uptime INTELLIGENCE

Executive summary

Uptime Intelligence report: July 2024

Uptime Institute Global Data Center Survey 2024

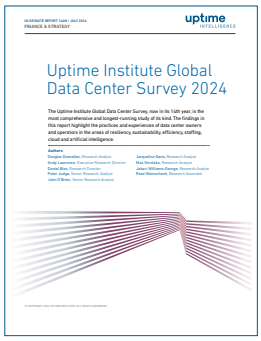
Now in its 14th year, this survey is the most comprehensive and longest-running study of its kind. This report highlights the practices and experiences of data center owners and operators in the areas of resiliency, sustainability, efficiency, staffing, cloud and artificial intelligence.

Synopsis

The Uptime Institute Global Data Center Survey 2024 reveals a confident, expanding industry but one that is also planning for major technological, economic and operational changes. Demand for digital services continues to grow — not just in volume but in compute intensity, challenging the power and cooling capabilities of much of the existing infrastructure. To meet rising demand, data center operators and their IT clients are investing and innovating more in their IT and facilities, as well as employing external services. The effectiveness of these investments will shape the industry in the years ahead.

Key findings

- Average PUE levels remain mostly flat for the fifth consecutive year, but this obscures advances in newer, larger facilities.
- Average server rack densities are increasing but remain below 8 kW. The majority of facilities do not have racks above 30 kW, and those that do have only a few. This is expected to change in coming years.
- Fewer than half of data center owners and operators are tracking the metrics needed to assess their sustainability and, in some cases, to meet pending regulatory requirements.
- Most operators recognize the benefits of AI and its potential. Despite many operators planning to host the technology, trust in AI for use in data center operations has declined for the third year in a row.
- The frequency and severity of data center outages remain mostly unchanged from 2023 or show small improvements. Operators are countering increases in complexity, density and extreme weather with investment and good management.
- Enterprises continue to meet their IT needs with hybrid architectures. More than half of workloads (55%) are now off-premises, continuing the gradual trend of recent years. Many continue to maintain their own data centers.
- Staffing challenges have neither improved nor worsened from 2023. More effort is needed to expand labor pools and skillsets to match the pace of capacity growth.



Report contents

Industry benchmarks

Average PUEs

The 2024 Uptime Institute data center survey tracks some of the key high-level operational and design metrics, such as facility energy performance, power density and server refreshes. Although each of these metrics has limitations, the data sheds light on large-scale trends.

In 2024, densified IT for generative AI and other applications is placing new demands on data center infrastructure and encouraging the use of new technologies, but not for all operators. As innovative IT and facility designs continue to grow in number, their influence on industry-wide averages will likely be apparent in a few years.

PUE: industry awaits a step change

Data center operators calculate PUE as a proxy for facility efficiency and a component of sustainability progress. PUE estimates the energy efficiency of a facility and helps track its change over time with a simple calculation: total facility power divided by power consumption of IT equipment. PUE was first defined by The Green Grid in 2007 and has since become the standard metric for facility energy efficiency.

