White Paper, sponsored by Broadcom | December 2024

On-Premises AI Infrastructure Balances Innovation and Security

Madhumitha Sathish
Research Manager,
High Performance Computing, IDC

Peter Rutten
Research Vice President,
Performance Intensive Computing,
Worldwide Infrastructure Research, IDC

Heather West, Ph.D.
Research Manager,
Performance Intensive Computing,
Worldwide Infrastructure Research, IDC

Table of Contents

CLICK ANY HEADING TO NAVIGATE DIRECTLY TO THAT PAGE.

Summary

As enterprises adopt generative AI (GenAI) to streamline processes, create content, automate IT, and refine supply chains, they often face a crucial choice: to build and deploy AI models on premises or rely on public cloud infrastructure. Although public clouds provide scalability and access to a variety of pre-trained AI models, they can raise concerns over data security, control, and unforeseen costs. In contrast, private AI environments offer a customizable, secure, and cost-efficient framework that prioritizes data privacy and compliance. This paper lays out the current AI infrastructure landscape, factors enterprises are considering for on-premises AI infrastructure deployments, and key factors for leveraging private AI to meet organizational objectives.

Scope of This Study

Global survey:

411
respondents

Director-level titles
or above

5
countries

For the purpose of this study, training and fine-tuning/customizing AI models are defined as developing AI, while inferencing on those models, with or without retrieval-augmented generation (RAG), is defined as deploying AI models.

IDC defines private cloud, traditional (non-cloud) deployments, and colocation providers as on premises. Public cloud providers, managed service providers, and special-purpose cloud providers are defined as public cloud deployments.

This IDC White Paper is based on data collected through a global survey of 411 individuals with director-level titles or above from Germany, Japan, Singapore, the United Kingdom, and the United States.

Situation Overview

As AI becomes increasingly central to organizational strategy, the landscape for AI infrastructure deployment continues to expand, bridging on-premises and public cloud environments. Some of the top AI use cases today include contact center efficiency, document search/summarization, content creation and marketing, code development, IT operations automation, risk mitigation, and supply chain optimization.

These use cases are deployed across public cloud and on-premises infrastructure, highlighting the dual approach companies often adopt to leverage the flexibility of the cloud while retaining control over sensitive data on premises.

These applications showcase AI's adaptability across industries and emphasize the increasing demand for optimized AI infrastructure that accommodates private and public cloud environments.

AI Models, Data, and Security

Organizations have varying preferences for sourcing GenAI foundation models. As seen in **Figure 1**, open source (51%) and cloud provider (46%) models remain top choices but for different reasons.

FIGURE 1

Preferred Sources for AI Models

What are your preferred sources for AI models?

(Percentage of respondents)

n = 411; Source: IDC's *AI Infrastructure Survey,* July 2024

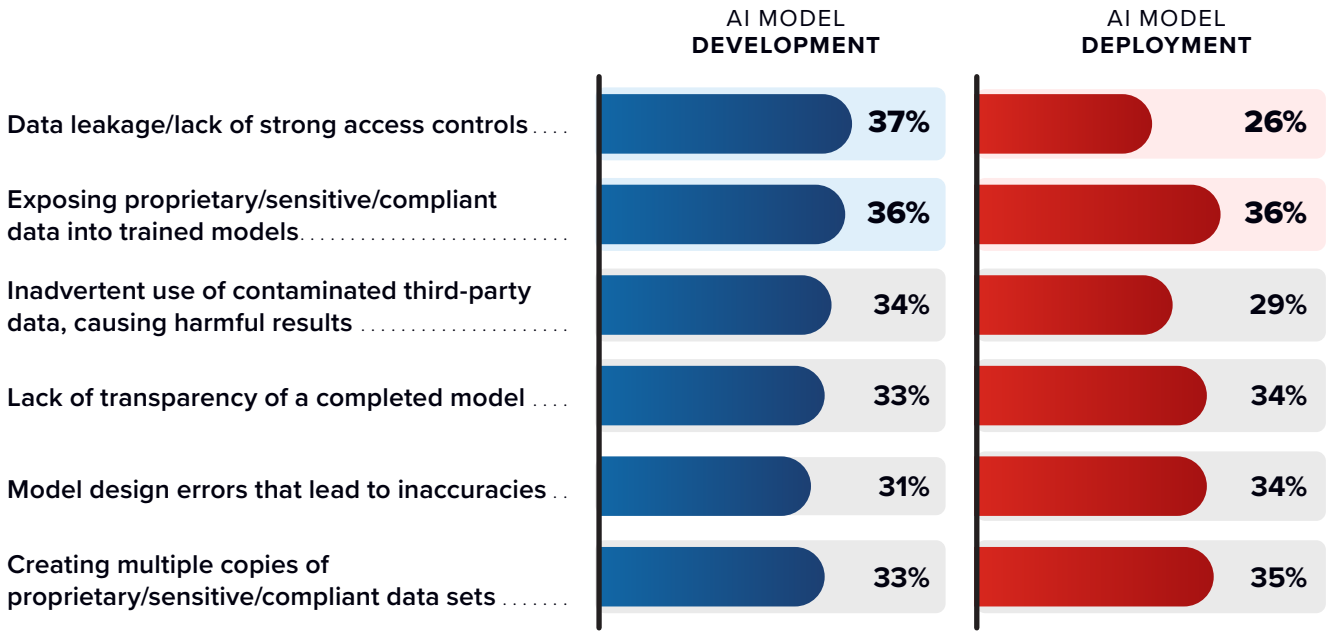
Public cloud and commercial ISV GenAI models can be attractive because of their quick deployment, ease of adoption, scalability, and performance. On the other hand, open source models appeal to organizations seeking greater customization, flexibility, and transparency. Open source models offer the ability to assess, modify, and troubleshoot models more easily, which appeals to teams that require a high degree of control and customization.

