Uptime Institute® INTELLIGENCE

PLANNING & STRATEGY

UI Intelligence report 38

Uptime Institute global data center survey 2020

Authors

Rhonda Ascierto, Vice President of Research, Uptime Institute

Andy Lawrence, Executive Director of Research, Uptime Institute

The tenth Uptime Institute annual survey is the largest and most comprehensive research study of its kind in the data center sector. The findings discussed in this report reveal what operators around the world are thinking, doing and planning in the areas of efficiency, resiliency, workload placement, staffing and new technology adoption.

UII-38 v1.1 published July 13, 2020 last updated July 20, 2020

This Uptime Institute Intelligence report includes:

Introduction

The tenth annual Uptime Institute Global Survey of IT and Data Center Managers provides an overview of the practices, experiences and underlying trends in the mission-critical digital infrastructure industry, today and in the future. This survey, the most comprehensive and longest-running of its kind, was conducted online during March and April 2020. For more details, see the **Appendix**.

The results show a sector that is, as ever, grappling with a number of difficult issues, such as staffing shortages, the move to cloud and increasing complexity. But it is an industry that is growing and adapting to rapid change on multiple levels. In almost every area under discussion — whether outages, resiliency, staffing, placement of workloads, deployment of innovation or use of cloud — there is considerable diversity in the strategies being employed.

KEY FINDINGS

- **The enterprise data center is neither dead nor dying.** The switch of critical loads to a public cloud is happening slowly, with more than half of workloads expected to remain in on-premises data centers in 2022.
- **Transparent clouds are good for business.** Cloud operators would win more mission-critical business if they were more open. Enterprises want greater visibility into facilities and how resiliency is achieved.
- **Edge is still on the edge.** Most organizations expect their edge computing requirements to increase somewhat in 2020, but fewer than 20% expect a significant increase.
- **Average site energy efficiency has flatlined.** Power usage effectiveness values have not improved much across the industry since 2013. But because more work is now done in big, efficient facilities, the overall energy efficiency of IT has improved.
- **Rack densities are rising, but facilities are not stretched.** The mean average density for 2020 was 8.4 kilowatts per rack. Densities are rising, but not enough to drive wholesale site-level changes in power distribution or cooling technologies.
- **Bigger outages are becoming more painful.** Outages generally continue to occur with disturbing frequency, and the bigger outages are becoming more damaging and expensive — a fact supported by Uptime Institute survey findings for three years running.
- **Operators admit most outages were their fault.** Three-quarters of respondents admit that, in hindsight, their most recent major outage was preventable. With more attention and investment, outage frequency would almost certainly fall significantly.
- **Power problems are still the biggest cause of major outages.** Systems/software and networks may be catching up, but power failures — which impact everything on-site and can cause knock-on effects — are the most likely cause of major outages.

Continues next page

KEY FINDINGS *(continued)*

- **Hardware refreshes are less frequent.** Operators are upgrading or replacing their servers less frequently. However, the slowdown in Moore's law means the potential energy savings from frequent refreshes are no longer very significant.
- **The data center staffing crisis is getting worse.** The portion of managers saying they have difficulty finding qualified candidates for open jobs has risen steadily over the past several years.
- **Artificial intelligence won't take over … yet.** Artificial intelligence and automation will not reduce data center operations staffing requirements in the next five years, according to the majority of respondents. After that, however, most think it will.
- **Water use is unmetered by many.** Despite the growing threat of water scarcity, only half of respondents say their organization collects water usage data for their IT/data center operations.
- **More work is needed to address the workforce gender imbalance.** The proportion of women in the data center industry remains very low. Despite pressure and good intent, relatively few operators have a plan or initiative in place to boost the hiring of women.
- **Use of availability zones is now mainstream.** The use of multi-data center availability zones is now common beyond hyperscale operators, with half of respondents saying they use this approach.

