White Paper

Performance, Agility, and Scalability for Succeeding in the AI and Hybrid IT Era

Sponsored by: IBM

Chris Drake
June 2025

IDC OPINION

As organizations look to advance their digital business strategies, they recognize the need to invest in IT infrastructure that can help them innovate faster and more efficiently while enabling them to manage business risk and reduce operational complexity. IDC's Future of Digital Infrastructure research offers organizations a framework for evaluating the opportunities for innovation and disruption that are made possible by strategic infrastructure technologies; these include AI-ready infrastructure, autonomous operations, and hybrid and multicloud interoperability. IBM's next-generation Power11 server platform addresses and incorporates several of the core elements of IDC's Future of Digital Infrastructure framework. The new Power11 platform has been specifically designed to help organizations harness opportunities in the era of AI and hybrid, multicloud operations. First, Power11 addresses both performance and compliance requirements of AI, including acceleration capabilities and the ability to access, transform, and manage enterprise data at scale, and support the seamless integration of generative AI (GenAI) capabilities with mission-critical enterprise processes. Second, the new platform also leverages new automation capabilities, bringing a range of benefits to Power platform customers that span management, reliability, security, and sustainability. Finally, IBM's Power11 is designed to support a distributed and flexible hybrid infrastructure with business-driven workload placement across on premises, public cloud, or private cloud based on performance, cost, and compliance needs.

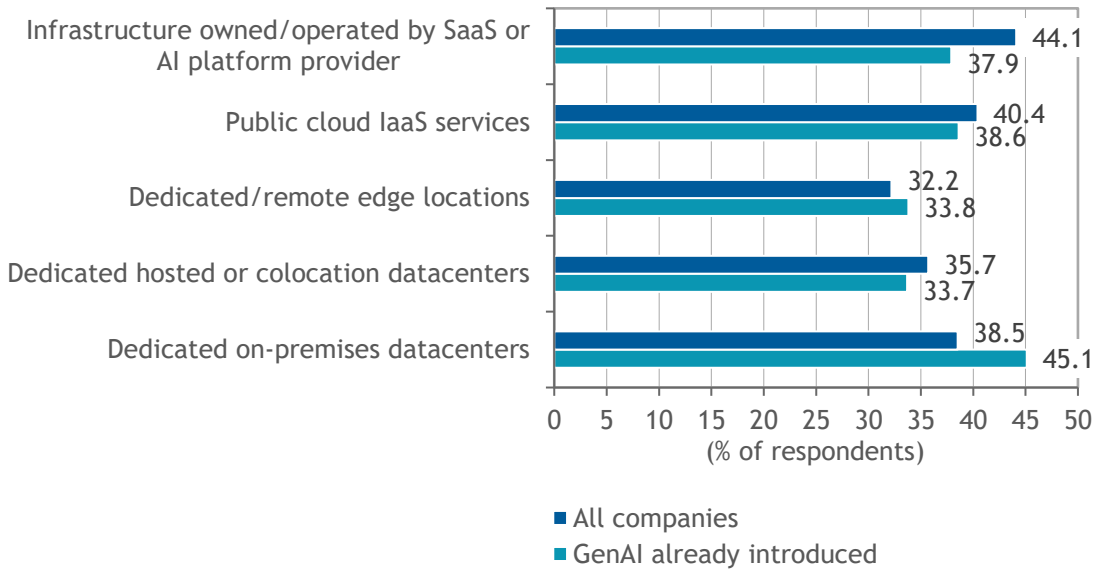
SITUATION OVERVIEW

As organizations become more experienced with GenAI tools, they increasingly recognize it as a major new corporate workload that requires dedicated infrastructure and partner strategies to support adoption and deployment. According to IDC research,

85% of organizations worldwide recognize GenAI as a major new corporate workload like ERP or ecommerce that will require increased technology spending over the next few years. 85% of organizations worldwide also agree that their ability to support GenAI as a strategic workload will require a dedicated vendor/partner strategy across infrastructure, software, data, cloud, and services that is separate from existing systems. The most experienced organizations demonstrate a preference for dedicated on-premises infrastructure to support GenAI model tuning, reflecting the importance they attribute to things such as data compliance and confidentiality (see Figure 1).