Despite these varied approaches and preferences for sourcing foundation models, there is universal concern about data privacy and security when using AI models. Fifty-four percent of survey respondents are extremely or very concerned about the possibility that the data used for AI development or deployment could be compromised. As seen in **Figure 2**, data leakage and exposing proprietary/sensitive/compliant data into trained models are cited as the top risks associated with developing or deploying AI models.

FIGURE 2

Top Risks Associated with AI Model Development and Deployment

What are the top risks that your organization associates with developing and/or deploying AI models? (Percentage of respondents)



n = 411; Source: IDC's *AI Infrastructure Survey,* July 2024

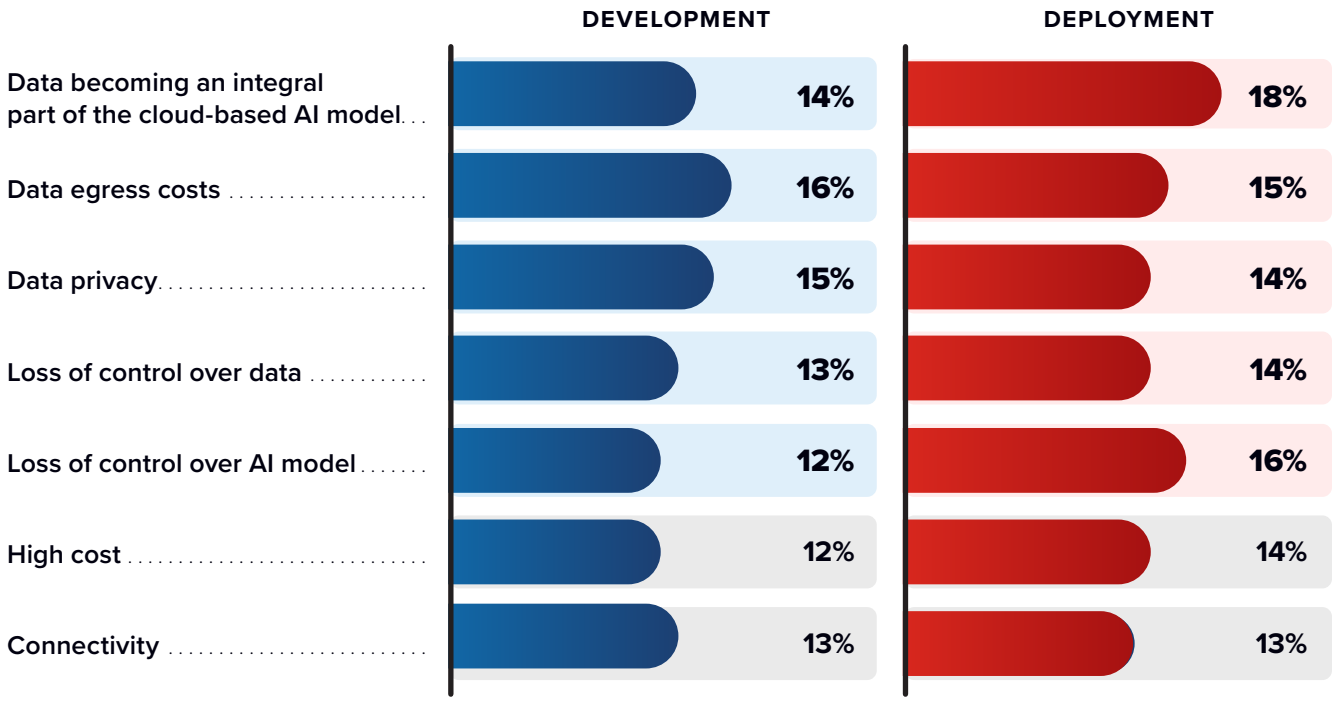
Organizations are wary of exposing proprietary or sensitive data into trained AI models, as this could result in data breaches or inadvertent leakage of confidential information. For many organizations, the concern is not limited to external threats but includes issues such as data leakage and unintentional use of contaminated or biased third-party data, which could compromise model integrity.

The choice between public cloud and on-premises environments introduces its own concerns. Data becoming an integral part of the cloud-based AI model and the subsequent loss of control over the AI model, egress costs, data privacy, and data control are top concerns about public cloud environments (see **Figure 3**).

