uptime INTELLIGENCE

Uptime Institute Global Data Center Survey 2024

The Uptime Institute Global Data Center Survey, now in its 14th year, is the most comprehensive and longest-running study of its kind. The findings in this report highlight the practices and experiences of data center owners and operators in the areas of resiliency, sustainability, efficiency, staffing, cloud and artificial intelligence.

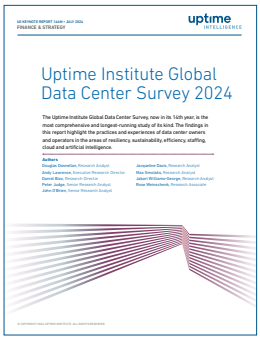
Authors

Douglas Donnellan, Research Analyst
Andy Lawrence, Executive Research Director
Daniel Bizo, Research Director
Peter Judge, Senior Research Analyst
John O'Brien, Senior Research Analyst
Jacqueline Davis, Research Analyst
Max Smolaks, Research Analyst
Jabari Williams-George, Research Analyst
Rose Weinschenk, Research Associate

Synopsis

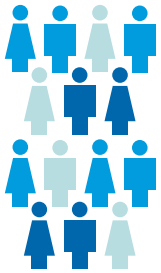
The Uptime Institute Global Data Center Survey 2024 reveals a confident, expanding industry but one that is also planning for major technological, economic and operational changes. Demand for digital services continues to grow — not just in volume but in compute intensity, challenging the power and cooling 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.

- 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.



Contents

Uptime Institute Intelligence is an independent unit of Uptime Institute dedicated to identifying, analyzing and explaining the trends, technologies, operational practices and changing business models of the mission-critical infrastructure industry. For more about Uptime Institute Intelligence, visit uptimeinstitute.com/ui-intelligence or contact research@uptimeinstitute.com.



Figures

Introduction

The 14th annual Uptime Institute Global Data Center Survey is the most comprehensive and longest-running study of its kind. The survey reveals the state of the industry in terms of resiliency, sustainability, efficiency, staffing, cloud and artificial intelligence.

The survey was conducted online from March 2024 to April 2024 and collected responses from more than 850 data center owners and operators, and over 750 vendors and consultants. This report focuses on owners and operators of digital infrastructure — an analysis of the experiences and views of vendors and consultants will be published separately. For more details, including demographics, see the **Appendix**.

For the second consecutive year, the survey asked operators to identify their digital infrastructure management team's key concerns. This year, additional response options relate to costs and the need to accommodate significantly denser IT.

Survey respondents indicate that cost is the primary concern for management — and by a significant margin (see **Figure 1**). Uptime Intelligence analysis from this and previous surveys suggests this is down to two main reasons: high prices and the need to invest to meet strong demand.

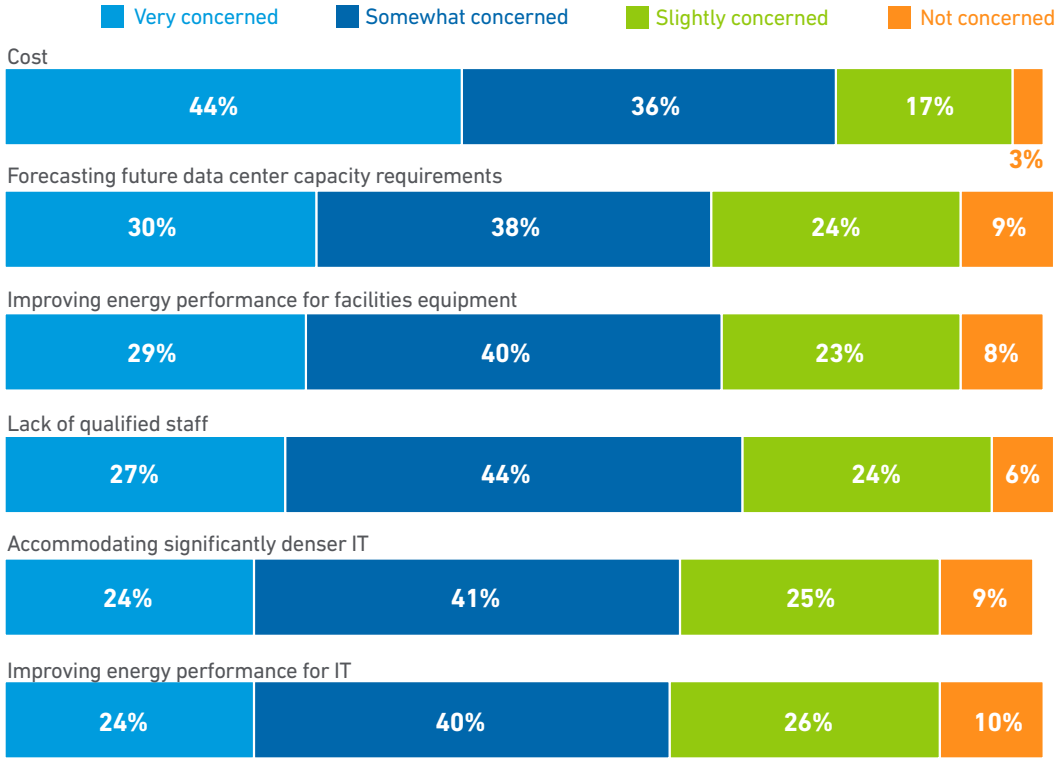
Owning and operating data centers has become increasingly expensive in 2024, continuing the upward trend since 2021. Although global inflation is slowing and some of the more severe supply chain issues are easing, these pressures remain. More than half of the survey's vendor respondents (51%, n=650) report higher than normal data center spending patterns. Prices have risen or are historically high for energy, equipment, labor, construction and infrastructure upgrades.

At the same time, there is continuing strong and growing demand for digital services (including AI). Across the sector, managers are weighing opportunities to expand and improve performance capabilities against budgetary limitations. Many are choosing to invest heavily to meet both current and expected rising demand.

Figure 1

Cost issues are the top concern for management in 2024

Looking at the next 12 months, how concerned is your digital infrastructure management regarding each of the following issues? (n=638)



UPTIME INSTITUTE GLOBAL SURVEY OF IT AND DATA CENTER MANAGERS 2024

Although AI and high-performance compute workloads garner significant media attention, their broader industry impact will likely take time to materialize. Managers also report that the need to accommodate higher density workloads is a less pressing concern than rising costs and the need for greater energy efficiency. While 65% of operators are at least somewhat concerned about their ability to accommodate rising densities, most are currently planning to meet the need for denser IT infrastructure by limiting their efforts to specific areas of their data halls.

Compared with 2023, other response categories show little change. About two-thirds of managers are at least somewhat concerned with forecasting future data center capacity requirements, a lack of qualified staff, and energy performance for IT and facilities equipment.

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.