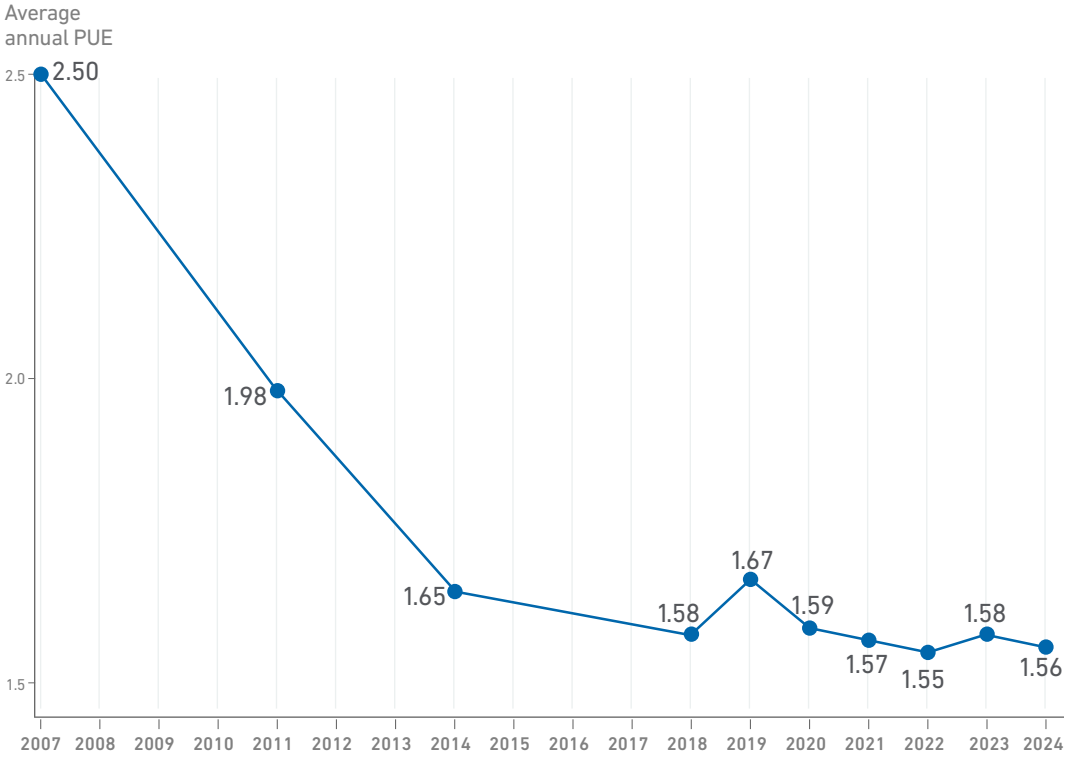
Key data center infrastructure characteristics that affect PUE can vary widely, including business objectives and climate — The Green Grid cautioned against making direct comparisons between individual sites for this reason. Trade-offs, such as supply temperature set points and water consumption, are outside of the metric's scope. Most importantly, PUE does not account for the energy performance of IT systems. The limitations of PUE as a useful metric will only grow as some future facilities specialize in denser IT architectures, often using direct liquid cooling.

Since 2007, Uptime Intelligence has been collecting average annual PUE figures from a large and diverse sample of data center operators. In the 2024 survey results, the industry average PUE of 1.56 (see **Figure 2**) reveals a continuing trend of inertia — although this headline number masks movements beneath the surface. While innovative facility and equipment designs are already demonstrating substantial efficiency gains and informing expectations for the next five years, their influence on average PUEs remains diluted. This is because of the large number of existing facilities worldwide, which include many aging legacy facilities.

Figure 2

Industry average PUE holds steady

What is the average annual PUE for the largest data center your organization owns / operates? (n=526)



UPTIME INSTITUTE GLOBAL SURVEY OF IT AND DATA CENTER MANAGERS 2007-2024

After rapid improvements in the industry average PUE between 2007 and 2014, progress lost momentum as the ratio approached 1.5. Data center designs have not approached physical limits of efficiency; nor have they standardized or become more similar to each other. Any gains in efficiency have been achieved by adopting relatively easier and more cost-effective measures, and these have largely run their course. Examples include the use of blanking panels, containment systems and variable frequency drives, as well as some relaxing of temperature set points.

For legacy data centers, more substantial upgrades are often cost-prohibitive and disruptive. Older facilities make up a considerable portion of the world's data center footprint — nearly half (47%) of respondents work primarily with a facility that is more than 11 years old.

New facility designs increase opportunities to optimize facility energy performance, and this is reflected in Uptime Intelligence's survey data. Many recent builds consistently achieve a PUE of 1.3 — and sometimes much better. With new data center construction activity at an all-time high to meet capacity demand, Uptime Intelligence expects these more efficient facilities to lower the average PUE in the coming years as their proportion in the survey sample grows. A key component to this shift is IT densification: a third of operators in our survey are developing new capacity to handle high-density cabinets.

Uptime Intelligence also expects an increase in demand for hyperscale colocation facilities through 2024 and beyond, as power, space and connectivity become strained in locations that have, up to now, been data center hotspots. This shift will likely drive further innovations and investments in both power and cooling infrastructure to support higher-density racks.