FIGURE 3

Top Hurdles to Developing and/or Deploying AI Services in a Public Cloud

What are the biggest hurdles that your organization is experiencing with developing and deploying AI services in a public cloud?
(Percentage of respondents)



n = 191 (Base: Respondents indicated organizations primarily develop/deploy AI models on premises);
Source: IDC's *AI Infrastructure Survey,* July 2024

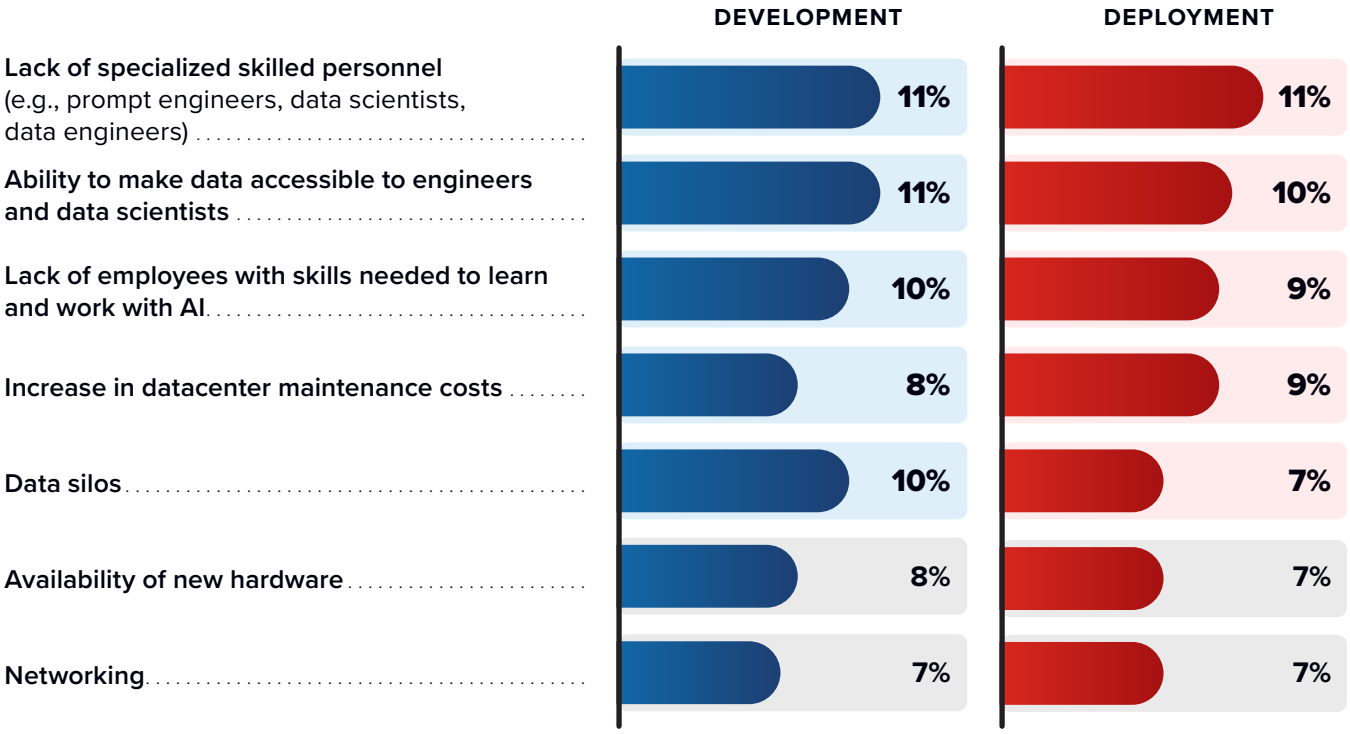
For on-premises AI deployments, companies cited challenges in data accessibility for engineers, limited specialized skill sets, AI/IT personnel shortages, data silos, and expanding datacenter maintenance costs (see **Figure 4**).

FIGURE 4

Top Hurdles to Developing and/or Deploying AI Services On Premises

What are the biggest hurdles that your organization is experiencing with developing and deploying AI services on premises?

(Percentage of respondents)



n = 191 (Base: Respondents indicated organizations primarily develop/deploy AI models on premises);
Source: IDC's *AI Infrastructure Survey,* July 2024

Adopting On-Premises AI Infrastructure

Organizations face crucial decisions about where and how to develop and deploy models effectively while also addressing security and compliance concerns.

Although public clouds provide scalability and access to a variety of pre-trained models, they can raise concerns over data security, control, and unforeseen costs. In contrast, private AI environments offer a customizable, secure, and cost-efficient framework that prioritizes data privacy and compliance while maximizing performance.

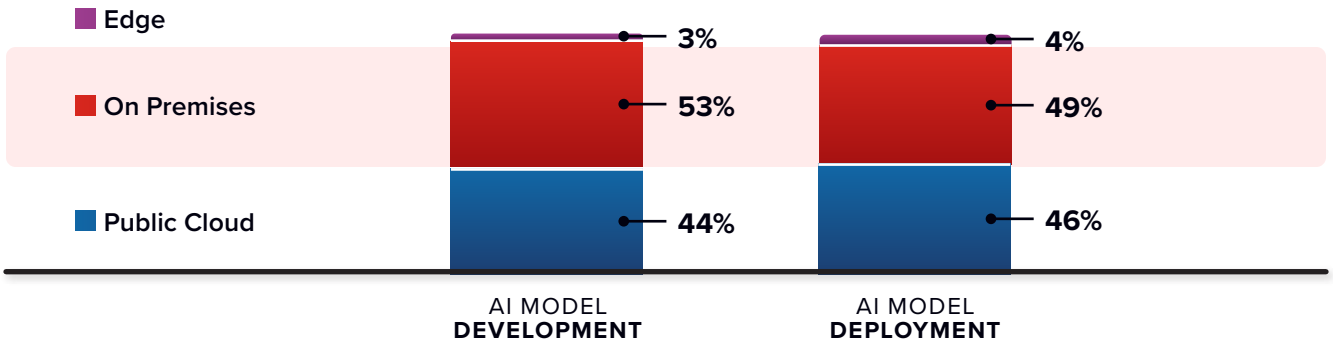
When it comes to infrastructure preferences, organizations demonstrate a modest preference for on-premises development (53%) and deployments (49%), underscoring the importance of maintaining data control, compliance, and low-latency access (see **Figure 5**, next page).