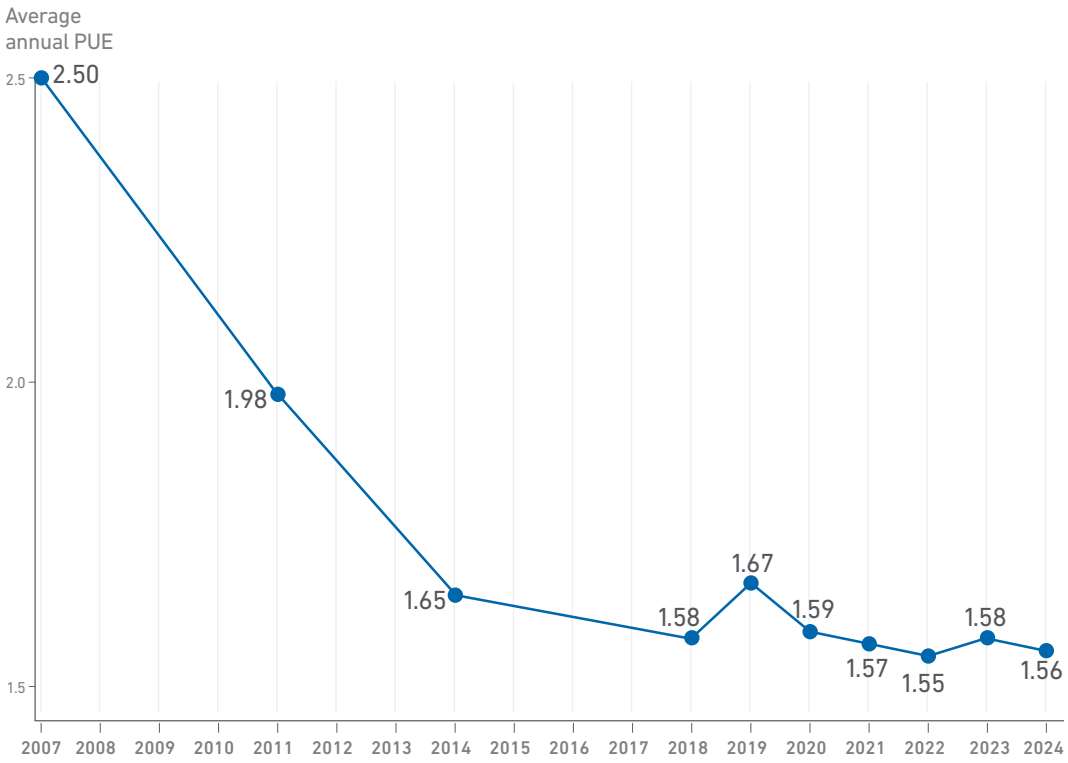
Key data center infrastructure characteristics that affect PUE can vary widely, including business objectives and climate — The Green Grid cautioned against making direct comparisons between individual sites for this reason. Trade-offs, such as supply temperature set points and water consumption, are 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)



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.

However, it cannot be assumed that PUEs will fall significantly. Some factors will, at least partially, offset industry average PUE gains. Climatic restrictions come into play as a growing share of development is being located in desert or subtropical climates. There is also an economic incentive to extend the lives of existing facilities, partly due to necessity: colocation capacity is scarce in many locations and equipment lead times are still lengthy. Furthermore, one in four data centers in our survey is typically under 40% utilized by available UPS capacity, which can undermine energy performance of both power distribution and cooling equipment.

Rack density — a steady climb

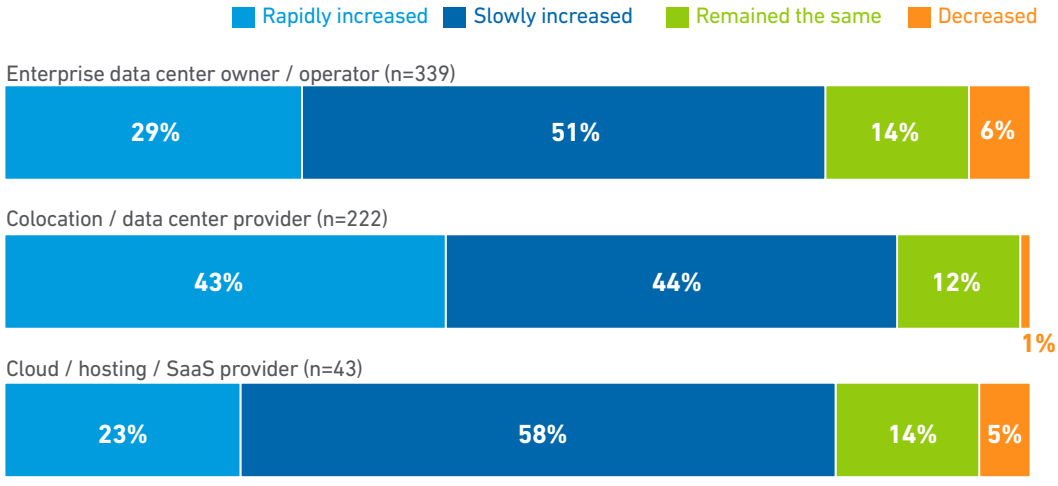
The 2024 Uptime Institute data center survey results continue to reflect a gradual shift toward more powerful racks. Typical rack power is edging up gradually, driven largely by the rise in silicon power — modern volume servers can use several hundreds of watts each when loaded, even without accelerators (such as GPUs).

Respondents to the survey also described the rate of change in rack density in their own facilities. In recent years, operators consistently reported sharp increases in rack power densities. This year is no different: nearly a third of respondents report rapid growth in rack power for recent deployments, with colocation deployments seeing the strongest uptick in rack power (see **Figure 3**).

Figure 3

Typical rack power increases have accelerated

Over the past three years, how has the most common (modal average) rack power density deployed in your data center changed?



(All figures rounded.)

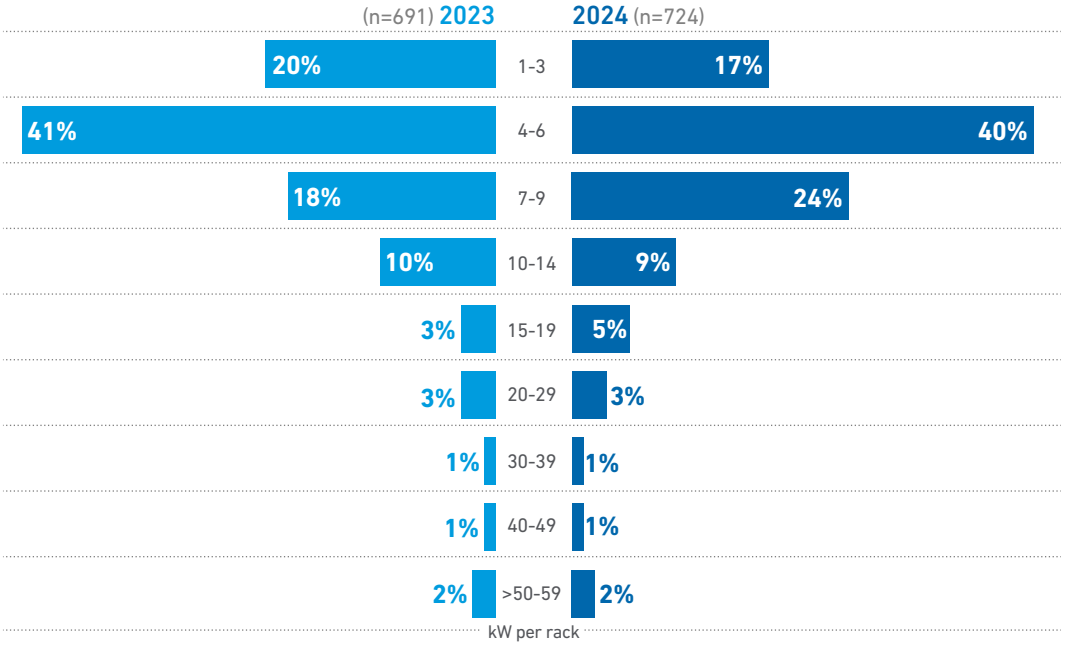
UPTIME INSTITUTE GLOBAL SURVEY OF IT AND DATA CENTER MANAGERS 2024

This increase is from a modest base and, in most cases, will not have major operational consequences. The 2024 survey shows 4 kW to 6 kW racks remain the most commonly deployed, consistent with previous years (see **Figure 4**). There is, however, a notable increase in 7 kW to 9 kW racks, with enterprise and colocation facilities largely reporting 7 kW rack deployments. This increase in higher-density racks marks a substantial shift from 2023 and aligns with the slow, but steady, rise in rack power densities observed in previous surveys.

Figure 4

Rack densities of 7 kW to 9 kW have become more common

What is the most common (modal average) server rack density deployed in your data center?



UPTIME INSTITUTE GLOBAL SURVEY OF IT AND DATA CENTER MANAGERS 2024

This upward trend is evident across enterprise, colocation and cloud / hosting segments. Also, a small base number underplays the significance of a major change: upgrading from 6 kW to 9 kW per rack, for example, represents a 50% increase, which can translate to a corresponding rise in power usage at many facilities. The density increase in the overall view was somewhat muted by some shifts in the makeup of the sample.

