Executive summary

Uptime Intelligence report: January 2025

Five data center predictions for 2025

In this report, Uptime Intelligence looks beyond the more obvious trends for 2025 and examines some of the latest developments and challenges shaping the data center industry. AI intensifies the already strong demand for IT but adds significant, new challenges for those designing, building and operating digital infrastructure. This is due to escalating power and cooling requirements, the need for complex retrofits, supply chain problems, high costs and technical uncertainty. These pressures are driving innovation and change in power distribution, cooling, and workload management, as operators seek to exploit opportunities against a background of grid limitations, sustainability pressures, and scrutiny. These challenges are being addressed with investment and enthusiasm: the industry continues to expand, fueled by optimism over the economic potential of AI and the development and deployment of transformative technologies.

Synopsis

The critical digital infrastructure sector continues to experience strong growth driven by surging demand for IT and the rise of AI. However, this growth presents significant challenges that will shape the industry in 2025 and beyond. This report explores these challenges and their implications, including the evolution of AI hardware, on-premises and cloud-based strategies for training and inference, innovations in power distribution, growing public opposition to new data center developments and the need to collaborate with power companies to address grid limitations.

Summary of the data center predictions for 2025

1. Data center resource use will raise deep questions — and opposition

Data center developments will become increasingly politicized in the coming years. Despite rising public opposition over environmental concerns and unmet promises of job creation, governments will support rapid expansion for the perceived economic and AI-driven benefits. As a result, climate commitments will be downgraded or postponed as optimism around AI continues to drive growth.

2. Most AI models will be trained in the cloud

Most investments in AI infrastructure for large-scale training will come from hyperscalers and cloud providers, as enterprises avoid the cost and complexity of on-premises GPU clusters. Enterprises will rely on public cloud services and pre-trained foundation models, fine-tuning them to reduce computational overhead.

3. Grid demand will require active participation from data centers

New and expanded data centers will increasingly be expected to provide or store power and possibly even shed loads to support grids. Data center operators running non-latency-sensitive workloads, such as specific AI training tasks, could be financially incentivized or mandated to reduce power use when required.

4. AI to trigger radical overhaul of data center electrification

Infrastructure requirements for next-generation AI will force operators to explore new power architectures. As a result, innovations in data center power delivery, such as deploying medium-voltage distribution to the IT space and solid-state transformers, will begin to emerge.

5. Nvidia's vision for data centers is not without alternatives

The performance of Nvidia's GPUs has led to the company's near-monopoly in the enterprise GPU market, but they are costly, scarce, and challenging to deploy. Some organizations will seek alternatives to these power-hungry GPUs, especially for inference tasks that require fewer computing resources. At the same time, there are signs that AI hardware will become more diverse in 2025.

Contents

Introduction

At the beginning of each calendar year, Uptime Intelligence compiles a short list of trends or predictions that will be relevant to the digital infrastructure sector for the year (and years) ahead. These lists aim to highlight crucial, yet often overlooked industry topics, which encourage a closer examination. Our five predictions for 2024 have proved to be largely accurate and will remain relevant for 2025. These are summarized in **Table 1**.

This year's predictions focus on the industry's rapid growth and its associated challenges, such as power availability, supply chain constraints, and rising costs. Unlike previous years, however, the global hype surrounding AI has thrust the data center industry into the public view. This attention is mostly due to the rapid growth of AI applications and the massive investments in AI infrastructure, which have captured widespread attention from the world's major media outlets.

The data center industry tends to be cautious about adopting new technologies, but is now experiencing unprecedented levels of investment driven by AI

The data center industry tends to be cautious about adopting new technologies, but is now experiencing unprecedented levels of investment driven by AI. This shift is occurring despite many unknowns surrounding capacity planning, power availability, supply chains and the performance requirements of increasingly densified IT racks.

Last year, Uptime Intelligence predicted that the impact of AI on the industry would be mostly limited to large-scale operators with the financial resources and infrastructure to support demanding AI workloads (see *Five data center predictions for 2024*). This prediction has predominantly borne out, with larger enterprise organizations better positioned to capitalize on the AI-driven demand than their smaller counterparts.