COVID-19 survey data

While the impact of COVID-19 on the sector was not the focus of this survey (and is not addressed in this report), Uptime Institute conducted additional surveys in April and July 2020 assessing the effects of the pandemic. See the **Appendix** for details of our COVID-19 survey results.

Corporate IT venues: A mixed picture

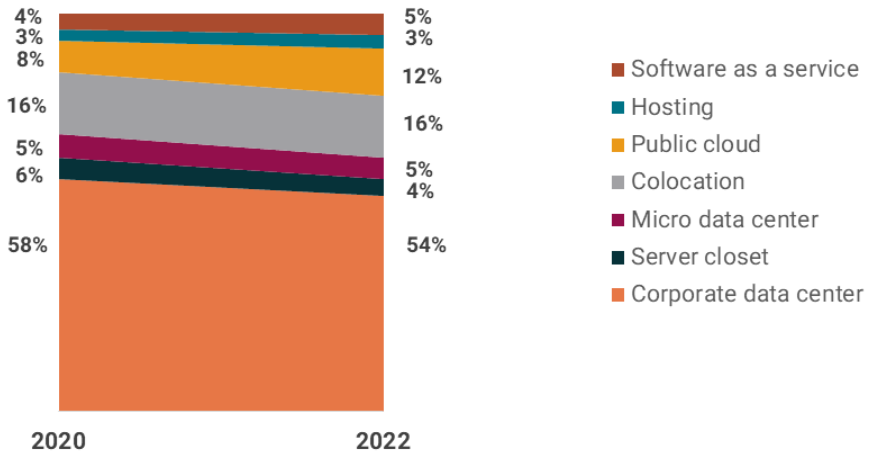
Data center capacity is rapidly expanding in outsourced, third-party IT venues, such as colocation data centers and public cloud. Whether measured in megawatts (MW) of uninterruptible power supply capacity or IT load (or by some other measure, such as square feet of white or leased space, or in units of compute or storage) overall capacity is growing rapidly.

There are many factors driving this growth, including newer cloud-based services, such as social media and streaming; mobile applications and services; new enterprise services and applications; the migration of corporate workloads into colocation sites; and the adoption of software as a service (SaaS) and public cloud platforms.

Does this mean that almost all IT workloads will — eventually — end up running in a third-party data center? In the 2020 Uptime Institute survey, as in years past, we asked respondents to estimate, by percentage, how

much of their workload/data is processed/stored in different types of data centers today and how this might look in two years' time.

The majority (58% today, 54% projected in two years) said that most of their workloads run in corporate data centers — that is, enterprise-owned, on-premises facilities (see Figure 1).



Approximately what percentage of your organization's total IT would you describe as running in the following IT environments today, versus in two years? (Your answers for each year must sum to 100%)

Source: Uptime Institute Global Survey of IT and Data Center Managers 2020 (n=390; for 2022, n=387)

Uptime Institute | INTELLIGENCE

Figure 1. **Corporate IT venues, 2020 versus 2022**

These findings are similar to those of our previous years' surveys. They confirm Uptime Institute's view that the enterprise-owned data center sector, while not necessarily the most innovative, will continue to be the foundation of enterprise IT for the next decade. In our survey, nearly two-thirds of IT workloads are expected to be running in privately owned environments (large data centers, server closets and micro data centers) by 2022, with the remainder contracted to external suppliers. Although the enterprise data center sector is falling as a percentage of the whole, it is still growing (slowly).

A mix of factors typically drive enterprise demand for third-party IT venues; similarly, multiple factors drive demand for on-premises data centers. Often some combination of the factors listed below come into play when deciding the best-execution venue for workloads.