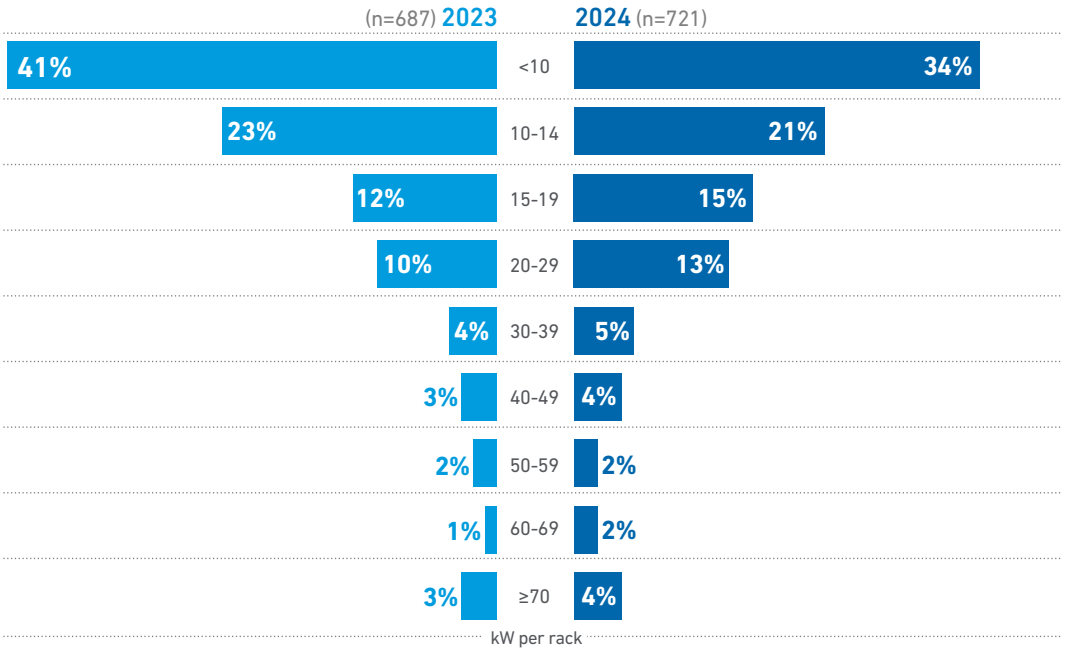
Uptime Intelligence calculates the average of typical racks densities across respondents to this survey to be 8 kW. However, this is skewed upwards by 11 sites that report their most rack density to be above 50 kW. If these sites are taken out as outliers, the average is 7.1 kW.

When asked about their highest rack densities deployed, survey respondents indicate modest increases across the density ranges above 15 kW, with a few reaching up to 100 kW or more (see **Figure 5**). These shifts can largely be attributed to the introduction of new high-powered server processors in 2022 and early 2023, which are now impacting the data center landscape. Uptime Intelligence expects this trend to accelerate in the coming years, as substantial shipments of dense GPU-servers (1 kW per rack unit or above) are installed and deployed across a range of applications.

Figure 5

Highest densities in the 15 kW to 29 kW range have become more common

What is the highest server rack density deployed in your data center?



UPTIME INSTITUTE GLOBAL DATA CENTER SURVEY 2024

Hyperscale and IT services data center operators are the dominant buyers of GPUs, yet make up only a small percentage of respondents. As these high-powered silicon components become available to a wider market, not only for use with generative AI models, Uptime expects reported peak densities to jump. However, it is important to note that the longer-term effect will depend on the development of viable workloads (use-cases) for these high-density GPU systems. The industry is still in the initial stages of this technology and economic cycle, and there is considerable uncertainty with how it will develop.

Delivering power and cooling to very high-density cabinets in existing facilities is challenging for many operators. To meet the needs of high-performance IT systems, more than one in four (29%) of operators are upgrading their existing data hall space. This suggests that rapid densification is the reality for many operators, and they need to react — even within the constraints of a retrofit.

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.

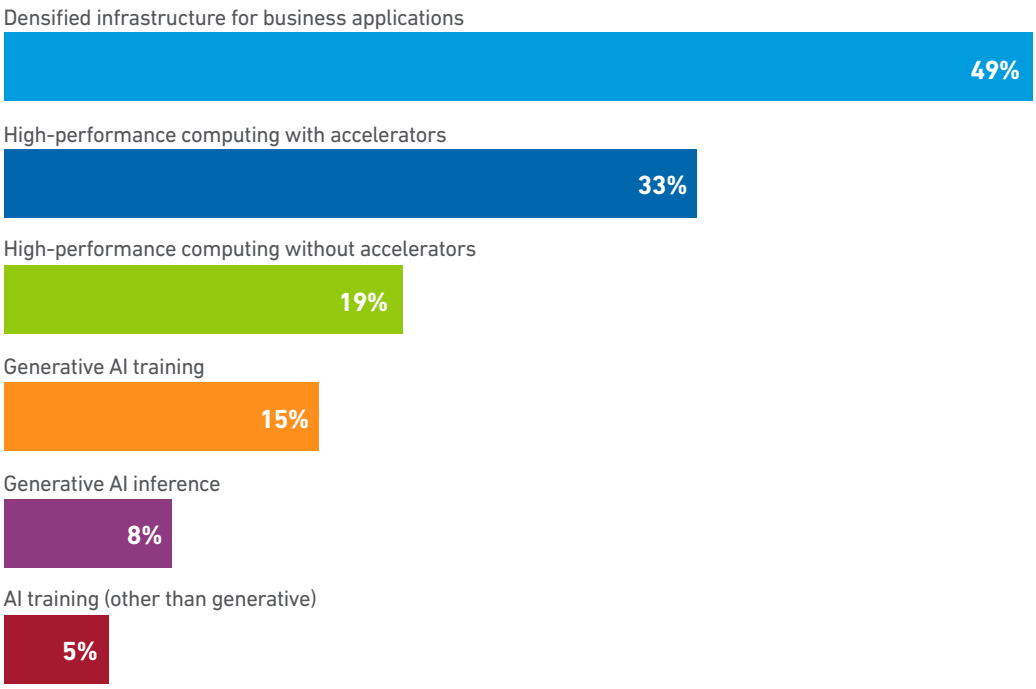
High-density workloads

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

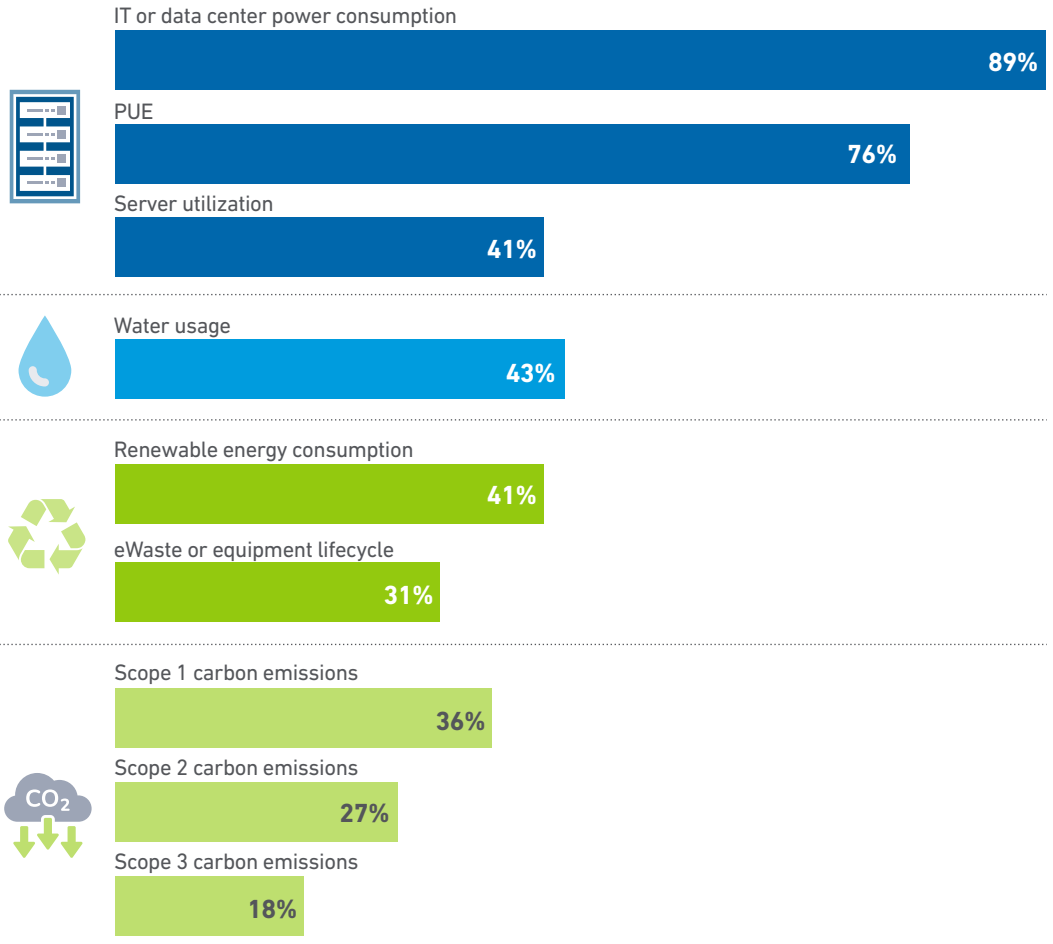
Uptime has reported for several years that the collection and reporting of sustainability-related data is patchy. Results from this year's survey are consistent with previous years, but a pattern is emerging. Operators are most able to report on just two well-established metrics, power consumption and 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