For example, the Uptime Institute Data Center and IT Spending Survey 2024 revealed that 84% of colocation providers plan to spend more in 2025 than the previous year. Similarly, 71% of enterprise operators expect their budgets for 2025 to increase, but many of these enterprises lack the financial flexibility to invest in capital projects, such as dedicated AI facilities, compared with their colocation peers.

AI projects face challenges: these include a limited supply of AI hardware, long lead times for power and cooling equipment, additional physical space requirements and shortages in skilled labor. Many enterprises are exploring alternative AI strategies, both on-premises and off-premises, while demand for general-purpose IT remains strong.

Additionally, as the industry expands, so too does its carbon footprint. Sustainability commitments once central to the sector are now under scrutiny, with some organizations scaling back or quietly retiring their ambitious carbon reduction goals.

This trend will likely fuel public opposition to new data center developments, particularly as the growing energy demands of AI put additional pressure on already strained power grids. Regulators, too, may face public backlash. Governments want to leverage the innovation and benefits of AI and data centers, but need to balance these goals with existing legislation around sustainability and energy security.

The coming year will test data center organizations' strategies as they navigate the opportunities and uncertainties presented by AI. Operators will need to balance this with maintaining existing service level agreements, achieving sustainability goals and meeting financial objectives.

Table 1

Predictions for 2024 recap

2024 prediction	Summary
Operators — prepare for a sustainability reckoning	New reporting laws will enforce stricter carbon accountability for data centers, challenging operators to prove their sustainability goals are realistic and evidence-based, often at high cost.
Demand for AI will have a limited impact on most operators	The AI boom will drive increased demand for power and cooling, with most data centers facing indirect impacts as they adapt to higher densities and varied resiliency tiers.
Data center software gets smarter, leverages data — at last	While operators are beginning to adopt software, connectivity and sensor technologies to optimize critical infrastructure, the market's complexity and risks remain significant.
Direct liquid cooling will not resolve efficiency challenges	Despite high expectations for direct liquid cooling, its slow adoption and the need to run mixed environments will limit its immediate impact on efficiency and sustainability.
Hyperscale campuses begin to redraw the data center map	Hyperscale colocation campuses, supported by wide-bandwidth fiber, will gradually reduce pressure on traditional data center hubs and eventually drive down colocation prices.

UPTIME INSTITUTE 2025

PREDICTION 2

Most AI models will be trained in the cloud

Key trends

- Most enterprises will use the public cloud for generative AI training, avoiding the cost and complexity of implementing and managing dedicated GPU clusters.
- Enterprises will use pre-trained foundation models to reduce computational overhead, customizing and fine-tuning them using short-term cloud services when necessary.
- Most organizations will compromise on their AI ambitions, choosing financial flexibility over greater control of the infrastructure, model and customization.

The rapid rise of generative AI has changed the landscape of AI infrastructure requirements. Training generative AI models, particularly large language models (LLMs), requires massive processing power, primarily through GPU server clusters. GPUs are essential in this task because they accelerate the processing of matrix multiplication calculations that underpin the neural network architectures behind generative AI (see *How generative AI learns and creates using GPUs*).

GPU clusters can be difficult to procure, expensive to purchase and complex to implement. Cloud providers offer access to GPU resources and AI development platforms on a pay-as-you-go basis.

The cost and complexity of deploying large-scale GPU clusters for generative AI training will drive many enterprises to the cloud. Most enterprises will use foundation models, pre-trained by third parties, to reduce computational overheads. Cloud services will be used for short-term and infrequent fine-tuning and customization tasks.

Creating and managing large-scale GPU clusters on-premises presents enormous challenges for enterprises. The financial burden alone is substantial: for instance, a single Nvidia H100 server can cost hundreds of thousands of dollars, and when setting up a functional AI cluster with even a few servers, the cost can reach millions. Other factors, such as storage, networking, power, cooling and labor, add significantly to the overall expense. Beyond cost, there are also operational complexities, as AI clusters require specialized data center infrastructure and teams of highly skilled engineers for maintenance, management, and troubleshooting. Additionally, supply chain issues continue to impact the availability of AI hardware, making it difficult for enterprises to acquire and deploy clusters quickly.