FIGURE 1

Primary Type of Computing and Storage Infrastructure Used to Support GenAI Model Tuning Over Next 18 Months



n = 889

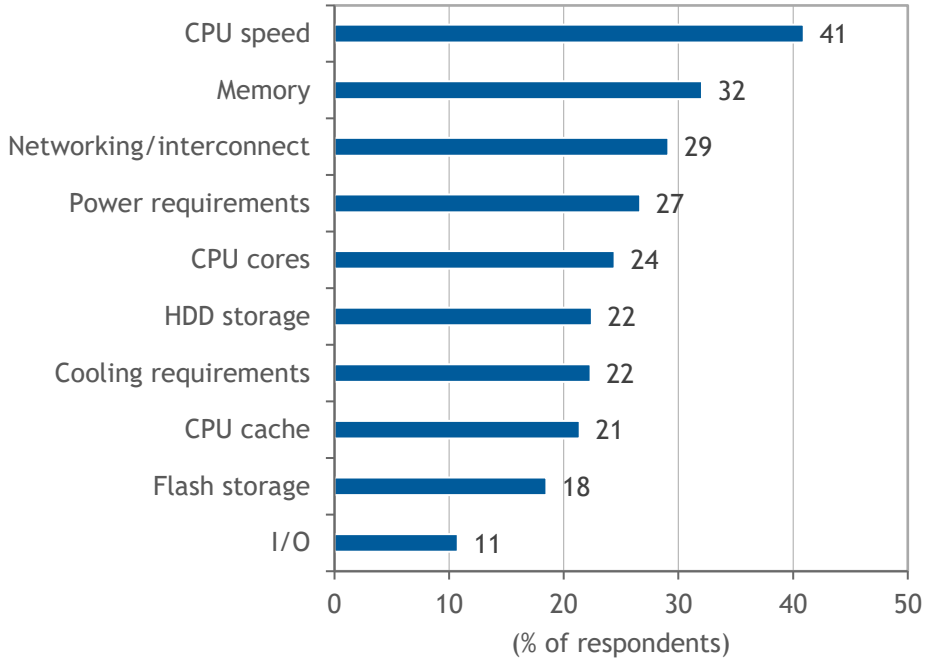
Source: IDC's *Future Enterprise Resiliency and Spending Survey, Wave 4,* April 2024

However, despite widespread recognition of the need for dedicated infrastructure and partner strategies to support the adoption and deployment of GenAI capabilities, most organizations are still at the stage of investing in and testing GenAI workloads in preparation for future launch. Many organizations continue to face a host of challenges within their on-premises compute environments, ranging from challenges related to CPU speeds and memory to those associated with their datacenter environments, notably networking, power, and cooling (see Figure 2). Organizations therefore need to ensure they are investing in future-proof technologies and strategies that can help them fully prepare their infrastructure environments for the AI era.

FIGURE 2

Greatest Resource Bottlenecks/Limitations Within On-Premises Compute Environments

Q. What is the greatest resource bottleneck or limitation for your on-premises compute/server infrastructure?



n = 855

Source: IDC's *Enterprise Infrastructure Pulse Compute Survey,* April 2024

The Future of Digital Infrastructure

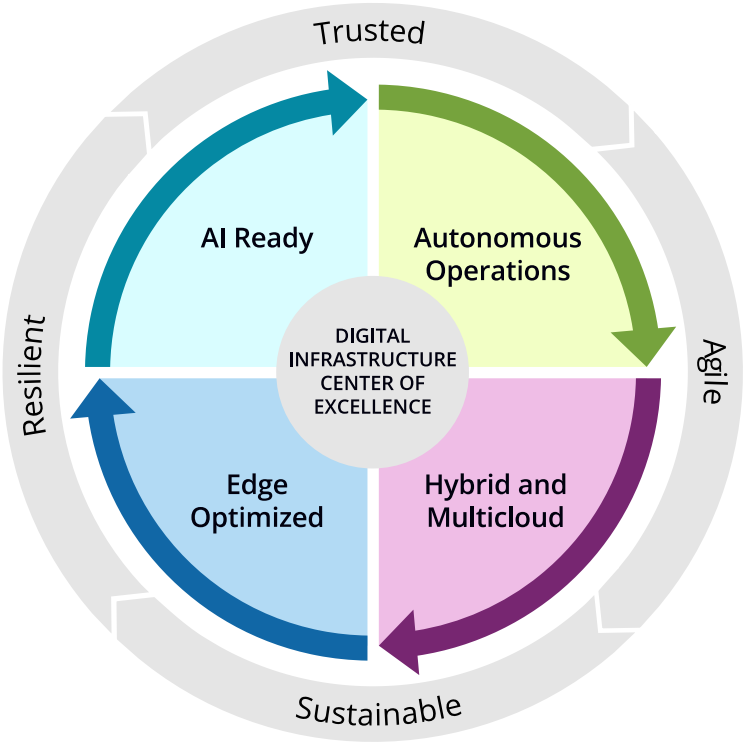
With the need to manage rising volumes of data and support workloads with growing computational requirements, the demands placed on server platforms and their supporting datacenter environments have never been greater. In addition to more powerful processing capabilities, server infrastructure must be capable of delivering high-performance levels of reliability, security, scalability, and interoperability while enabling organizations to adapt to changing demands in ways that are both agile and sustainable. With these multiple requirements, it is essential that server platforms are simple to manage and maintain and are able to help users reduce operational complexity.