COMMON DRIVERS FOR IT VENUE DECISIONS
(Select examples)

Factors driving outsourcing to third-party data center services

- **Cost:** Outsourcing can lower costs in the short to medium term. For organizations "born" in a public cloud or colo, it typically is cost-prohibitive to move to an enterprise data center.
- **Cost allocation:** Outsourcing shifts cost allocations from capex toward more repeatable opex models.
- **IT agility and flexibility:** Outsourcing provides the ability to readily and quickly adapt to changing capacity needs without the burden of managing the full stack of IT and applications; IT can be used for a project's duration only (e.g., for test and development).
- **Access to resources:** Third-parties may provide access to a wider range of resources, including technology, interconnections, software tools, services and application environments.
- **Security:** Third-parties can offer the most advanced, highly resourced security features.

Factors driving demand for on-premises enterprise data centers

- **Cost:** Ownership delivers total cost of ownership benefits over the long term; in the shorter term, owners avoid the data transport costs of moving to an outsourced venue.
- **Governance:** On-premises environments may be necessary for compliance with data governance and regulatory requirements.
- **Control:** Owners can closely monitor and control factors such as latency, availability and application performance. While most outsourced venues are strong in these areas, service level agreements vary and are limited.
- **Risk**: Ownership ensures full visibility into (and the ability to adjust) the risk profile of every workload.
- **Security:** Ownership provides the ability to maintain control and governance (dedicated rather than shared physical infrastructure) over security features.

Source: Uptime Institute Intelligence 2020

Migration and repatriation of workloads

While many IT workloads will move permanently to a third-party venue, for other workloads the move may be temporary.

The migration to public cloud is widely tracked and understood. But we asked if respondents' organizations had moved workloads or data from a public cloud to a private cloud or private on-premises/colocation environment during the past year and, if so, what was the primary reason?

As Figure 2 shows, the majority (70%) of organizations surveyed had not moved a workload from a public cloud. For those that had, the most common primary reason for doing so was cost (one-third cited this reason), followed closely by regulatory compliance — for example, financial governance or compliance with privacy laws.

Other reasons included performance issues (accounting for over 16% of moved workloads) and perceived concerns over security. Actual security breaches were responsible for just a tiny portion of repatriated workloads. This underscores the argument that effective security can be a reason why some workloads are moved to commercial IT services such as public cloud.