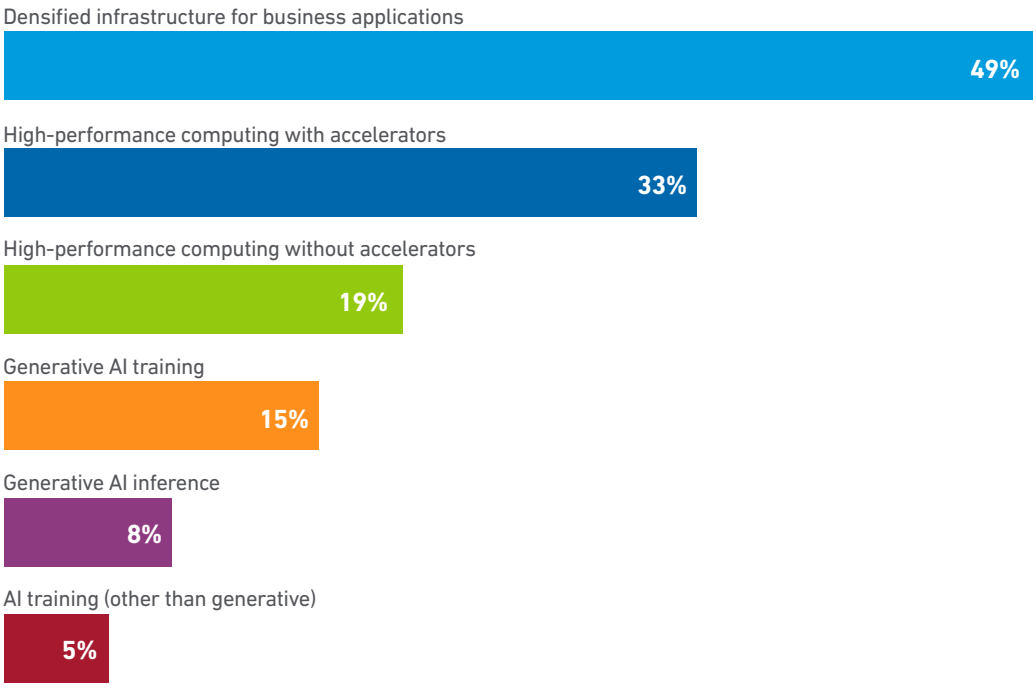
High-density workloads

Although many industry stakeholders anticipate unprecedented densification ahead of growth in generative AI — this is not the only workload that uses high-density IT. Uptime asked operators to classify the workloads supported by the densest deployments in their data center, and the outcome suggests generative AI has still to dominate (see **Figure 6**). The densest IT workloads supported today are still primarily business applications and high-powered computing (HPC).

Figure 6

Most dense IT runs business or HPC, not AI

Which of these workloads drive the highest density deployments in your data center? Choose no more than two. (n=711)



UPTIME INSTITUTE GLOBAL SURVEY OF IT AND DATA CENTER MANAGERS 2024

Just over half of respondents are supporting HPC: 33% using hardware with accelerators and 19% without. Further, nearly half (49%) pointed to business applications (such as high-performance, in-memory transaction processing workloads), while generative AI algorithms (training and inference) combined to only 23%. Uptime will continue to examine AI and other factors underpinning the industry's capacity and density decisions as they fluctuate.