FIGURE 5

Deployment Location for the Development and Deployment of AI Models

Where does your organization primarily develop and deploy AI models?

(Percentage of respondents)



Note: Public cloud includes public cloud provider, managed services provider, and special-purpose cloud. On premises includes private cloud, traditional infrastructure, and colocation provider. n = 411; Source: IDC's *AI Infrastructure Survey*, July 2024.
For an accessible version of the data in this figure, see Figure 5 Supplemental Data in Appendix 1.

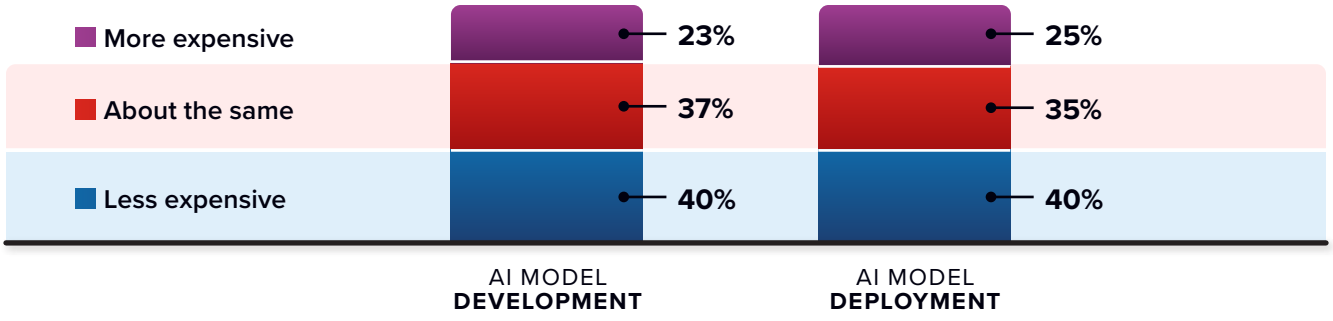
In addition to improved data privacy and security, organizations see on-premises AI infrastructure as cost competitive versus public cloud. Sixty percent of organizations that IDC surveyed perceived developing and deploying AI models using on-premises AI infrastructure as costing less than or about the same as the public cloud (see **Figure 6**).

FIGURE 6

Perceived Costs of Public Cloud Versus On-Premises AI Models

Does your organization perceive the cost of developing or deploying AI models in the public cloud as more expensive or less expensive than developing or deploying AI models on premises?

(Percentage of respondents)



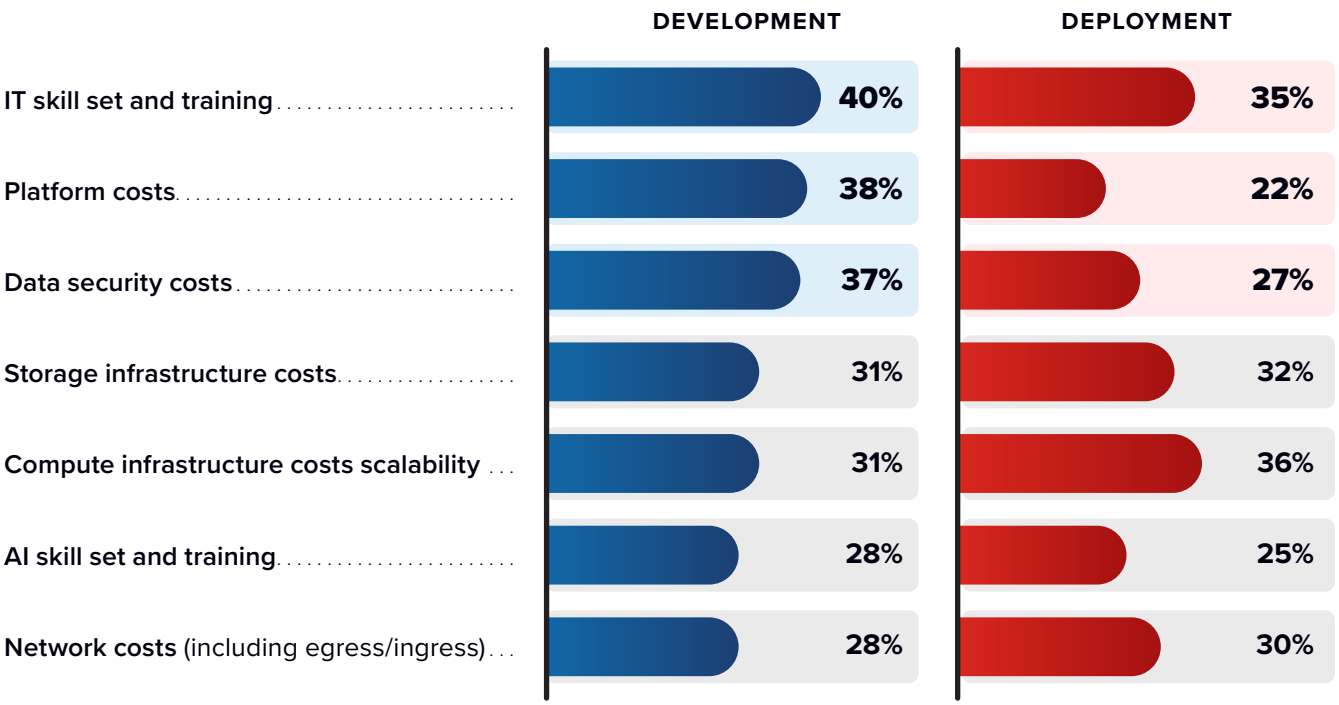
Note: Public cloud includes public cloud provider, managed services provider, and special-purpose cloud. On premises includes private cloud, traditional infrastructure, and colocation provider. n = 411; Source: IDC's *AI Infrastructure Survey,* July 2024
For an accessible version of the data in this figure, see Figure 6 Supplemental Data in Appendix 1.