In IDC's most recent *Worldwide Future of Digital Infrastructure Sentiment Survey,* 77% of organizations indicated that they believe digital infrastructure is "important" or "mission critical" for the success of their digital business strategies. IDC's Future of Digital Infrastructure research offers organizations a framework for evaluating the

opportunities for innovation and disruption that are enabled by strategic infrastructure technologies, including AI-ready infrastructure, autonomous operations, and hybrid and multicloud interoperability (see Figure 3). Key takeaways from this research include the argument that emerging digital infrastructure technologies will dramatically improve next-generation workload performance, security, scale, and total cost of ownership and that organizations need to determine how best to take advantage of these innovations based on the unique needs of their own workloads. The research also asserts that architecting, implementing, operating, and continually refreshing digital infrastructure requires IT leaders to anticipate ways in which datacenter and cloud infrastructure will evolve and continually disrupt the status quo. Based on this, they are then better positioned to proactively harness those technologies for business advantage.

FIGURE 3

IDC's Future of Digital Infrastructure Framework, 2024



Source: IDC, 2024

IDC's Future of Digital Infrastructure comprises several core elements:

- **AI-ready infrastructure.** It refers to technologies, products, and cloud services optimized for the scale, performance, cost, sustainability, and interoperability

requirements of emerging AI and other high-performance, data-intensive workloads. These technologies are designed for deployment and consumption across dedicated datacenters, colocation and hosting sites, edge locations, and public cloud footprints, depending on the needs of specific workloads and businesses. AI-ready infrastructure comprises different building blocks, including fit-for-purpose advanced coprocessors and accelerators that provide superior price/performance and price/efficiency ratios, next-generation storage, high-performance GenAI network fabrics, and sustainable datacenter technologies that address the power and thermal challenges created by emerging high-performance computing technologies and AI workloads.

- **Autonomous operations.** These are infrastructure operations that take full advantage of AI, observability, and automation to enable organizations to manage, scale, and secure infrastructure consistently across datacenters, colocation and hosting sites, mobile and edge locations, and public cloud infrastructure-as-a-service (IaaS) and software-as-a-service (SaaS) platforms. Autonomous operations represent an emerging operational model that allows organizations to more effectively scale and consistently manage and secure end-to-end infrastructure management and controls across interoperable hybrid, multicloud, and fit-for-purpose platforms and services. Autonomous operations can incorporate various capabilities and strategic foci, including predictive observability, self-driving infrastructure automation, GenAI-enabled end-user self-service, and IT staff and skills modernization.
- **Hybrid and multicloud interoperability.** It recognizes that applications and data are deployed according to the specific needs of individual workloads, as well as the interactions that are required when multiple applications and data repositories need to interconnect. IDC research consistently shows that many organizations are fully committed to hybrid and multicloud infrastructure strategies that involve deploying and managing workloads across a range of dedicated and shared footprints, according to the specific performance, cost, security, and scalability requirements of the individual application and data sets. Interoperable environments can benefit the entire organization, by enabling workload portability and modernization as well as supporting seamless data and process links.
- **Edge-optimized architectures.** These architectures anticipate and accommodate the increasingly distributed nature of enterprise computing and data management required by the convergence of IT and operating technology (OT) and growing demand for widely distributed network connectivity and location-independent workloads. Edge-optimized infrastructure is designed to be deployed and maintained in remote locations, away from core datacenters and cloud computing sites. Edge technologies typically need to be remotely managed

and designed to accommodate significant limitations in terms of access to power, cooling, and space. As organizations depend more on edge infrastructure, the need for uptime and resilience will increase, along with capabilities for data management and control.