Sustainability and metrics

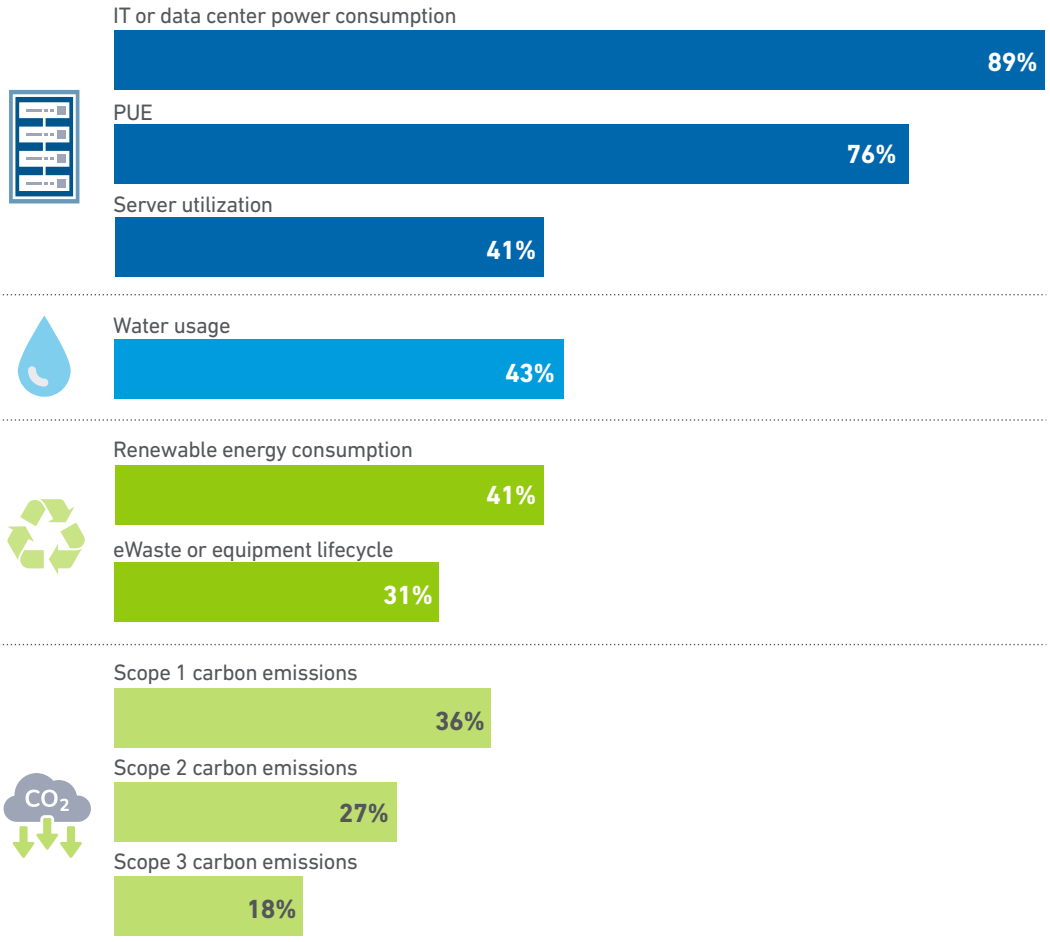
Uptime has reported for several years that the collection and reporting of sustainability-related data is patchy. Results from this year's survey are consistent with previous years, but a pattern is emerging. Operators are most able to report on just two well-established metrics, power consumption and PUE. However, these are not adequate in themselves to track progress towards sustainability.

There are obvious reasons why these two are the most reported: the data is collected easily and is of most interest to executives. The energy used or wasted has a direct impact on operational costs and improving efficiency has a direct impact on business performance and environmental impact. All other metrics relating to sustainability of facilities are reported by less than half of the survey respondents (see **Figure 7**).

Figure 7

Real sustainability metrics still lag behind PUE and power

Which of the following IT or data center metrics does your organization compile and collect for corporate sustainability purposes? Choose all that apply. (n=670)



UPTIME INSTITUTE GLOBAL SURVEY OF IT AND DATA CENTER MANAGERS 2024

Innovation and impact

Throughout 2023 and into 2024, the topic of AI and its potential impact on the future of business, politics and culture has continued to dominate the headlines. This has created both opportunities and challenges for data center operators, who welcome the additional demand for capacity but are now expected to host AI hardware that requires much more power and cooling than traditional enterprise IT. There are also opportunities to use AI to manage their own facilities.

The AI hype has prompted many businesses into experimenting with AI internally — and data center operators are no exception

Today, the industry is undergoing the largest speculative build-out of data center capacity in history. It is too early to know if current generative AI technologies will turn this abundance of specialized compute into business value. In the meantime, data center operators have been proceeding with older generation AI-based systems that have, up to now, proven suitable for mission-critical applications.