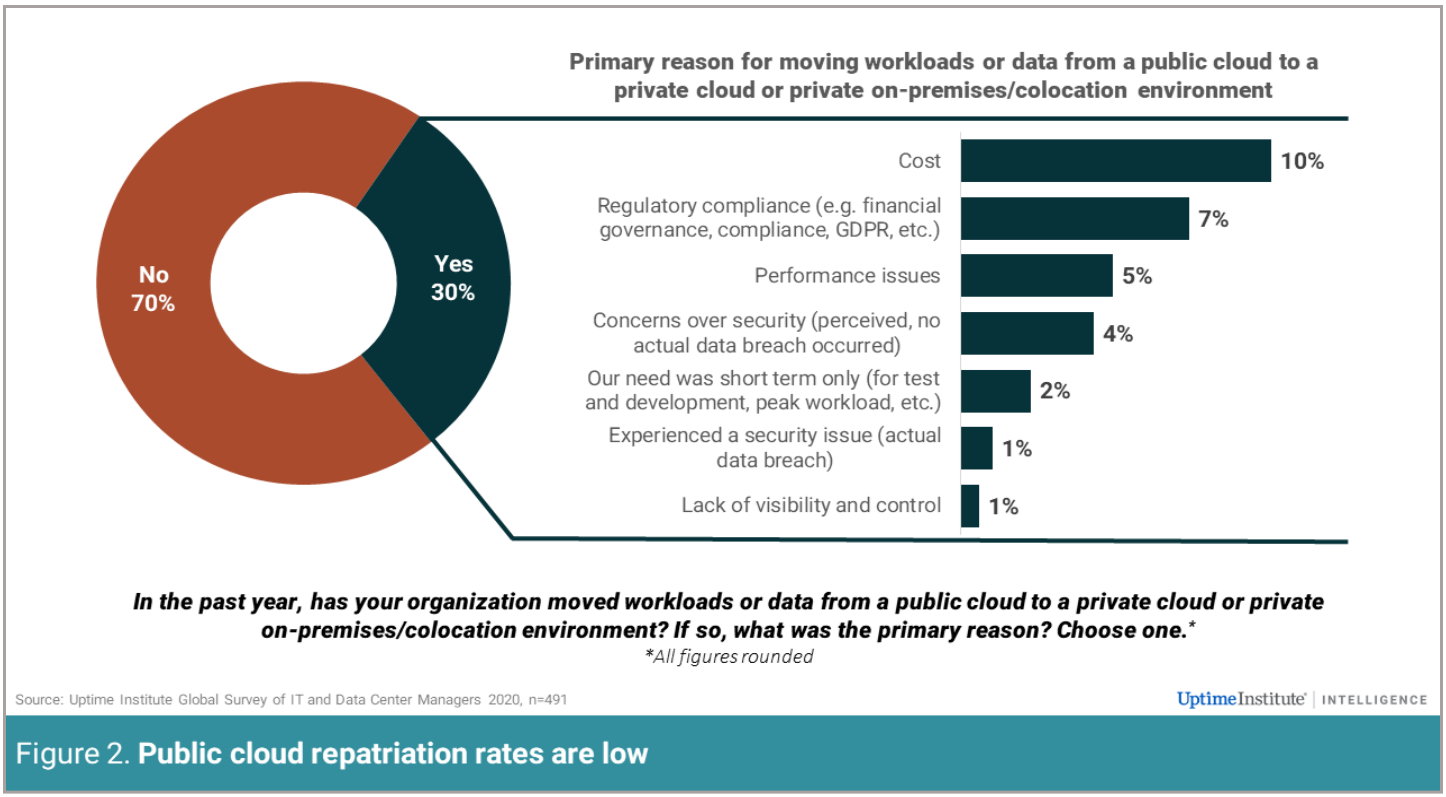


Figure 2. **Public cloud repatriation rates are low**

Cloud visibility

As shown in **Figure 1**, the use of public cloud is expected to increase from 8% of workloads today to 12% in two years. Although this may seem undramatic when set alongside other industry forecasts, public cloud represents the biggest increase of usage for an IT venue in our survey. Yet, as a percentage of enterprise IT workloads, public cloud usage is relatively minimal.

What could change this?

According to our survey, a lack of visibility, transparency and accountability of public cloud services is clearly a major issue for enterprises with mission-critical applications.

We asked respondents whether their organization had adequate visibility into the resiliency of public cloud operations in terms of architecture, availability record, management processes and options. The results were nearly identical to the results of the same question in 2019, with the same conclusion: Enterprises want more visibility into how cloud operators manage their operations. If they had that visibility, they would be more likely to use a public cloud.

As shown in Figure 3, just 17% say their organization has adequate visibility and they place mission-critical workloads into a public cloud. Another 10% say they do not have enough visibility, but they still place critical workloads in the cloud. Most importantly, 21% say they would be more likely to run mission-critical workloads in a public cloud if there were a higher level of visibility into the operational resiliency of the service.