These challenges make on-premises training of large generative AI models feasible only for a few organizations that can justify the high initial investment and ongoing operational costs. Consequently, many enterprises seek more accessible, scalable and cost-effective ways of supporting their AI training needs. Cloud providers such as Amazon Web Services, Google Cloud and Microsoft Azure offer infrastructure as a service options that enable enterprises to access and utilize high-powered GPUs and other advanced AI infrastructure on a pay-as-you-go basis. A new breed of cloud providers, such as CoreWeave, have emerged to deliver large-scale GPU clusters as a service.

Hyperscalers also offer platform as a service (PaaS) options that enable access to AI capabilities without the responsibility of managing the model or the infrastructure. Pre-trained foundation models are also offered, reducing enterprises' burden of training.

Cloud providers are ideally positioned to fulfill the rising demand for AI training infrastructure

Balancing cost, flexibility and customization

Given the prohibitive costs and technical complexities of on-premises GPU infrastructure, most generative AI training will be forced to take place in the cloud for the foreseeable future. With their massive, centralized data centers and extensive GPU resources, cloud providers are ideally positioned to fulfill the rising demand for AI training infrastructure. These hyperscalers have made substantial investments in high-performance computing hardware, providing access to cutting-edge GPUs and new AI frameworks without capital expenditure.

By using these cloud services, companies can develop their own AI models from scratch without purchasing, installing or managing hardware themselves. They can also utilize other companies' models for AI capabilities. These cloud services and foundation models will not necessarily be cheap, but — for most buyers — they will be cheaper than dedicated equipment.

Foundation models, which have been pre-trained by software vendors or cloud providers, can further reduce computational overhead. Companies will use these models as the basis for small-scale customization or fine-tuning, ideally suited for the cloud, where training infrastructure can be consumed for short periods without upfront purchase.

Inevitably, some organizations will need to compromise on their AI ambitions if using a cloud service. Dedicated infrastructure provides full customization across all levels of the stack, while cloud services and foundation models provide general-purpose capabilities.

Some niche use cases, such as governmental and health care industries prioritizing security and compliance, or companies that rely heavily on proprietary AI capabilities, will require full customization. In these niche cases, the perceived risks of using shared infrastructure may outweigh the benefits of the cloud, prompting some organizations to maintain their own GPU clusters. But these cases will be the minority. For most, cloud-based models (whether fine-tuned foundation models or PaaS) will offer a "good enough" solution, balancing capability and cost without requiring massive investments in dedicated hardware.

Hyperscalers lead the way

Most investments in AI infrastructure to support large-scale training will be made by hyperscalers and cloud providers rather than by enterprises.

Training is a non-interactive batch job that demands large-scale, cutting-edge infrastructure. It can be performed without regard to proximity to model end-users. However, inference — the process of using the model in production — needs to be integrated with applications and provide a quick response to end-users. As such, inference will occur near where end-user applications are hosted, whether in the cloud, on-premises data centers, consumer devices or at the edge. Enterprises will continue to require infrastructure to support inference.

Over time, the cost of GPU clusters will likely fall due to more efficient hardware and improved supply. This reduction will change the dynamic slightly in that more enterprises might consider training their models using their own infrastructure. However, cloud providers and hyperscalers will also benefit from these cost reductions, offsetting the dedicated cost advantage.

Hyperscalers will likely continue investing in advanced AI infrastructure for training and inference, such as their own AI application-specific integrated circuits (ASICs), potentially lowering prices over time and improving service offerings. Cloud providers may also expand their portfolio of foundation models and pre-built AI services, making it easier for enterprises to integrate AI capabilities with minimal customization.

The GPU cloud market could consolidate as hyperscalers acquire specialist GPU cloud providers to meet enterprise demands (see *What is the outlook for GPU cloud providers*).