Figure 7 shows that IT skill set and training, platform costs, and data security costs were the top 3 reasons that public cloud was viewed as more expensive for AI model development. When it came to AI model deployment, compute infrastructure costs rose to the top reason, which indicates organizations are likely concerned about surprise cloud bills due to AI models. In addition to overall costs, organizations may also see on-premises AI infrastructure as offering more cost predictability and stability, especially over time.

FIGURE 7

Reasons for the Perception That Public Cloud AI Models Are More Expensive than On Premises

Why does your organization perceive developing or deploying AI models in the public cloud as more expensive than on premises?
(Percentage of respondents)



n = 95 (Base: Respondents indicated organizations perceive developing models in the public cloud as more expensive);
Source: IDC's *AI Infrastructure Survey,* July 2024

Essential Features for On-Premises AI Infrastructure

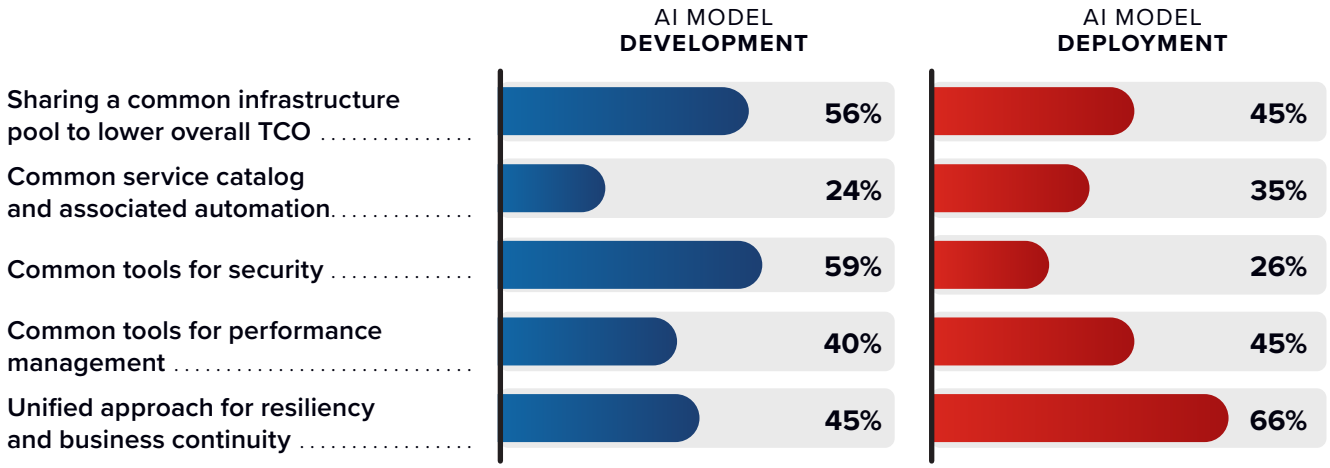
Today, AI workloads, especially GenAI workloads, are viewed as discrete from other workloads by many organizations.

However, in the future, IDC believes that organizations will want private cloud AI workloads managed similarly to any other enterprise workload. This includes a common platform with a comprehensive feature set. As seen below in **Figure 8** (next page), the common platform should include features for resiliency and business continuity, shared infrastructure, automation, performance management, and security.

FIGURE 8

Features Sought for a Common Platform to Develop and/or Deploy AI and Non-AI Workloads

Which of the following features would cause your organization to invest in a private cloud with a common platform to develop and/or deploy AI and non-AI workloads?



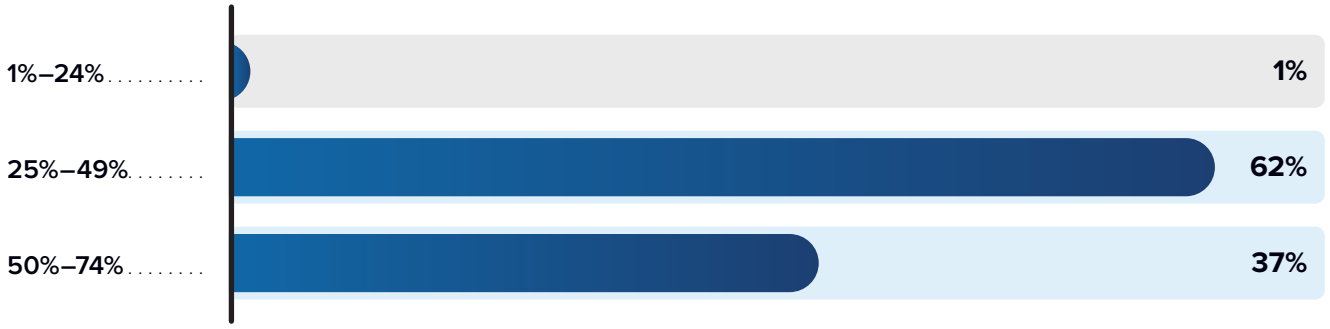
n = 124 (Base: Respondents indicated their organization's AI operations and management are not currently integrated within the overall IT environment.);
Source: *AI Infrastructure Survey Sponsored by Broadcom,* July 2024

A seen in **Figure 9**, survey respondents universally plan to virtualize some amount of their private environments in the next 24 months. Among them, 62% foresee virtualizing 25%–49% of their environment, whereas 37% plan for 50%–74% virtualization.

