IDC

SPOTLIGHT

Sponsored by: Penguin Solutions

While the most experienced organizations favor dedicated on-premises infrastructures for their private AI environments, many organizations lack the necessary skill sets to implement and manage these systems effectively.

Powering Innovation: Private AI Infrastructure in the Enterprise

August 2024

Written by: Sean Graham, Research Director, Cloud to Edge Datacenter Trends

Introduction

Generative AI (GenAI) has captured the industry's imagination, evidenced by the fanfare surrounding foundational large language models (LLMs) exemplified by popular platforms like ChatGPT, swiftly followed by Google Gemini, GitHub Copilot, Claude, and others. With its ability to construct and predict complex scenarios, GenAI promises to revolutionize multiple industries. While GenAI has captured the public imagination, it represents just one facet of AI, which includes machine learning (ML), natural language processing, computer vision, robotics, and expert systems.

AT A GLANCE

WHAT'S IMPORTANT

Private AI offers numerous advantages, including enhanced data privacy, tailored solutions, compliance with regulatory standards, and cost control.

KEY TAKEAWAY

Implementing private AI requires significant changes to datacenter infrastructure, including cooling and power management, which takes specialized resources and skills.

The GenAI fanfare has heightened the interest and investment of enterprise CXOs who are keen to leverage this technology to unlock unprecedented productivity levels and generate new revenue by creating compelling digital experiences. With its ability to construct and predict complex scenarios, AI has the potential to revolutionize multiple industries. From generating hyper-personalized content in media and entertainment to facilitating drug discovery in healthcare to enhancing predictive maintenance in manufacturing, its transformative possibilities are staggering. AI is not uniquely the domain of hyperscalers and platform vendors. Numerous enterprises will create smaller specialized models and applications or tune foundational LLMs for their specific needs, with many choosing private AI environments. Still, in doing so, they will face challenges. AI requires new technologies and datacenter designs, which demand specialized resources for implementation and ongoing management.

Benefits of Private AI

Private AI uses enterprise datacenter infrastructure and AI framework capabilities to enable enterprise-specific AI/ML workflows. Whether datacenter assets are hosted in interconnection providers or in enterprise-owned facilities, private AI offers numerous advantages, including enhanced data privacy, tailored solutions, compliance with regulatory standards, and most importantly, control over AI initiatives.

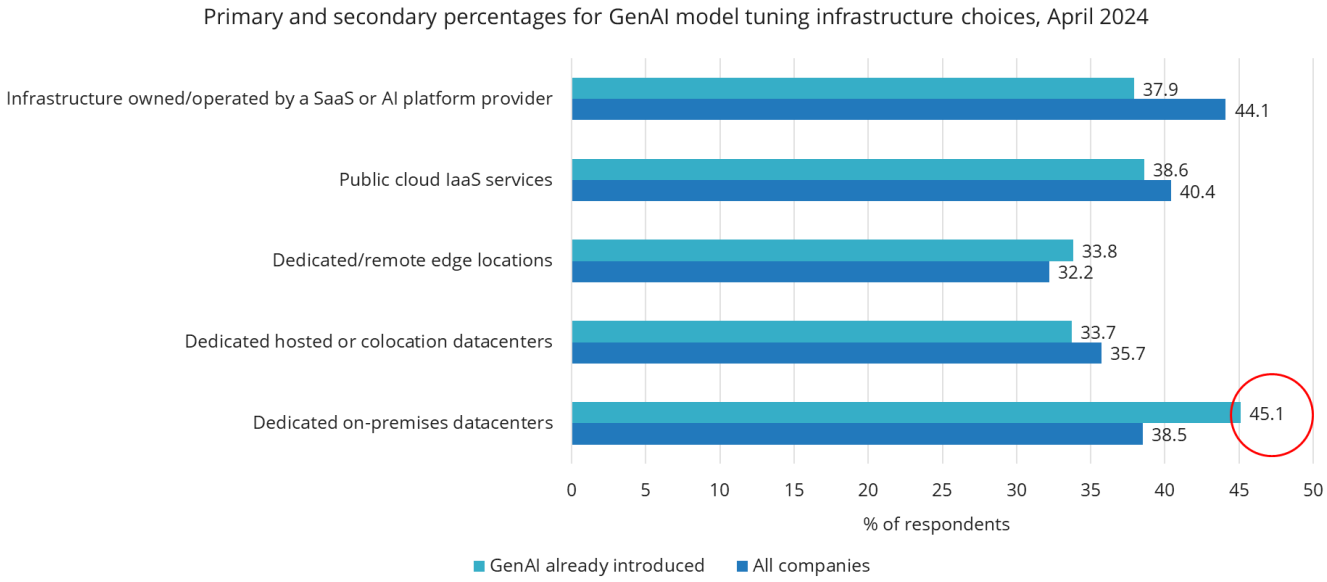
This control empowers enterprises to steer their AI projects in the direction that best serves their goals and objectives. In detail:

- **Competitive advantage:** Control over an AI infrastructure can provide a competitive advantage by enabling faster innovation and better integration with existing systems, allowing for the ability to leverage proprietary data more effectively. This can lead to improved decision-making and product development, enabling better strategic decision-making.
- **Customization and flexibility:** A private environment offers the flexibility to customize AI models and algorithms to suit the organization's unique needs. This can involve fine-tuning hardware specifics, networking configurations, and tailoring security policies to an organization's needs. Public providers often have limited customization, and therefore, integration with existing systems could require a major revamp of legacy setups or platform upgrades.
- **Data sovereignty:** The need for sovereignty, or the desire for greater control and ownership over data and infrastructure, is a major driver for private AI infrastructure adoption. While public cloud offerings provide robust security measures, some organizations, particularly those in highly regulated industries or with sensitive data, opt for private AI infrastructure to maintain sovereignty and ensure compliance with data protection regulations.
- **Costs:** Private AI can offer significant cost benefits compared with cloud-based solutions. Organizations can reduce ongoing subscription fees associated with cloud services by deploying AI on premises. Businesses can also repurpose their existing infrastructure for AI tasks, maximizing their current capacity and avoiding the expense of new facilities. For companies with large-scale data processing needs, this approach may be more economical over time than the recurring expenses of cloud usage.

Given these benefits, it is unsurprising that a growing number of organizations, especially those more experienced with GenAI, prefer on-premises datacenters. According to IDC's April 2024 *Future Enterprise Resiliency and Spending Survey, Wave 4,* the most experienced organizations prefer dedicated on-premises infrastructures for private AI environments (see Figure 1).

FIGURE 1: ***Computing and Storage Preferences for GenAI Tuning***

Q ***Over the next 18 months, what will be the primary type of computing and storage infrastructure used by your organization to support Gen AI model tuning? What will be your secondary approach?***



n = 889

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

Trends

While organizations have compelling business reasons for creating private AI environments, IT infrastructure and datacenter architecture differ significantly from general-purpose computing. AI requires performance-intensive computing (PIC) — servers, storage, and networking — that delivers greater performance and is achieved with high core count CPUs, coprocessors such as graphics processing units (GPUs), special interconnects, and high-speed networking (typically densely clustered). PIC factors organizations usually need to consider include:

- **Chip requirements (GPU, TPU, NPU):** AI is often enabled by accelerators such as GPUs equipped to provide the high-performance and efficient processing necessary to train these models. However, they correspondingly demand significantly greater power, which increases demands on the datacenter infrastructure (power and cooling). For example, a single NVIDIA H100 GPU draws 700W of power.
- **High-density racks:** The servers employed for AI can have configurations of up to eight GPUs per server. These attach rates can cause rack densities to escalate rapidly to 50kW per rack, which starkly contrasts with the industry average of 7–10kW per rack today. This necessitates either upgraded power distribution equipment to maximize

space or a change to the datacenter layout for those choosing to sacrifice physical space to circulate proper cooling to dispersed servers within their racks.

- **Liquid cooling:** Increased power density in AI systems leads to greater heat dissipation demands, while traditional air cooling falls short. This has led to a rise of alternative cooling methods (see Table 1), led by liquid cooling.
- **Modular designs:** Modular designs enable the rapid addition of upgraded capacity for datacenter operators through prefabricated, easily deployable units. They are ideal for evolving demands and essential for enterprises wanting to adapt to AI advancements while maintaining flexibility.
- **Electrical and mechanical equipment placement:** The advent of high-density racks in datacenters has led to a new design trend of placing cooling equipment, such as chillers, outside the datacenter, improving efficiency, reducing noise, and optimizing space. In addition, the strategic external placement of lithium-ion uninterruptible power supply systems offers a smaller footprint, longer life span, and higher energy density, further refining the datacenter's overall footprint and equipment needs.
- **Energy consumption and power scarcity:** With AI's intense power needs, it follows that energy consumption will rise significantly. IDC estimates that by 2027, AI energy consumption will reach 146TWh, growing at a CAGR of 42% from 2023 to 2027 and fueling fears of power scarcity.
- **Sustainability:** As businesses prioritize digitalization and environmental, social, and governance (ESG) strategies to drive business value, sustainability has emerged as the predominant challenge of the datacenter industry today. This becomes an even greater challenge with AI.