Easier access to AI can enable even small and mid-sized enterprises to leverage powerful AI capabilities without needing extensive in-house resources. However, reliance on the cloud also introduces some challenges, particularly around data sovereignty and regulatory compliance. For companies managing highly sensitive information, cloud-based training might require strategies such as data anonymization, which could reduce the quality or specificity of model outcomes.

PREDICTION 4

AI to trigger radical overhaul of data center electrification

Key trends

- Hardware for generative AI training is evolving in supercomputing style, pushing rack densities to heights previously unseen in mainstream facilities.
- The densification push is throwing data center designs off balance, with the same IT space requiring increasingly larger support infrastructure.
- Ongoing research and development into medium-voltage gear, solid-state transformers and other power delivery innovations will emerge in response to these challenges.

As we enter 2025, the IT and data center industry remains latched onto generative AI. The influence of a single type of workload reshaping an entire industry is unprecedented. The closest comparison is bitcoin, but this attracted specialist operators' development of "shadow" facilities, rather than a transformation in mainstream data centers.

While some remain skeptical about the economic viability and utility of massive generative AI models, many see the past two years as just the beginning. For now, supersized generative AI models are here to stay, placing onerous demands on infrastructure — both in terms of IT and data center facilities. Initially, the industry focused on solving cooling issues, but the more pressing challenge for next-generation AI infrastructure will be power, forcing operators to explore new electrification architectures.

The crushing forces of AI

Much of the technical direction of AI hardware development centers on densification. The current generation of AI compute racks in mass deployment (built around Nvidia H-series products) are typically around 40 kW per standard 19-inch rack — already well above typical (average) rack densities. Most organizations do not have a single high-density rack above 30 kW. There are even higher density deployment options for operators that want to compress cluster footprint and optimize cabling. But the current generation of AI compute racks (mostly) remain within the realm of standard power distribution equipment — notably busways and breakers.

However, AI hardware product roadmaps project densities to jump with each generation. Nvidia is spearheading this trend, pushing its generative AI hardware to adopt supercomputing-style system architecture, that prioritizes high-speed data sharing directly between its GPUs. This objective requires tight integration of many GPUs, linked via copper interconnects in a local mesh. This has a direct consequence for rack power density — more than any other factor.

Nvidia is on track to begin volume shipments of its first rack-scale GPU-systems in 2025. The high-end configurations with 72 GPUs per rack are rated at 130 kW with half-sized versions in the 60 kW to 70 kW range. Subsequent chip upgrades demand even greater power in 2025, with the 2026 generation projected to push thermal power ratings beyond 2 kW per GPU.

But the real power challenge that lies ahead will come from even tighter integration of compute chips. Nvidia's product roadmap calls for doubling the number of GPUs per rack — and then doubling again. These rack-scale systems, 300 kW and above, are slated to start shipping as early as 2026. These extreme levels of densities have only ever been seen before in the leading edge of supercomputing.

Data center power needs a step change

Demand for such AI training-focused rack systems remains uncertain. However, many operators and equipment suppliers consider hundreds of kilowatts of power per rack to be a likely scenario. While most organizations developing large AI models may not require the most advanced hardware, the AI market continues to be dominated by a handful of technology firms locked in an infrastructure development race. Even for mainstream organizations, typical AI racks are expected to start at around 80 kW, with many versions reaching the 100 kW to 200 kW range within a few years.

The real power challenge that lies ahead will come from even tighter integration of compute chips

The prospect of this level of power densities becoming widespread is forcing a rethink in the data center industry. An order of magnitude jump in rack power presents several challenges, particularly the proportion of space required for low-voltage distribution (automatic transfer switches, switchboards, UPS systems, distribution boards, batteries) relative to the technical space. Without changes to the power architecture, many data centers risk becoming electrical plants built around a relatively small IT room, where every aisle (or row even) requires a corresponding large UPS system, energy storage and other associated electrical equipment.