- **Digital infrastructure centers of excellence (COEs).** These refer to collaborative governance models within organizations, which provide collaborative governance and strategic coordination across IT, cloud, line-of-business, DevOps, and data science teams; these teams must promote tech debt avoidance, interoperability, and coordinated engagement with strategic vendors and ecosystem partners. COEs typically include representation from IT, cloud, DevOps, data science, and line-of-business teams and are tasked with defining critical operational and architectural strategies. They are responsible for implementing key processes for guiding decisions about a range of topics that impact the broader organization, including cloud adoption, work deployment, data classification, integration templates, and tech debt remediation.

Power11 — IBM's Next-Gen Server Platform for the AI Era

Power11, the next generation of IBM's Power server platform, is designed to help organizations harness the opportunities of the era of AI and hybrid, multicloud operations. With innovations across its processor, hardware architecture, and virtualization software stack, IBM's Power platform delivers new capabilities like availability and resiliency while delivering the performance and scale needed to help organizations scale their workloads with fewer servers, optimizing energy and licensing costs. Hybrid deployment options allow organizations to run applications wherever their business requires and enables them to bring AI workflows securely and reliably to where their data resides.

IBM's Power11 solution comprises several differentiating features:

- **AI acceleration and insight.** Power11 is designed to meet performance and compliance requirements for AI inferencing and workflows across hybrid IT environments. The IBM Power11 processor offers enhanced performance, reliability, and efficiency compared with its predecessor, the Power10. It achieves this through higher clock speeds, a 25% increase in cores per chip, and improvements in power consumption and quantum security. Power11 will also add off-chip acceleration with the IBM Spyre Accelerator, which will be available in the near future. It is a purpose-built, enterprise-grade accelerator designed for AI inferencing tasks with high efficiency and scalability, particularly for complex models and generative AI. This will enable seamless AI integration into existing applications and workflows running on Power and deploy a broader range of AI use cases. IBM is also looking to use AI to modernize Power applications with the

upcoming IBM watsonx Code Assistant for i. It is expected to accelerate RPG code modernization tasks for IBM i applications with AI-powered capabilities made available directly in the integrated development environment.

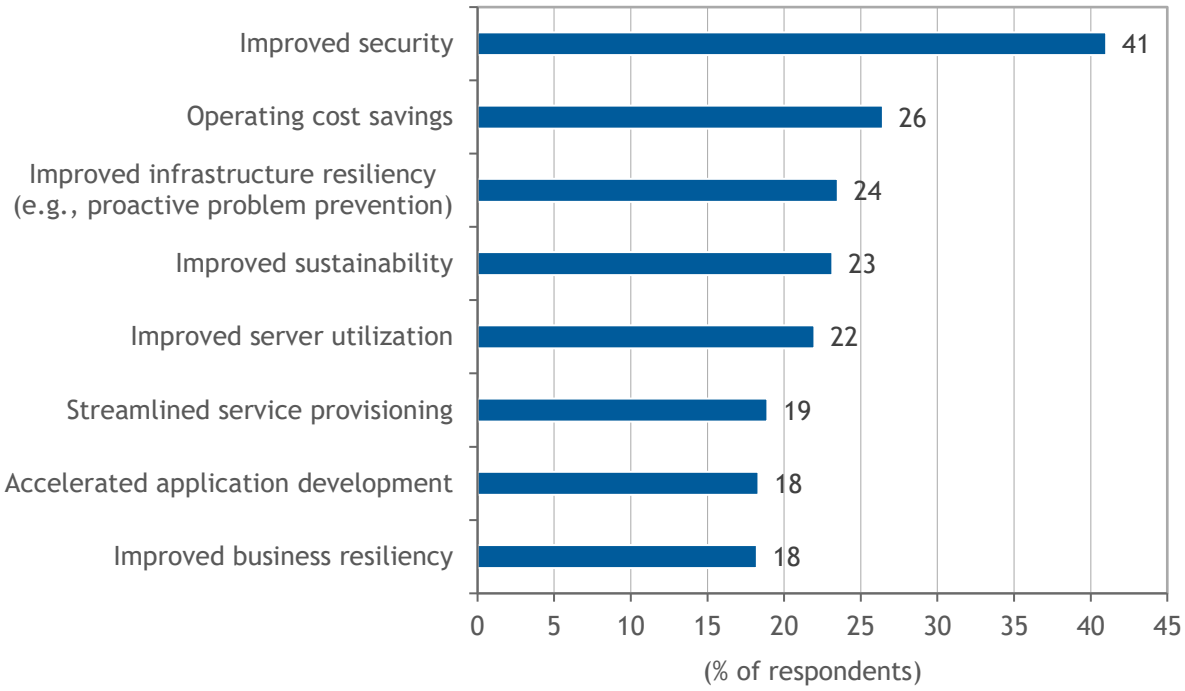
- **Autonomous IT efficiency.** IDC research indicates that organizations associate automation with a range of benefits, the most important of which include improved security, resiliency, and sustainability (see Figure 4). By leveraging new automation capabilities, Power11 is intended to bring users a host of new benefits, which include automated server, storage, and network upgrades and autonomous error resolution tools that self-identify problems and self-heal predictively and reactively, eliminating disruptions to system availability. New automated system maintenance capabilities include live updates, rolling upgrades, and autonomous patching for necessary maintenance without the need to take critical applications offline. This allows planned downtime to be reduced to zero while boosting productivity without requiring skilled expertise to manage the Power platform. Autonomous capabilities also extend to security and sustainability. IBM Power Cyber Vault — a unified solution for Power11 with Storage, Software, and Expert Labs services — is built to provide operational resilience to detect, respond, and recover from evolving cyberthreats while complying with regulatory standards. Power11 customers will also be able to integrate quantum-safe cryptography across Power infrastructure with automated tools for cryptographic inventory discovery, risk prioritization, and full-stack quantum-safe protection. Meanwhile, the use of smart energy scheduling allows customers to optimize server energy use and improve energy efficiency. Additional sustainability features include innovations in cooling, including the use of improved heat sinks and more efficient fans to optimize energy delivery.
- **Hybrid cloud consistency and flexibility.** Power11 is designed to support a distributed hybrid infrastructure. The new platform is a highly resilient runtime environment with a consistent experience and architecture that can span multiple deployments. It offers the ability to create, use, and manage IBM Power virtual machines/environments across IBM public cloud and private cloud deployments with a single user experience. This can extend existing application and data workloads into an as-a-service environment with a streamlined onboarding experience and without the added cost or complexity associated with refactoring or replatforming. With IBM Power Virtual Server (PowerVS), customers can provision and manage virtual servers running in IBM Cloud, with benefits including on-demand provisioning, flexible scalability, and automated deployment, security, and compliance. With Power11, PowerVS will also introduce quantum-safe compliance and Power Cyber Vault to strengthen business resilience use cases, including more efficient disaster recovery, geographic separation, or third site instances driven by regulatory requirements.