This is concerning because several of these measurements will be required by regulation that is already passed or pending. Notably, the EU's Energy Efficiency Directive (EED) will, either this year or for 2025, require operators to report their renewable energy consumption and water usage (both potable and non-potable), which are currently collected by only slightly more than 40% of respondents.

Collecting the data for these metrics can be difficult because:

- There are no standardized methods for generating and capturing IT utilization data and reports are likely to be estimates — a fact that the EU acknowledges by also asking operators to report their level of confidence with their reported figures.
- Water usage figures are also hard to assess because some data centers will consume both potable and non-potable water, and return some of it in varying conditions.
- Renewable energy consumption is complicated since most data centers use some form of energy attribute certificates (EACs) to offset the carbon intensity of the electricity they use, but the use, quality and acceptability of different types of EAC is currently being debated.

Greenhouse gases are still under-reported

Only a minority of organizations report on carbon emissions under the established Greenhouse Gas (GHG) Protocol

Only a minority of organizations report on carbon emissions under the established Greenhouse Gas (GHG) Protocol, despite this being an essential part in substantiating claims of progress to carbon neutrality. This data is also required under climate reporting laws in EU, the UK, many Asian countries and parts of the US. This survey's finding suggests that the majority of operators do not have the data to either make these submissions or backup their corporate net-zero goals.

Under the GHG Protocol, carbon reporting is divided into three categories: Scope 1 (direct emissions), Scope 2 (emissions embedded in electricity) and Scope 3 (emissions attributable to the supply chain and embedded in products). Just over a third of respondents are reporting Scope 1 emissions, which includes emissions from diesel fuel burnt on-site for testing backup power systems and makes up just a small proportion of total emissions.

Scope 2 shows the greatest increase of any metric, jumping from 19 to 27 percentage points — an increase of 40% of respondents. These emissions can be minimized by reducing energy use, and by switching to lower-carbon electricity sources.

Scope 3 involves the collection of upstream and downstream emissions associated with supply chains (including transport), construction, and manufacturing of equipment, as well as the use and end-of-life processing of equipment used in the data center.

Scope 3 reporting, at 18%, remains low and reflects the greater difficulty in collecting the relevant data. Discussions are ongoing about how to collect what data and, perhaps more importantly, what the benefits are in collecting this data. However, the requirement to collect Scope 3 data is established, in spite of these issues (see *Scope 3 accounting: once is not enough*).

Some of those data center operators reporting Scope 3 emissions have made a striking observation: Scope 3 emissions can be as large as, or larger, than emissions from Scopes 1 and 2. However, this reveals the idiosyncrasies of carbon reporting: by buying offsets and/or using renewable energy, the Scope 1 and 2 emissions of these operators can potentially be eliminated — even if they primarily use fossil-fuel powered electricity for their data centers.

More work needed

Basic power consumption and PUE data has been the basis for data center efficiency reporting for several years, but is not adequate for the level of reporting that is increasingly being required by both regulatory authorities and customers.

In 2023, Uptime Intelligence predicted that reporting of sustainability metrics would increase rapidly. This has yet to happen, but legislation and public pressure will continue to demand this information.

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

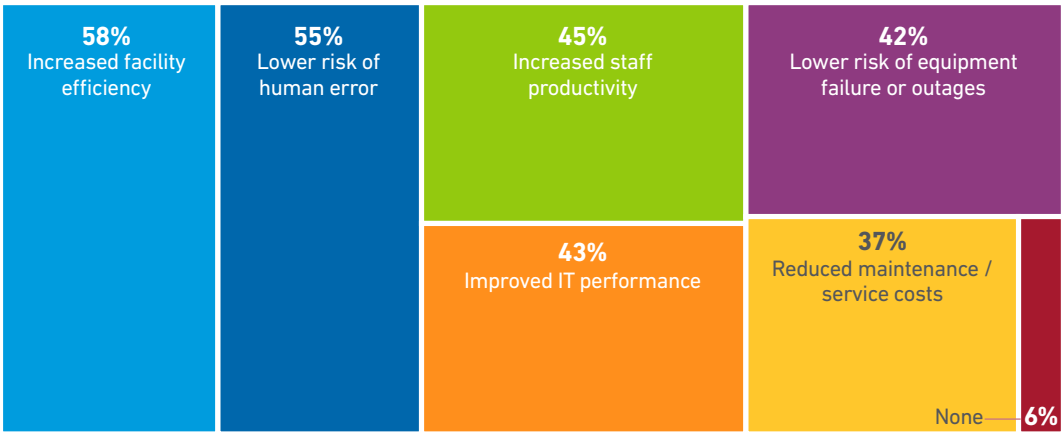
The AI hype has prompted many businesses into experimenting with AI internally — and data center operators are no exception. The number of operators that have deployed AI in production has been growing rapidly, and includes some of the world's largest colocation companies. Case studies detailing implementation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

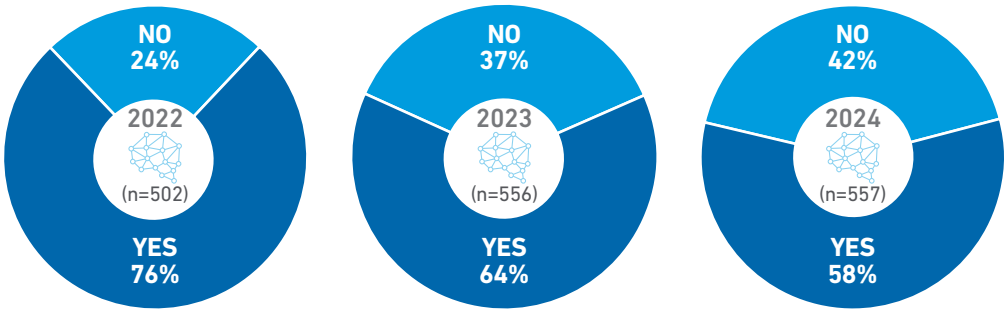
At the same time, operators' trust in AI as a tool for operational decision-making has seen its third year-on-year decline. The majority of respondents still say they would trust an adequately trained AI model to make operational decisions in the data center, but the size of this group has shrunk by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

However, resiliency improvements are not guaranteed. Aging utility infrastructure means generators and UPS may be tested more. In addition, extreme weather events are increasing and the surging demand for new capacity is adding complexity and reducing ride-through times. As the industry expands further, addressing these issues will be crucial to maintaining and enhancing the stability and reliability of data center operations.