Operators are using more AI

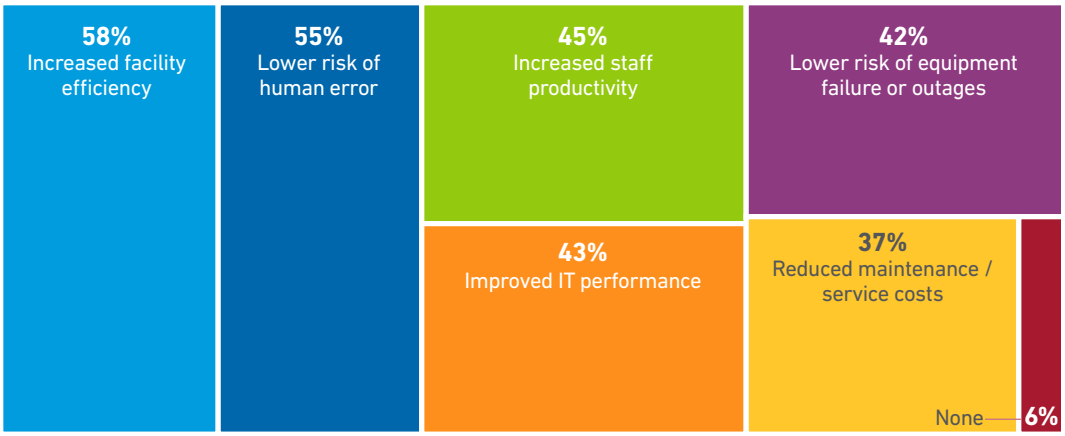
The AI hype has prompted many businesses into experimenting with AI internally — and data center operators are no exception. The number of operators that have deployed AI in production has been growing rapidly, and includes some of the world's largest colocation companies. Case studies detailing implementation of AI in operations are becoming more accessible.

The number of data center software vendors offering AI-based functionality in their products is increasing too, with AI-based cooling optimization in particular emerging as a distinct software category. According to the survey (see **Figure 8**), the three primary drivers that motivate AI deployments are the desire to improve facility efficiency (58%), followed by the need to reduce human error (55%) and as a means to improve staff productivity (45%).

Figure 8

Perceived benefits of using AI in operations

Which of the following — if any — do you consider to be benefits for using AI in your data center operations? Choose all that apply. (n=689)



UPTIME INSTITUTE GLOBAL SURVEY OF IT AND DATA CENTER MANAGERS 2024

Trust in AI continues to decline

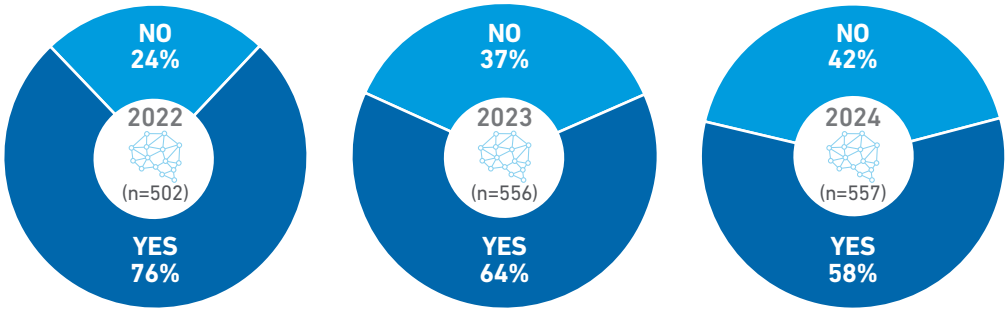
At the same time, operators' trust in AI as a tool for operational decision-making has seen its third year-on-year decline. The majority of respondents still say they would trust an adequately trained AI model to make operational decisions in the data center, but the size of this group has shrunk by almost 20 percentage points between 2022 and 2024 (see **Figure 9**).

Some of the negative aspects of AI typically cited include the lack of decision-making transparency and accountability, the cybersecurity risks introduced by additional network connections and the potential for AI-based control mechanisms to create additional points of failure.