TABLE 1: ***Alternative Datacenter Cooling Methods***

Method	Description	Strengths	Weaknesses
Free air cooling	This method uses outside air to cool servers, often via fans pulling air through server racks.	The setup cost is low, and energy consumption is reduced if the climate is favorable.	Limited by geographic location, air quality can affect the component life span. It needs redundant cooling from alternative sources due to weather uncertainty.
Rear door heat exchanger	This method uses a door-like device attached to the back of server racks that uses water to absorb heat.	It's very efficient, especially in high-density areas, and it doesn't disrupt existing server or rack designs.	It requires significant up-front capital investment and limits on scalability as densities increase.
Single-phase liquid immersion	Servers are immersed in a nonconductive, single-phase coolant fluid such as oils, fluorocarbons, or synthetic esters, which absorbs heat.	It has high energy efficiency, greater server density, and an enhanced hardware life span. It is eco-friendly and operates quietly.	It requires extensive reworking of the datacenter (i.e., tubs on the floor with vertical rack designs, hoists, and heat exchangers) and comes with a risk of leaks.
Two-phase immersion	Servers are immersed in a bath of dielectric fluid that boils off to remove heat.	It's even more efficient than single-phase liquid cooling.	It requires extensive reworking of the datacenter (i.e., tubs on the floor with vertical rack designs, hoists, and heat exchangers) and comes with the risk of leaks and liquid loss.
Direct to chip	This method directly cools servers by pumping coolant to a cold plate that contacts components directly.	It has efficient heat dissipation, compact form factor, effective heat spreading, noise reduction, energy efficiency, compatibility with various chip designs, and scalability.	It requires a datacenter redesign, distribution and return mechanism liquid, and server redesign.

Considering Penguin Solutions

Penguin Solutions specializes in designing, building, deploying, and managing AI and accelerated computing infrastructures at scale. With over 25 years of experience in high-performance computing and more than 75,000 GPUs deployed and managed to date, Penguin Solutions positions itself as a trusted strategic partner for AI infrastructure solutions and services.

Penguin Solutions' Professional Services is tailored to expedite the design and deployment of new AI systems, ensuring immediate and efficient use with support from certified engineers who bring extensive experience in datacenter infrastructure, enabling successful implementation from small clusters to large supercomputers.

Penguin Solutions also provides managed services for AI environments, with more than six years of experience managing large NVIDIA DGX clusters as an NVIDIA Elite Partner. Its NVIDIA DGX DevOps and Services teams provide automated monitoring, ticketing, and regular improvements, ensuring seamless operations. The managed services team ensures enhanced system availability and stable operations for AI workloads, significantly improving ROI. Penguin engineers have been deploying and managing AI infrastructure since 2017.

Penguin Solutions' extensive AI infrastructure expertise helps reduce complexity and accelerate ROI, providing CEOs and CIOs with the essential and reliable infrastructures they need to deploy and manage demanding AI workloads at scale in the datacenter and at the edge.

Challenges

Despite Penguin Solutions' extensive expertise, the company faces hurdles in the AI market. The company will face competition from major cloud providers that are aggressively investing and expanding their AI and platform services. Cloud providers market themselves and are often perceived as a straightforward solution for AI infrastructure, offering low-risk, scalable options that demand minimal initial capital investment. The key for Penguin Solutions is to assist in developing a compelling business case that articulates the benefits of private AI known to the most experienced organizations and is supported by a measurable return justifying the investment in private AI infrastructure.

Conclusion

AI has emerged as a transformative force across industries, with its impact growing exponentially. As organizations seek to harness the power of AI, many are turning to private AI solutions to maintain control over their data, ensure compliance with regulations, and tailor AI models to their needs. Private AI offers numerous advantages, including enhanced data privacy, customized solutions, and potentially lower long-term costs compared with cloud-based alternatives.

In this complex and fast-paced environment, organizations need specialized services with deep experience in AI infrastructure. As AI continues to evolve and reshape industries, the role of expert partners in guiding organizations through implementing private AI infrastructures is increasingly vital.

AI requires new technologies and datacenter designs, which involve specialized resources for implementation and ongoing management.

About the Analyst

Sean Graham, *Research Director, Cloud to Edge Datacenter Trends*

Sean Graham is a research director for Cloud to Edge Datacenter Trends at IDC. He focuses on providing insights and analysis to the IT infrastructure vendors, datacenter and colocation providers, cloud service providers, and datacenter services firms.

MESSAGE FROM THE SPONSOR

Penguin Solutions designs, builds, deploys, and manages AI and accelerated computing infrastructures at scale. With 25+ years of HPC experience — and more than 75,000 GPUs deployed and managed to date — Penguin is a trusted strategic partner for AI solutions and services.

Designing, deploying, and operating "AI factories" is an incredibly complex endeavor. Penguin has successfully delivered AI factories at scale since 2017. Penguin's OriginAI infrastructure, backed by its specialized intelligent cluster management software and expert services, streamlines AI implementation and management, enabling predictable AI cluster performance that supports customers' business needs for clusters ranging in size from hundreds to thousands of GPUs.

Penguin's extensive AI infrastructure expertise helps to reduce complexity and accelerate ROI, providing CEOs and CIOs the essential and reliable infrastructure they need to deploy and manage demanding AI workloads at scale in the data center and at the edge. Learn more at: https://go.penguinsolutions.com/ai-consultation.

Visit us at: https://www.penguinsolutions.com/.

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
Twitter @IDC
blogs.idc.com
www.idc.com