Outage frequency and severity

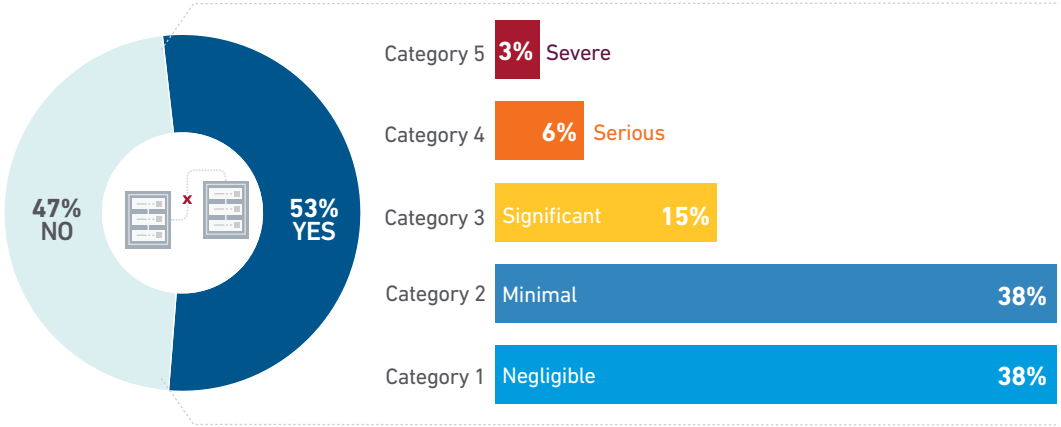
The growing importance of digital services and the expanding global data center footprint suggest that outages are becoming more widespread and impactful. While this is true in absolute terms, it is not the case relative to the overall rise in IT. Uptime data shows that, for the fourth year in a row, the frequency and severity of outages is decreasing on a per-site basis.

In the 2024 Uptime Institute data center survey, more than half (53%) of operators say their organization experienced an outage in the past three years (see **Figure 10**). While this figure has fallen significantly from 60% in 2022, 69% in 2021 and 78% in 2020, the year-on-year improvement has slowed to just two percentage points from 2023 (55%).

Figure 10

Most outages in past three years had little impact

Has your organization had an impactful outage in the past three years? If so, how would you classify the most impactful outage on a scale of 1 (negligible) to 5 (severe)? (n=768)



(Responses for "Don't know" are not included.)

UPTIME INSTITUTE GLOBAL SURVEY OF IT AND DATA CENTER MANAGERS 2024

Large outages can make headlines and have major consequences for the organization, but Uptime's surveys reveal that most outages have a limited impact. Of those who experienced an outage, only 9% report that it was either serious or severe — the lowest reading recorded by Uptime so far.

There are many causes of outages (see **Figure 12**), and factors that drive overall industry outage rates. The data suggests that the overall rate is stabilizing, with serious / severe outages at their lowest ever level, justifying the high priority and investment that availability has been given over the past two decades.

The cost of outages

When outages occur, they are often expensive. For the second consecutive year, 54% of respondents say their most recent, significant outage cost more than $100,000 (see **Figure 11**). There has also been a marginal (four percentage points) increase from 2023 in those reporting an outage costing more than $1 million.

Figure 11

One in five impactful outages cost more than $1 million

Estimate the total cost of this downtime incident (from outage to full recovery), including direct, opportunity and reputation costs. (n=91)

(Responses shown are only from those who reported experiencing either a significant, serious or severe outage in the past three years.)
(All figures rounded)

UPTIME INSTITUTE GLOBAL SURVEY OF IT AND DATA CENTER MANAGERS 2024

Many persistent factors have contributed to these high costs in recent years, including inflation, penalties for breaching service level agreements (SLAs), labor costs, and expenses from replacing and installing hardware. The increasingly critical nature of digital services is likely to drive these costs even higher: more people and services are dependent on digital services and the impact from a major outage is therefore wider and more consequential.

In recent years, Uptime has identified a trend referred to as "creeping criticality", where services and applications become more important with increased usage and are embedded in business and social practices. Expectations from consumers, businesses and regulators have become more stringent and this is reflected in stronger SLAs with harsher financial penalties for outages and, in some cases, fines from regulators. These higher costs associated with outages will continue to validate investment in maintaining service availability.

The causes of outages

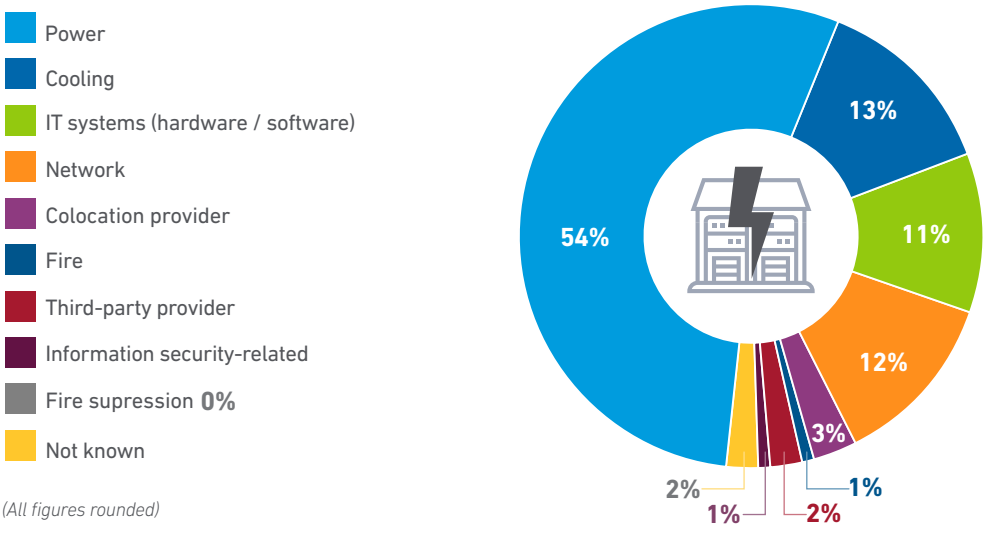
Establishing the root causes of system failures is crucial for preventing future downtime and identifying areas requiring investment.

Uptime data consistently show that disruptions to on-site power distribution are the most common cause of impactful outages, accounting for more than half (54%) of these outages in 2024 (see **Figure 12**). As discussed in previous Uptime Intelligence reports, challenges with electrical grids, exacerbated by aging infrastructure, rising demand, severe weather events and a reliance on intermittent renewable energy sources, may worsen this trend.

Figure 12

Power is still the leading cause of impactful outages

What was the primary cause of your data center's most recent impactful incident or outage? (n=97)



UPTIME INSTITUTE GLOBAL SURVEY OF IT AND DATA CENTER MANAGERS 2024

Taken together, failures stemming from IT and networking systems increased by eight percentage points compared with 2023 and now account for approximately one in four (23%) impactful outages. This is partly due to the increasing complexity of networks, and the growing role of software-defined networking and distributed IT to maintain services if one data center is unavailable or experiencing problems. In effect, the improving reliability to single sites may be highlighting problems in the distributed, multi-site approach.

Cooling failures can generally be tolerated due to thermal ride-through, which delays the immediate impact of any problem. Ride-through times have become shorter in recent years, however, as density increases, outside temperatures rise, and (in some cases) as data hall set points are increased. Continuing increases in density could result in more failures in the years ahead, although there is some evidence that more cooling systems are being supported by a UPS, and greater redundancy is being used (see below and *Annual outage analysis 2024*). The reliability of liquid cooling is, as yet, untested at scale in the field.