Another issue is the size and weight of the conductors delivering power to the IT space. Each row of future high-performance racks will have equivalent power demand to entire megawatt-scale data centers of the past, concentrating busways and cables into compact areas. This also mandates tight coupling of low-voltage (LV) electrical rooms and IT rooms to minimize costs and distribution losses. As the size of electrical rooms increases, site planning and design may become more difficult, especially when retrofitting existing buildings or operating in urban environments.

Transforming data center power systems

Major electrification suppliers are developing several product innovations that will help data center operators optimize their power distribution:

- Medium-voltage (MV) distribution to the IT space.
- Novel IT power distribution topologies.
- Solid-state transformers.

A key solution to the challenges outlined above is switching to MV UPS systems and downstream distribution — generally defined as any voltage level between 1,000 volts and 35 kilovolts. There are several benefits:

- Dramatically reduces conductor sizes and distribution losses.
- Helps site planning by enabling greater distances between the UPS rooms and the IT load, without severely impacting cost and losses.
- Compresses total electrical plant footprint.

Some MV UPS systems are currently available from select vendors, including both diesel-rotary and static — but choice is limited. For MV distribution to become more appealing, a larger number of established UPS vendors need to enter the segment, which Uptime Intelligence views as a likely outcome in coming years.

MV equipment is not new in data centers: MV engine generators and switchgears are common in multi-megawatt facilities. But as infrastructure power densification continues, the closer MV distribution is to the IT load, the better — providing segregation can be maintained between the MV and LV sides to comply with electrical safety codes. Potentially, future data center designs will incorporate novel power topologies that deliver MV power immediately next to power racks in the data hall to which IT racks connect.

MV power distribution designs are currently constrained by the need to install coupling transformers, which require considerable space and cannot be installed directly within the IT space. However, the development of sold-state transformers (SSTs) which is in progress at several power electronics suppliers for commercial deployment in a range of applications, may be a solution.

SSTs promise to be much smaller and lighter than a comparable core-based transformer, but also offer other power quality benefits. Furthermore, SSTs also make it possible to produce direct current output, simplifying IT power supply units and reducing total conversion losses in the power chain.

The overhaul of data center power chains will take several years to unfold, with 2025 shaping up to be a pivotal year. If investments into AI and data center infrastructure continue unabated, funding of the commercialization of MV UPS systems, SSTs and other innovations will be secure.

Find out more

We hope you found this executive summary of our Five Data Center Predictions for 2025 report valuable.

To access the entire Five Data Center Predictions for 2025 report and Uptime Intelligence on an evaluation basis, please visit intelligence.uptimeinstitute.com/request-evaluation

The full report is available to Uptime Institute members and Uptime Intelligence subscribers.

To enquire about an annual subscription, which includes this report; or to purchase this report, please contact info@uptimeinstitute.com

For information on becoming a Uptime Network Member, please visit: uptimeinstitute.com/ui-network

For media enquiries, please contact: publicrelations@uptimeinstitute.com

To discuss issues with the authors, please email: research@uptimeinstitute.com

For more information:
info@uptimeinstitute.com
+1 212 505 3030
Or contact a local Uptime Institute representative

All general queries
Uptime Institute
405 Lexington Avenue
9th Floor
New York, NY 10174, USA
+1 212 505 3030
info@uptimeinstitute.com

About Uptime Institute

Uptime Institute is the Global Digital Infrastructure Authority. Its Tier Standard is the IT industry's most trusted and adopted global standard for the proper design, construction, and operation of data centers — the backbone of the digital economy. For over 30 years, the company has served as the standard for data center reliability, sustainability, and efficiency, providing customers assurance that their digital infrastructure can perform at a level that is consistent with their business needs across a wide array of operating conditions. With its data center Tier Standard & Certifications, Management & Operations reviews, broad range of related risk and performance assessments, and accredited educational curriculum completed by over 10,000 data center professionals, Uptime Institute has helped thousands of companies, in over 100 countries to optimize critical IT assets while managing costs, resources, and efficiency.

Uptime Institute is headquartered in New York, NY, with offices in London, Sao Paulo, Dubai, Riyadh, Singapore, and Taipei.

For more information, please visit **www.uptimeinstitute.com**