Figure 9

Trust in AI dips for the third year in a row

Would you trust artificial intelligence (AI) to make operational decisions in a data center, assuming the AI has been adequately trained with historic data?



UPTIME INSTITUTE GLOBAL SURVEY OF IT AND DATA CENTER MANAGERS 2024

The findings relating to AI over the past three years are counterintuitive, given the dramatic increase in AI use generally. It appears that the more operators learn about AI, the less they trust the technology. Part of the problem is likely due to the quality and focus of much of the AI coverage, with highly publicized failures of generative AI systems throughout 2023 and continuing into 2024. The impressive results produced by simple, well-proven AI models in industrial settings are often ignored.

Perhaps this is a healthy degree of skepticism from an industry that has been misled before. This is not the first time a technology has promised to revolutionize data center management: similar transformational effects were once ascribed to DCIM, Internet of Things, digital twins, augmented reality and several other technologies — with only modest results.

The need to avoid outages at a site level and maintain IT service, despite the high cost, remains a critical issue for operators in 2024

Vendors confident AI will be widely used

Adoption of any new technology requires an ecosystem of hardware and software vendors. Nine out of 10 (91%) vendor respondents believe that it is likely that AI will be widely used in the data center in the next five years to improve operational efficiency and availability. This response has remained consistent over the past three years, suggesting that a new generation of AI-based products and services is under development and will arrive in the data center soon (in addition to the many products and services currently on offer).

Some of the new products will deliver value for money, but others will inevitably be promoted to ride the AI wave, confusing customers with machine learning terminology, yet offering no substantial benefit.

Resiliency and outages

The need for resiliency is well understood by all data center operators and across the supply chain. Although advances in IT, and software-based distributed resiliency, have offered the potential for operators to de-emphasize site-level resiliency, this has not happened. The need to avoid outages at a site level and maintain IT service, despite the high cost, remains a critical issue for operators in 2024.

Outage data can be challenging to track. Definitions of what constitutes an outage vary, as can measuring its severity and tracking the causes. Growing complexity stemming from increasingly interconnected facilities and IT systems can make outage impacts more widespread and difficult to diagnose. Under-reporting of outages, whether due to incomplete or undisclosed data, complicates outage analysis.

Uptime's survey data has been consistent over the years, showing gradual yet significant improvements in resiliency. This is partly due to improving management and processes; partly due to better maintained and monitored equipment; and partly due to consistent investment in facility resiliency, including redundancy (see *Annual outage analysis 2024*).

Building more resilient systems

Greater reliance on digital services boosts the business case for improving resiliency. Uptime survey data consistently shows that data center operators are investing more to increase site-level redundancy (see *Annual outage analysis 2024*).

Uptime expects distributed resiliency strategies to play an increasingly important role in mitigating the effects of outages in the coming years. With further investments in cloud-style application architecture and software-based approaches, these approaches will improve over time.

Four in five operators believe their most recent significant downtime incidents were preventable with better management, processes, or configuration

It can be argued, however, that resiliency efforts can also benefit most from operators improving training, processes and greater management attention on the importance of availability. Uptime's survey finds that four in five operators believe their most recent significant downtime incidents were preventable with better management, processes, or configuration — and this is consistent with previous years' data.

This data highlights the need for more testing and training, and a continued re-examination of existing systems and processes. There is also an opportunity to learn from the experience of previous outages, and from the industry's progress in adapting to an expanding risk landscape.