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

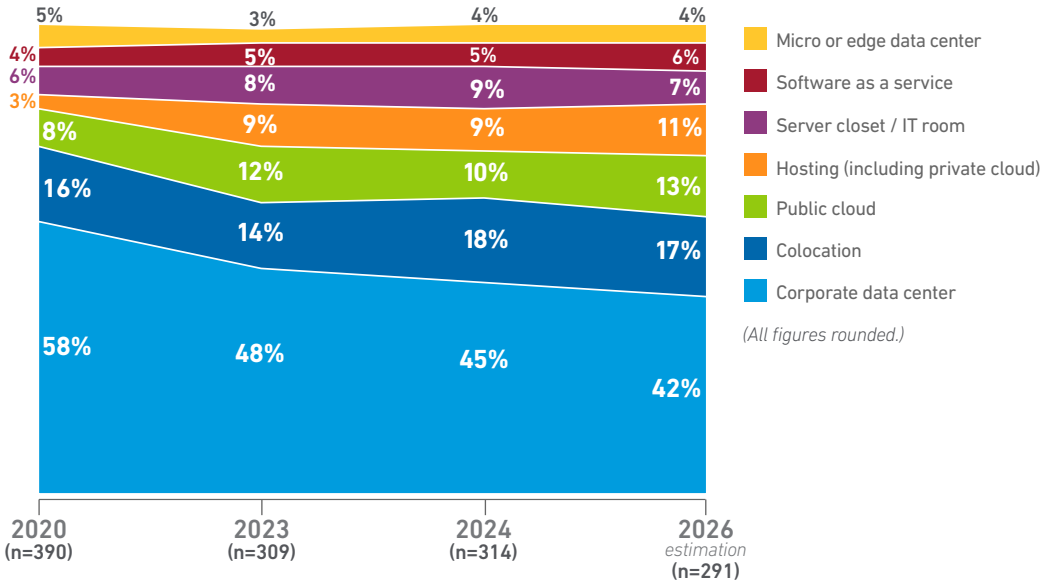
The proportion of IT workloads hosted remotely in off-premises facilities will continue to rise for most operators. In the 2020 Uptime Institute data center survey, respondents reported that, on average, 42% of their organization's IT workloads were hosted off-premises — in 2024, this percentage increased 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.

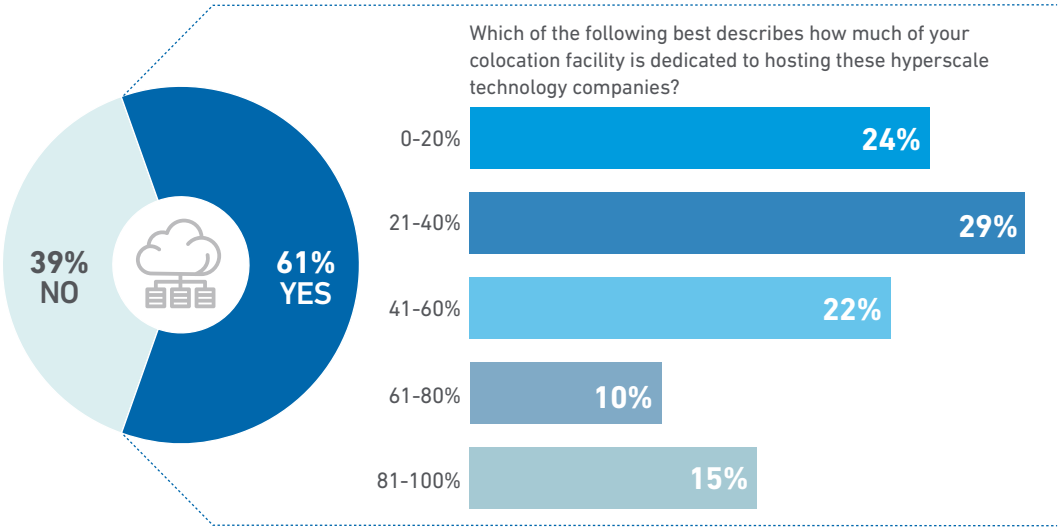
The proportion of workloads hosted in the public cloud is expected to be lower than previously anticipated. This year, respondents forecast that 13% of their workloads will be in the public cloud by 2026, compared with last year's expectation of 15% by 2025. Note that cloud and software-as-a-service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

This observation is backed up by Uptime research, which reveals that some operators are rebalancing their usage of public cloud, while repatriating certain workloads back on-premises. Increasingly, larger enterprises are replacing "cloud first" strategies with "cloud appropriate", where operators use a combination of on-premises, cloud and colocation provisioning, according to the specific workload requirements (i.e., hybrid IT). The reasons for this rebalancing could include cost, complexity or regulatory requirements. However, very few operators, if any, are moving away from the public cloud completely (see *Capacity expands rapidly, but complexity is challenging*).

Figure 14

Nearly two-thirds of colocation providers host hyperscale tenants

Does your colocation facility host any of the following hyperscale technology companies: AWS, Google, Microsoft, IBM, Meta, Tencent, Alibaba or Baidu (n=182)? If so, how much of your facility is dedicated to hosting these companies?



UPTIME INSTITUTE GLOBAL SURVEY OF IT AND DATA CENTER MANAGERS 2024

Once threatened by public cloud vendors, colocation providers are being increasingly relied upon to host larger amounts of the cloud IT footprint. Almost two-thirds (61%) of colocation providers currently host hyperscale tenants (see **Figure 14**). More than half (53%) dedicate up to 40% of their facility to hyperscale technology companies, and a quarter dedicate most of their facility. This demand by hyperscale tenants is undoubtedly supporting the continued expansion of some large colocation campuses.

Many hyperscale cloud providers rely on colocation partners to access new markets or expand within existing regions. This helps them provide a service to customers more quickly and economically than building new sites, which, in many cases, can take years (although many are also building new sites in parallel). Several colocation operators have reported witnessing a significant increase in pre-leasing activity to hyperscale providers, to support cloud and AI services.

Half of operators use on-premises cloud infrastructure

As hybrid IT provisioning becomes more widely used by data center operators, Uptime observes more use of on-premises private cloud infrastructure offered by major public cloud vendors. This involves the cloud provider running dedicated infrastructure and services in the operator's facility. The aim is to offer the flexibility and scalability of public cloud services, but under the control of the operator, rather than the cloud provider.

Just under half of the survey respondents (44%) report using on-premises private cloud infrastructure. Colocation providers were the largest category of hosts for on-premises private cloud infrastructure (30%), followed by telecommunications and financial services (12%).

On-premises private cloud infrastructure is typically a combination of hardware, software and support provided as an integrated offering or managed service from the cloud provider. They extend cloud services, such as compute, storage, networking, security and, in some cases, platform-as-a-service development environments, into the on-premises data center. On-premises private cloud infrastructure services can be either completely or partially disconnected from the internet and public cloud.

In Uptime's annual survey, data security (60%) and regulatory and compliance (44%), consistently stand out as the top reasons for operators choosing not to host mission-critical workloads in the public cloud. The EU's general data protection regulation (GDPR), for example, puts restrictions on personal data being moved outside of the country or the EU region.

Staffing

Data centers still seeking staff

Data center operators need to attract and retain workers for crucial roles, as they have for more than a decade. Targeted recruitment initiatives have proliferated during this period of vigorous growth in data center capacity and competition for skilled resources, but these efforts have yet to meaningfully shrink a high and long-standing vacancy rate. In 2024, half of operators (51%) in our survey report difficulty in finding qualified candidates to fill their job openings. This is the third consecutive year that this figure has not risen — but neither has it fallen.

With the data center industry still exhibiting strong capacity growth, even stabilization of skills shortages suggests some success in hiring and retention efforts

With the data center industry still exhibiting strong capacity growth, even stabilization of skills shortages suggests some success in hiring and retention efforts. However, industry stakeholders also describe significant planned capacity that is paid for but cannot yet be built. The reasons for this include lack of available power, difficulty obtaining environmental permits, or tight supply chains. The persistent shortage of skilled staff may be another factor impeding data center construction, where organizations delay construction of facilities because they do not have sufficient staff to operate them effectively.

Operators diversify strategies

Faced with acute skills shortages in the data halls, operators will draw from all available labor pools, using a variety of strategies. Historically, many organizations have opted to increase salaries for operations teams as a straightforward way to retain their skilled staff. Survey data suggests this strategy motivates workers to stay in the data center industry; only one in 10 operators report staff leaving the industry for non-data-center work, which is down from 17% in 2022.