Power11's integration with Kernel-based KVM virtualization also enhances the solution's Linux compatibility and the ability to leverage hybrid virtualization solutions. Meanwhile, customers can also take advantage of flexible consumption and delivery options — with IBM Power Virtual Server Private Cloud, IBM provides the physical infrastructure (compute, network, and storage) and operates it within the customer's own datacenter.

FIGURE 4

Top Benefits of Server Automation

Q. What top benefits does server automation provide to your organization?



n = 851

Source: IDC's *Enterprise Infrastructure Pulse Compute Survey,* April 2024

CONCLUSION

The new Power11 server platform is well positioned to help IBM support customer efforts to capitalize on the era of AI and hybrid, multicloud operations. IBM's AI everywhere strategy positions the company as a leading enabler of enterprise AI initiatives. IBM's commitment to embedding AI across a range of solutions spanning, not only Power Systems but also IBM Z and IBM Storage, puts it in a leading position to

ensure mission-critical security, scalability, and compliance for AI workloads across hybrid and multicloud environments.

IBM's new Power11 solution offers the performance, efficiency, security, and flexibility that organizations require from next-generation infrastructure. Together with its portfolio of hybrid cloud technologies that includes watsonx AI, OpenShift AI, and Red Hat hybrid cloud solutions, IBM is well placed to provide customers with a flexible, cost-effective, and operationally efficient AI ecosystem that supports AI development and deployment across diverse cloud, distributed, and sovereign IT environments. By integrating AI-driven automation into IT, cybersecurity, and data management, IBM will be a driver of AI adoption across a range of regulated industries — including financial services and healthcare — where trust and governance are essential.

IBM's biggest challenge will be to demonstrate the scaling AI solutions and customer bases beyond the company's traditional enterprise base. Other leading cloud providers and hyperscale digital service providers are also deploying and aggressively promoting AI-native cloud services that offer seamless integration with broader ecosystems. IBM must therefore ensure that it is seen as a core AI enabler of AI across hybrid, interoperable environments.

The strategic focus of IBM Power11 on hybrid, multicloud environments, where the complexity of AI workload optimization is often identified by organizations as a risk that can slow adoption, must also be directly addressed. To support the successful growth of Power11 and its long-term competitiveness in the fast-changing enterprise AI landscape, IBM must also continue to focus on refining its developer ecosystem and AI orchestration capabilities.

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