Cloud and provisioning

Off-premises locations dominate IT

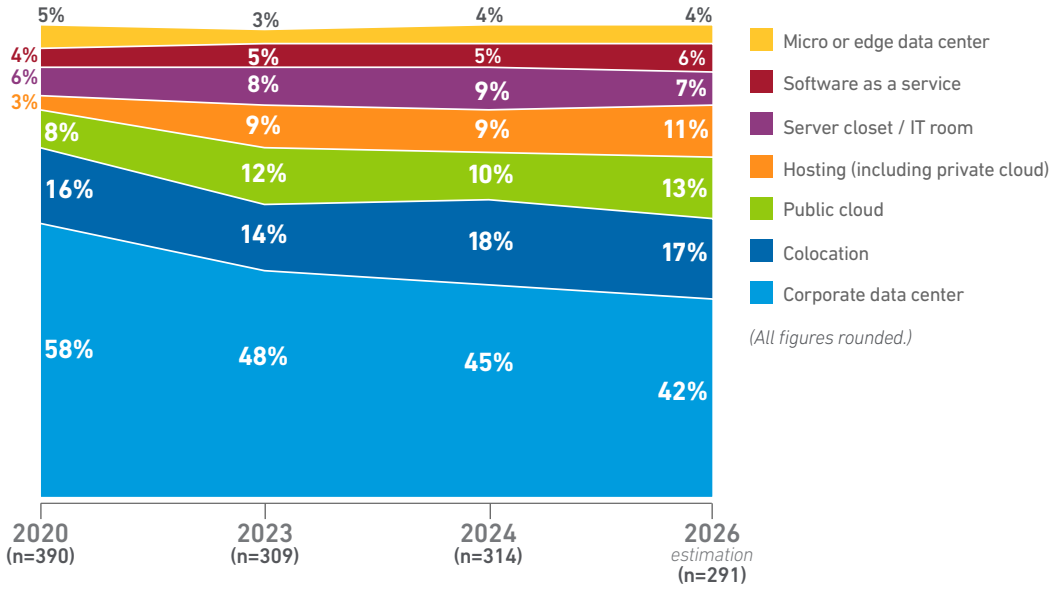
The proportion of IT workloads hosted remotely in off-premises facilities will continue to rise for most operators. In the 2020 Uptime Institute data center survey, respondents reported that, on average, 42% of their organization's IT workloads were hosted off-premises — in 2024, this percentage increased to 55%. Respondents forecast that 58% of workloads will be hosted in off-premises data centers by 2026 (see **Figure 13**).

Total IT capacities continue to rise overall — for both on-premises and off-premises environments. Well over half of owner operators (54%) and colocation providers (56%) cited capacity expansion as their main driver of spending increases, according to separate Uptime survey data (see *Most operators plan to spend more on rising demand*).

Figure 13

Colocation growth accelerates faster than other market segments

What percentage of your organization's total IT would you describe as running in the following environments today versus in two years from now?



UPTIME INSTITUTE GLOBAL SURVEY OF IT AND DATA CENTER MANAGERS 2024

Figure 13 shows a sharp increase in the proportion of colocation workloads between 2023 and 2024, compared with other data center environments. Respondents forecast that 18% of their workloads will be hosted in colocation facilities in 2024 and will change only slightly through 2026. This is a significant increase from last year's results when respondents forecast that 14% of their workloads would be in colocation facilities in 2025. Many of the cloud and hosting workloads are also ultimately housed in colocation facilities.

Uptime Intelligence's report on the growth of hyperscale colocation campuses (see *Hyperscale colocation: the emergence of gigawatt campuses*) backs up the observation that more IT and workloads are being placed in colocation facilities. The size and capacity of these hyperscale developments is unprecedented, as they take their place on the global data center map alongside established and fast-growing data center hotspots.

The proportion of workloads hosted in the public cloud is expected to be lower than previously anticipated. This year, respondents forecast that 13% of their workloads will be in the public cloud by 2026, compared with last year's expectation of 15% by 2025. Note that cloud and software-as-a-service providers make up a smaller proportion of the survey sample.