This is not a complete solution: rising salaries are subject to economic constraints and the return on investment eventually diminishes. Pay increases cannot meaningfully grow the pool of qualified new workers and can contribute to competitors "poaching" employees — one in five operators (22%) report losing staff to other data center organizations.

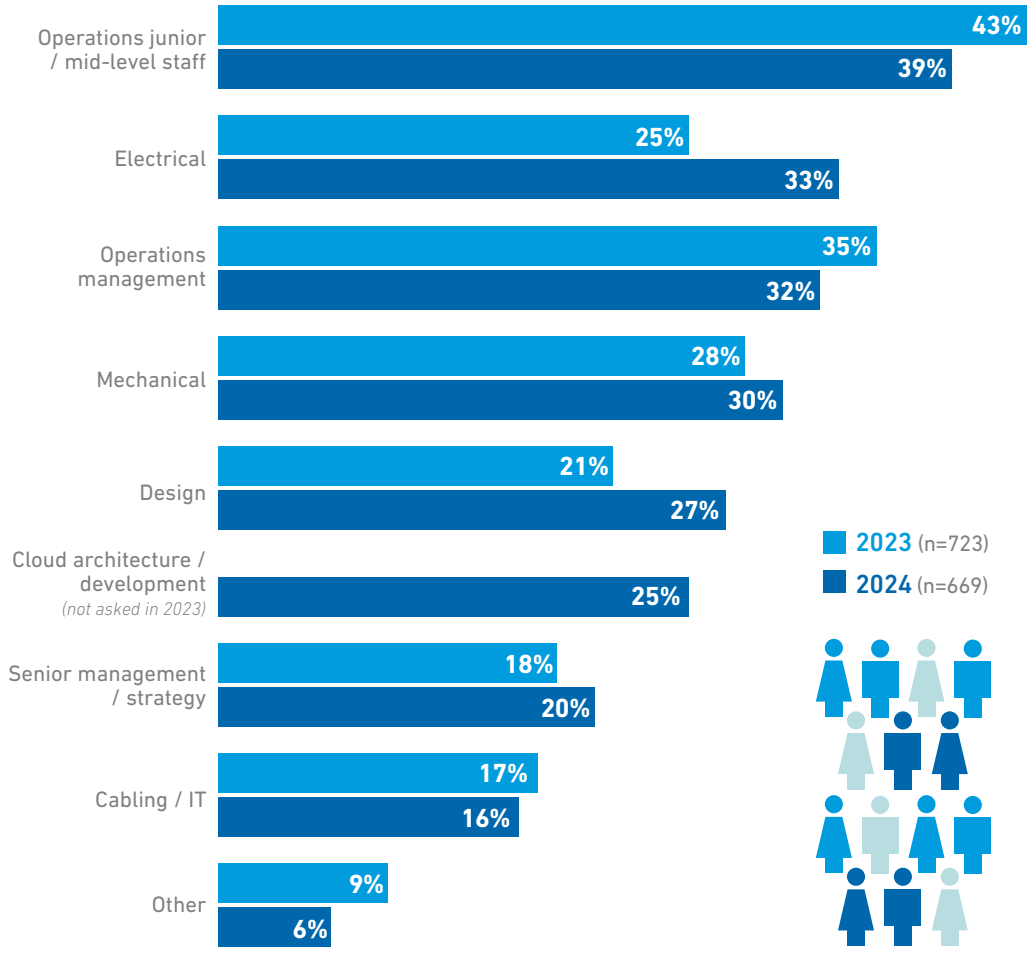
More operators are investing in mentorship, employee outreach and hiring partnerships. The most effective programs respond to the most pressing skill requirements of the employer. Though data center operators describe shortfalls in every area of worker competency, including trades, operations, and management, these shortfalls are not evenly distributed (see **Figure 15**).

Respondents to the survey say their most significant skills gaps affect electrical (33%) and mechanical (30%) roles, junior and mid-level operations (39%), and operations management (32%). Organizations can take advantage of programs such as internships to meet their most pressing labor needs — if the skills required do not require immediate experience or professional qualifications. For instance, if open roles in junior operations do not require a college degree, recruiters may choose to shift investment away from college internships in favor of trade schools or on-the-job training.

Figure 15

Shortfalls persist in trades and management

In which of the following areas is your organization experiencing significant gaps in staff skills? Choose all that apply.



UPTIME INSTITUTE GLOBAL SURVEY OF IT AND DATA CENTER MANAGERS 2024

The ranking of roles in skill shortages has shifted since 2023 — notably, electrical labor climbed to the second highest concern. Outfitting data center space for high-powered, high-density IT for AI and similar applications requires electrical distribution skills for both IT and cooling. Operators making these upgrades in 2024 are likely to feel a greater requirement for skilled electrical workers than in earlier years. This underscores how a potential nascent AI boom can affect the data center industry in multiple ways. Although most operators doubt that AI in the data center will reduce the need for staff within the next five years, adapting to increasing AI workloads may require additional staff today.

Data center staff shortages also vary by region. While operators from North America and Europe describe difficulty in finding skilled junior-level operators, those from China and the Middle East report a disproportionate lack of experienced operations managers.

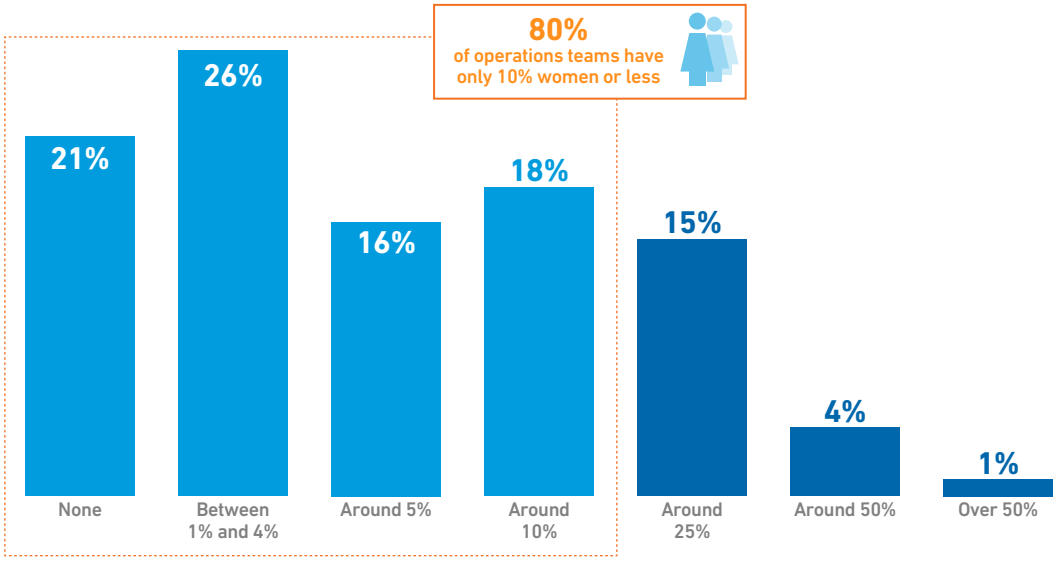
This is somewhat intuitive: North America and Europe have more developed and mature data center markets and have accumulated staff with longer-term experience. Uptime's past surveys also find workers aged 55 years and older tend to outnumber those younger than 35 years in these markets. Operators in other regions where the workforce is generally younger tend to focus recruitment on trade schools and this may serve as a useful model for organizations in any region seeking a pipeline of new workers.

Women are still an underutilized talent pool for data centers. In our survey sample, women continue to represent a very small minority of design, build and operations staff. A scant 4% of respondents report an equal ratio of men to women in their operations staff (see **Figure 16**). Far more commonly (80% of our sample), data center teams employ around 10% or less women and 20% of organizations employ no women at all.