FIGURE 9

Virtualization of Private AI Environment

What percentage of your private AI environment will be virtualized?
(Percentage of Private AI Environment Virtualized)



Note: The mean percentage of on-premises virtualization is shown.
n = 220 (Base: Respondents indicated organizations primarily develop/deploy AI models on premises); Source: *AI Infrastructure Survey Sponsored by Broadcom,* July 2024
For an accessible version of the data in this figure, see Figure 9 Supplemental Data in Appendix 1.

The survey results imply that organizations expect virtualization to deliver similar benefits to private AI infrastructure as it has to traditional servers over the past two decades.

IDC sees this trend continuing because the virtualization of private AI servers offers substantial benefits, such as hardware and maintenance cost savings, improved resource utilization and management, improved disaster recovery and business continuity, and security.

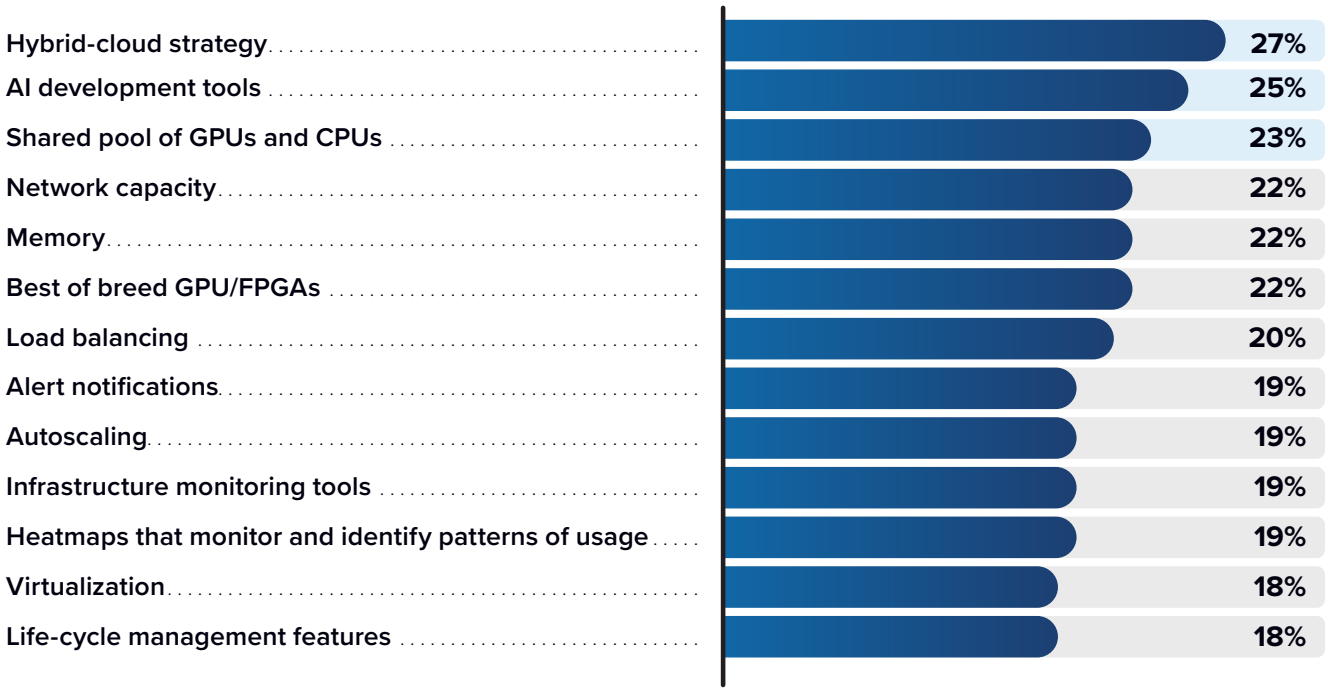
In IDC's survey, the respondents were asked about several server and storage features that would best help their organization achieve its goal of optimizing the cost and performance of on-premises AI infrastructure.

The data in **Figure 10** shows that a hybrid cloud strategy, enterprises' access to AI development tools, a shared pool of GPUs and CPUsare considered important features for optimizing the cost and performance of on-premises AI infrastructure.

FIGURE 10

Server and Storage Features to Optimize On-Premises AI Infrastructure

Which of the following server and storage features would best help your organization in its goal to optimize the cost and performance of on-premises AI infrastructure?
(Percentage of respondents)



. n = 220 (Base: Respondents indicated organizations primarily develop/deploy AI models on premises);
Source: *AI Infrastructure Survey Sponsored by Broadcom,* July 2024

Considering VMware Private AI Foundation with NVIDIA

Many of the features and attributes surveyed organizations are seeking with on-premises AI model development and deployment are found within the VMware Private AI Foundation with NVIDIA. VMware, part of Broadcom, and NVIDIA have collaborated to develop the joint GenAI platform VMware Private AI Foundation with NVIDIA.