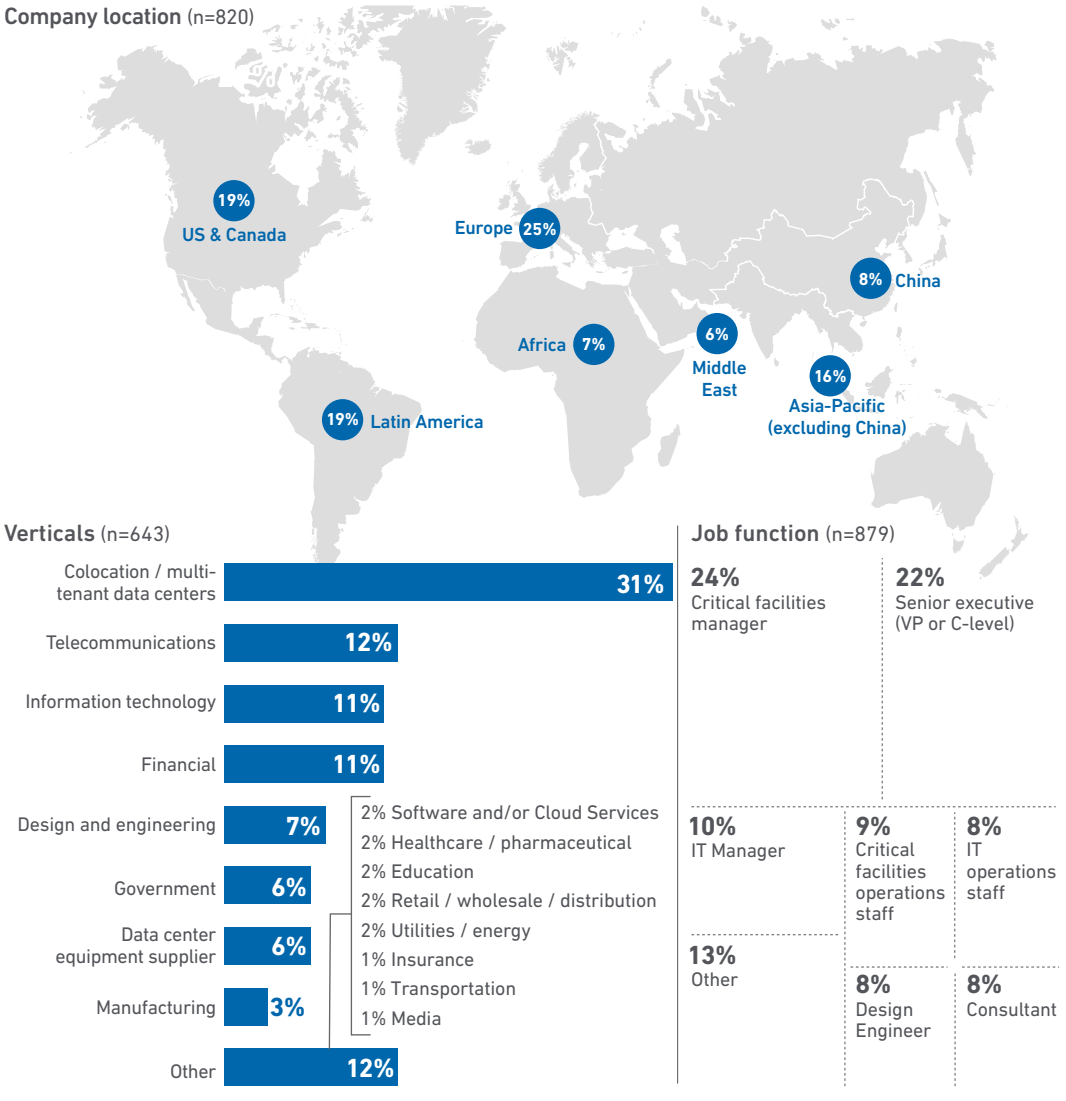
Survey methodology and demographics

Uptime Institute's Global Data Center Survey, now in its 14th year, is conducted annually online and by email. The 2024 survey was conducted in the first half of the year.

This report focuses on responses from the owners and operators of data centers, including those responsible for managing infrastructure at the world's largest IT organizations. Job titles include senior executive, IT manager, IT operations staff, critical facilities manager, critical facilities operations staff, design engineer and consultant.

Figure 17

Respondents by location, industry vertical and job function



UPTIME INSTITUTE GLOBAL SURVEY OF IT AND DATA CENTER MANAGERS 2024

Find out more

We hope you found this executive summary of our *Uptime Institute Global Data Center Survey 2024* report valuable.

Based on recent survey data and research, the full report examines the practices and experiences of data center owners and operators in the areas of resiliency, sustainability, efficiency, staffing, cloud and AI.

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 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**