Figure 16

Women remain scarce in data center operations

What portion of your organization's data center design, build, or operations staff is women? (n=694)



UPTIME INSTITUTE GLOBAL SURVEY OF IT AND DATA CENTER MANAGERS 2024

Uptime and others have reported on the data center gender imbalance before. The lack of movement strongly suggests there are qualified female candidates that data center organizations are failing to attract. There is no one clear reason why efforts have not been successful, but there are likely multiple factors at play. Data center owners and recruiters express a willingness to rely on all available labor pools, including women — and in our survey, 40% have a formalized initiative to recruit more women.

However, there remains a serious deficit of actionable data on best practices to address the industry's gender imbalance. Without a proven framework for success, many operators have little trust that expenditure on these initiatives will yield results. Many efforts presently focus on networking opportunities between women already in the field. These can improve retention and foster a supportive environment, but they do little to forge a direct entry path into the data center workforce. The challenge of attracting qualified women to the industry, similarly to the staff shortages generally, will likely require multiple forms of outreach — and the establishment of a robust pipeline.

Appendix

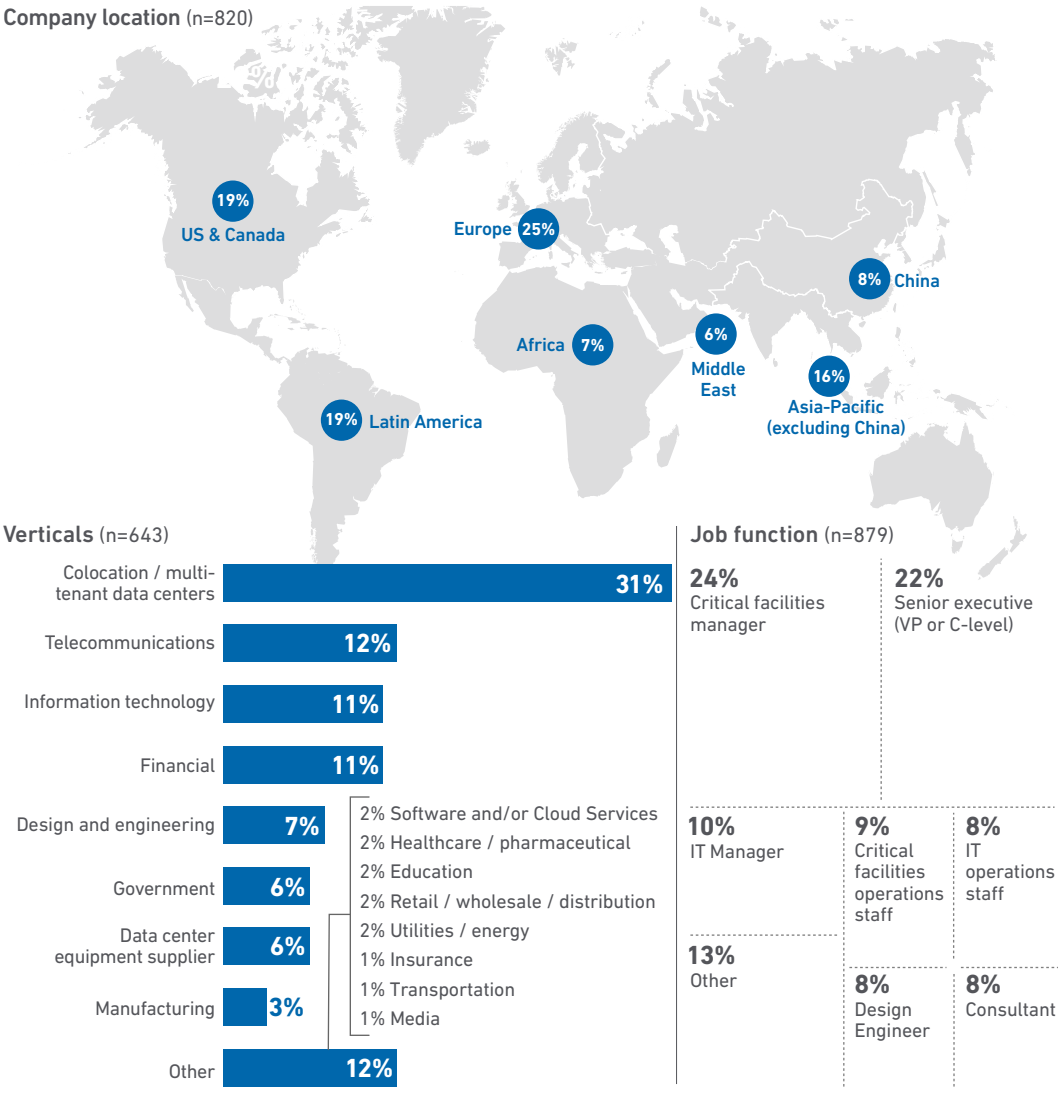
Survey methodology and demographics

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

Appendix (*continued*)

The participants represent a wide range of industry verticals in multiple countries.

Nearly half are located in North America and Europe. Approximately one third of respondents work for professional IT / data center service providers — that is, staff with operational or executive responsibilities for a third-party data center, such as those offering colocation, wholesale, software or cloud computing services.

A total of 879 end users registered for the survey and answered at least one question.

The number of respondents ('n') varies between individual questions because respondents are not required to answer every question.

Findings of previous surveys are available **here**

If you have questions, comments or seek further insights, please contact **research@uptimeinstitute.com**

**This report is based on the entire survey sample. If you are interested in additional data (e.g., by region, sector, data center size, etc.) please contact: Simon Carruthers at scarruthers@uptimeinstitute.com*

About the authors

Douglas Donnellan

Douglas Donnellan is a Research Analyst at Uptime Institute covering sustainability in data centers. His background includes environmental research and communications, with a strong focus on education.

ddonnellan@uptimeinstitute.com

Andy Lawrence

Andy Lawrence is Uptime Institute's Executive Director of Research. He is Uptime Institute's Executive Director of Research and has spent three decades analyzing developments in IT, emerging technologies, data centers and infrastructure. He also advises companies on their technical and business strategy.

alawrence@uptimeinstitute.com

Daniel Bizo

Daniel Bizo is Uptime Institute's Research Director. He has been covering the business and technology of enterprise IT and infrastructure in various roles, including more than a decade as an industry analyst and advisor.

dbizo@uptimeinstitute.com

Peter Judge

Peter Judge is a Senior Research Analyst at Uptime Intelligence. His expertise includes sustainability, energy efficiency, power and cooling in data centers. He has been a technology journalist for 30 years and has specialized in data centers for the past 10 years.

pjudge@uptimeinstitute.com

John O'Brien

John O'Brien is Uptime Intelligence's Senior Research Analyst for Cloud and Software Automation. As a technology industry analyst for over two decades, John has been analyzing the impact of cloud migration, modernization and optimization for the past decade. John covers hybrid and multi-cloud infrastructure, automation, and emerging AIOps, DataOps and FinOps practices.

jobrien@uptimeinstitute.com

Jacqueline Davis

Jacqueline Davis is a Research Analyst at Uptime Institute covering global trends and technologies that underpin critical digital infrastructure. Her background includes environmental monitoring and data interpretation in the environmental compliance and health and safety fields.

jdavis@uptimeinstitute.com

Max Smolaks

Max Smolaks is a Research Analyst at Uptime Institute. His expertise spans digital infrastructure management software, power and cooling equipment, and regulations and standards. He has 10 years' experience as a technology journalist, reporting on innovation in IT and data center infrastructure.

msmolaks@uptimeinstitute.com

Jabari Williams-George

Jabari Williams-George is a Senior Research Analyst for Uptime Intelligence. He provides critical analysis on a variety of topics relating to digital critical infrastructure, particularly in the areas of mechanical and electrical data center engineering. He has a decade of experience in oil and gas, and critical facility design and construction.

jgeorge@uptimeinstitute.com

Rose Weinschenk

Rose Weinschenk is a Research Associate at Uptime Institute covering staffing and education in data centers. Her background includes psychology research, with a focus on ethics.

RWeinschenk@uptimeinstitute.com

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 25 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 Seattle, London, Sao Paulo, Dubai, Singapore, and Taipei.

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