This joint GenAI platform enables enterprises to fine-tune LLM models, deploy RAG workflows, and run inference workloads in their datacenters, addressing privacy, choice, cost, performance, and compliance concerns. This joint platform simplifies GenAI deployments for enterprises by offering an intuitive automation tool, deep-learning virtual-machine images, a vector database, and GPU monitoring capabilities.

Built and run on the private cloud solution VMware Cloud Foundation (VCF), VMware Private AI Foundation with NVIDIA includes the NVIDIA Enterprise, NVIDIA NIM (included with NVIDIA AI Enterprise), and NVIDIA LLMs and provides access to other community models (such as Hugging Face models). VCF, VMware's comprehensive private cloud infrastructure solution, offers a secure, comprehensive, and scalable platform for building and operating GenAI workloads, providing organizations with the agility, flexibility, and scalability to meet their evolving business needs. VCF provides a flexible platform to enable enterprises to easily integrate existing data pipelines, workloads, internal AI applications, and ISV applications onto a common platform and unify infrastructure resources and data. VMware Private AI Foundation with NVIDIA is an add-on service for VCF.

- **Secure AI stack:**
  A private and secure AI stack is crucial for securing data privacy and supporting compliance requirements. It also addresses cost optimization by reducing data transfer and security management expenses.

- **GPU monitoring and performance tools:**
  Dedicated dashboards for GPU utilization and common tools for performance management allow organizations to oversee usage, minimize bottlenecks, and make data-driven adjustments, ensuring the infrastructure remains efficient.

- **Privacy and control:**
  Organizations can deploy and run AI models adjacent to their private data. They do not have to move their data to a provider's infrastructure and proprietary data stores to obtain the benefits of AI.

- **Open and extensible:**
  Application developers and data scientists can interact with the platform using familiar APIs and CLIs (e.g., Kubernetes, NVIDIA, and Hugging Face).

- **Automated resource sharing:**
  AI workloads are about more than the GPU. Organizations must intelligently provision, monitor, and potentially resize GPUs, networks, data I/O, and CPUs. VMware's distributed resource scheduler technology has grown and evolved over the past 20 years, and this is bringing many bare-metal deployments to VMware.

- **Diverse AI tools:**
  Organizations should have access to a broad set of AI tools and frameworks, which empowers teams to experiment and implement customized AI workflows that best meet project needs.

- **Advanced capabilities for LLM fine-tuning and RAG workflows:**
  The offering must enable enterprises to fine-tune LLM models and deploy and run inference workloads in their datacenters. This approach ensures data control, compliance, and optimized costs.

- **Software-defined architecture for versatile workload management:**
  With software-defined architecture, administrators can assign specific workload domains, enhance GPU utilization, support varied AI applications, and tailor resources to specific performance requirements.

- **Hybrid cloud–ready platforms:**
  A unified platform that supports hybrid environments enables organizations to utilize cloud and on-premises resources based on workload demands. This flexibility is critical for adapting to fluctuating demands without sacrificing performance or cost control.

USSFCU Accelerates Its GenAI Journey Using the VMware Private AI Foundation With NVIDIA

As part of its commitment to innovation, United States Senate Federal Credit Union (USSFCU) embarked on a journey to integrate generative AI into its operations. Leading this initiative is Mark Fournier, the chief information officer, who spoke with IDC about the organization's GenAI strategy and technology.

USSFCU began its GenAI journey with a single dedicated workstation dedicated to experimentation. This initial setup allowed the IT team to explore the potential of GenAI and understand its capabilities. Recognizing the potential of GenAI, USSFCU sought to expand its capabilities so that departments across the organization could create and test use cases. The existing VMware Cloud Foundation platform played a crucial role in this expansion. With the infrastructure already in place, adding the new GenAI capabilities required only minor tweaks, making the process nearly seamless.

Currently, 145 employees at USSFCU are actively experimenting with GenAI with the Meta LLama model, but the team remains open to experimenting with various models to find the best fit for their needs. This flexibility to adjust as the foundation model landscape evolves was an important consideration for the USSFCU. USSFCU operates their deployment of VMware Private AI Foundation with NVIDIA, as a controlled sandbox environment setup that provides a safe space for innovation, allowing staff to explore and test new ideas without the risks of exposing confidential data. This controlled environment has been pivotal in fostering a culture of experimentation and learning within the organization.

The integration of VMware Private AI Foundation with NVIDIA has offered several significant benefits to USSFCU:

- **Simplified adoption:**
  Using their existing infrastructure and expertise, USSFCU found the adoption of VMware Private AI Foundation with NVIDIA solutions straightforward. The familiarity with VMware's ecosystem and the seamless integration with NVIDIA's tools made the transition smooth, minimizing the learning curve for the staff.

- **Access to NVIDIA's Neural Network Exchange Management System (NEMS):**
  The access to NVIDIA's NIM simplified the deployment of the AI services within the USSFCU datacenter and will be essential as the organization develops various use cases.

- **Utilizing existing infrastructure:**
  One of the key advantages of VMware Private AI Foundation with NVIDIA is its ability to leverage existing infrastructure. This compatibility allowed USSFCU to expand its GenAI pilot without the need for extensive new investments. The API connectivity to cloud-based GenAI models further streamlined the process, enabling quick and efficient integration.

