as TensorFlow, PyTorch, and ONNX, simplifying AI solutions' development and deployment.

AMD Ryzen™ AI CPUs enable enhanced PC and workstation productivity and creativity by supporting widespread upgrades, as well as the flexibility to intelligently adapt new use cases as future AI models become available. Ryzen™ AI products combining CPU, GPU, and NPU IP can deliver unmatched performance needed for real-time, interactive, compute-intensive Gen AI workloads, maximizing productivity without sacrificing best-in-class power efficiency.

Finally, AMD combines its engineering skills and customer service/support with a corporate commitment to responsible AI use, sustainability, security, and privacy. Its Responsible AI framework addresses important issues such as fairness, protecting confidential information, safety, and transparency through its human-centric AI design. AMD is focused on advancing research and development, product design, and collaboration to promote AI innovation for good.

Conclusion

Committing to strategic use cases based on GenAI technology is a key step for organizations looking to improve their competitive position and transform how they do business. To achieve those goals, organizations must ensure they have the right technical foundation in place; that means making smart, long-term decisions on AI infrastructure.

AI infrastructure choices must take into account a wide range of factors, such as performance, scalability, security, energy efficiency, the presence of an open and deep software framework, and an AI ecosystem that promotes innovation at every step. Making the right infrastructure selection—including picking the right infrastructure partner—helps organizations reduce risk and increase opportunity over the long haul.

AMD is a market-proven supplier of state-of-the-art end-to-end AI infrastructure—both hardware and software—for training and inference compute engines. Its broad portfolio of AI infrastructure solutions makes it the ideal partner for performance-intensive use cases that also require strong performance, robust security, and energy efficiency.

For more information about AMD AI solutions, please visit Empowering Enterprise with Generative AI and AI Solutions sites.

Information contained in this publication has been obtained by sources TechTarget considers to be reliable but is not warranted by TechTarget. This publication may contain opinions of TechTarget, which are subject to change. This publication may include forecasts, projections, and other predictive statements that represent TechTarget's assumptions and expectations in light of currently available information. These forecasts are based on industry trends and involve variables and uncertainties. Consequently, TechTarget makes no warranty as to the accuracy of specific forecasts, projections or predictive statements contained herein.

About Enterprise Strategy Group
TechTarget's Enterprise Strategy Group provides focused and actionable market intelligence, demand-side research, analyst advisory services, GTM strategy guidance, solution validations, and custom content supporting enterprise technology buying and selling.

✉ contact@esg-global.com
🖥 www.esg-global.com