With the foundational infrastructure and tools in place, USSFCU is now focused on mapping out specific use cases and developing a comprehensive roadmap for GenAI integration across the organization. The confidence gained from the initial phase has laid a strong foundation for scaling up these efforts. The organization is exploring various applications of GenAI, from enhancing customer service to improving internal processes and decision-making.

The journey of USSFCU exemplifies how leveraging existing infrastructure and strategic partnerships can drive affordable and secure innovation. As USSFCU continues to explore and implement new use cases, the organization is well-equipped to scale its GenAI initiatives with VMware Private AI Foundation with NVIDIA.

Conclusion

As companies aim to unlock AI's transformative potential, they must carefully balance cloud and on-premises resources, selecting models that not only optimize performance but also meet stringent requirements for data security, compliance, and cost control.

By carefully planning and managing these aspects, organizations can build a resilient AI infrastructure that supports their goals. As AI infrastructure continues to mature, these choices are becoming more strategic, with businesses investing in robust platforms to realize AI's capabilities while proactively managing associated risks.

Appendix 1: Supplemental Data

This appendix provides an accessible versions of the data for the complex figures in this document. Click "Return to original figure" below each table to get back to the original data figure.

FIGURE 5 SUPPLEMENTAL DATA

Deployment Location for the Development and Deployment of AI Models

	Public Cloud	On-Premises	Edge
AI model development	44%	53%	3%
AI model deployment	46%	49%	4%

Note: Public cloud includes public cloud provider, managed services provider, and special-purpose cloud. On premises includes private cloud, traditional infrastructure, and colocation provider. n = 411; Source: IDC's *AI Infrastructure Survey,* July 2024.

Return to original figure

FIGURE 6 SUPPLEMENTAL DATA

Perceived Costs of Public Cloud Versus On-Premises AI Models

	More Expensive	About the Same	Less Expensive
AI model development	23%	37%	40%
AI model deployment	25%	35%	40%

Note: Public cloud includes public cloud provider, managed services provider, and special-purpose cloud. On premises includes private cloud, traditional infrastructure, and colocation provider. n = 411; Source: IDC's *AI Infrastructure Survey,* July 2024

Return to original figure

FIGURE 9 SUPPLEMENTAL DATA

Virtualization of Private AI Environment

	Virtualization
In 12 months	35%
In 24 months	45%

Base: Respondents indicated organizations primarily develop/deploy AI models on premises. Note: The mean percentage of on-premises virtualization is shown. n = 220; Source: IDC's *AI Infrastructure Survey,* July 2024

Return to original figure

About the IDC Analysts

Madhumitha Sathish
Research Manager, High Performance Computing, IDC

Madhumitha Sathish is a research manager within IDC's worldwide infrastructure research organization and part of the performance-intensive computing (PIC) practice. She leads IDC's AI infrastructure research and plays a supporting role in IDC's research on high-performance computing (HPC) infrastructure stacks and deployments. In her role, Madhumitha delivers syndicated and custom qualitative and quantitative insights, including market size, forecast studies, and custom market models.

More about Madhumitha Sathish

Peter Rutten
Research Vice President,
Performance Intensive Computing, Worldwide Infrastructure Research, IDC

Peter Rutten is research vice president within IDC's worldwide infrastructure research organization and global research lead for the performance-intensive computing (PIC) practice. IDC's PIC coverage includes research on high-performance computing (HPC), Artificial Intelligence (AI) and Generative AI (GenAI), big data and analytics (BDA) and quantum computing (QC) infrastructure stacks, deployments, solutions, workloads, and use cases. It includes coverage of classical and hybrid quantum–classical supercomputing and institutional and mainstream HPC. Peter and his team take a keen interest in emerging infrastructure domains — including quantum, analog, and neuromorphic computing — that are highly disruptive to mature infrastructure markets. As a member of IDC's worldwide compute infrastructure research practice, Peter covers high-end, accelerated, in-memory, and heterogeneous computing infrastructure systems, platforms, and technologies. These include servers with discrete and embedded accelerators (e.g., GPUs, FPGAs, and ASICs) used in AI and HPC environments. In his role, he performs quantitative (market sizing and forecasting) and qualitative (primary research based) analysis as well as custom market sizing for IDC's clients.

More about Peter Rutten

Heather West, Ph.D.
Research Manager,
Performance Intensive Computing, Worldwide Infrastructure Research, IDC

Heather West, Ph.D., is research manager within IDC's worldwide infrastructure research organization and part of the performance-intensive computing (PIC) practice. She leads IDC's quantum, analog, and neuromorphic computing research and plays a supporting role in IDC's research on AI and high-performance computing (HPC) infrastructure stacks and deployments. Dr. West is deeply engaged with her clients on their solutions and services as well as on their business and technology strategies. Her domain knowledge of the quantum computing industry, including proficiency in related workloads and use cases, has made Dr. West a trusted advisor to several emerging quantum and analog computing vendors and positioned IDC as the go-to vendor for market research on quantum computing.

More about Heather West, Ph.D.

IDC Custom Solutions

IDC Custom Solutions produced this publication. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis that IDC independently conducted and published, unless specific vendor sponsorship is noted. IDC Custom Solutions makes IDC content available in a wide range of formats for distribution by various companies. This IDC material is licensed for external use and in no way does the use or publication of IDC research indicate IDC's endorsement of the sponsor's or licensee's products or strategies.

IDC Research, Inc.
140 Kendrick Street, Building B, Needham, MA 02494, USA
T +1 508 872 8200

idc.com @idc @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 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.