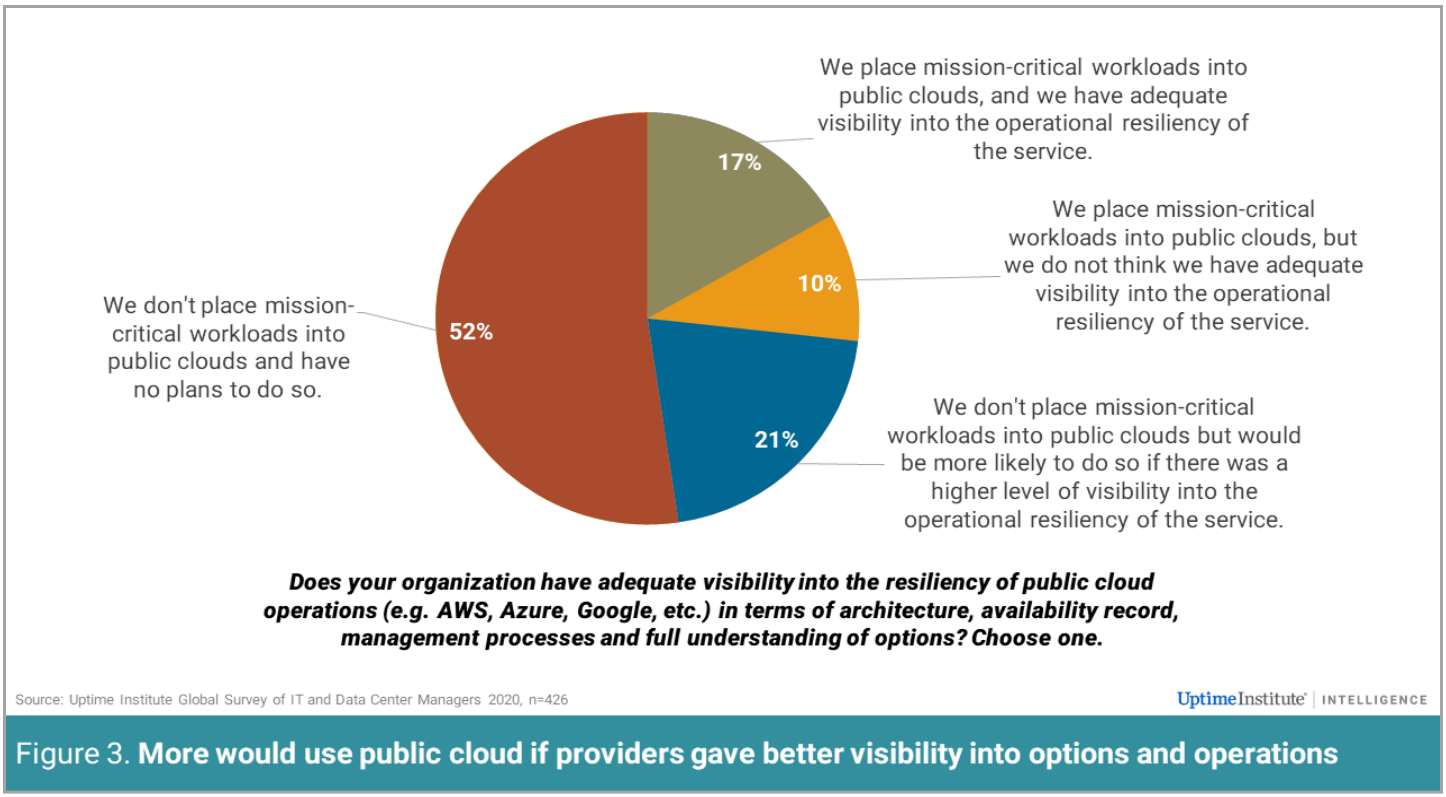


Figure 3. **More would use public cloud if providers gave better visibility into options and operations**

The findings also clearly show that for many organizations, the public cloud is simply not an option: over half have no plans to put any mission-critical workloads in the public cloud.

Edge demand is growing slowly, steadily

One of the most widely anticipated, widely discussed trends in IT and infrastructure is a new wave of demand for edge computing, fueled by new technologies such as 5G, the internet of things (IoT) and artificial intelligence (AI). All these may require much more computing of data gathered near the point of use — which will, in turn, require many more small data centers.

In our report **Ten data center industry trends for 2020**, we identified edge data centers as a significant new development. However, the market is in the early stages: we do not anticipate sudden mass deployment, nor a single standout use case or single design dominating before 2022.

This is supported by our survey findings. Most respondents say their organization's edge requirements will increase in 2020, but 40% anticipate it to increase only somewhat. Just 18% say it will increase significantly. Slightly more than a third (37%) do not expect their organization will require edge computing capabilities in 2020 (see Figure 4).

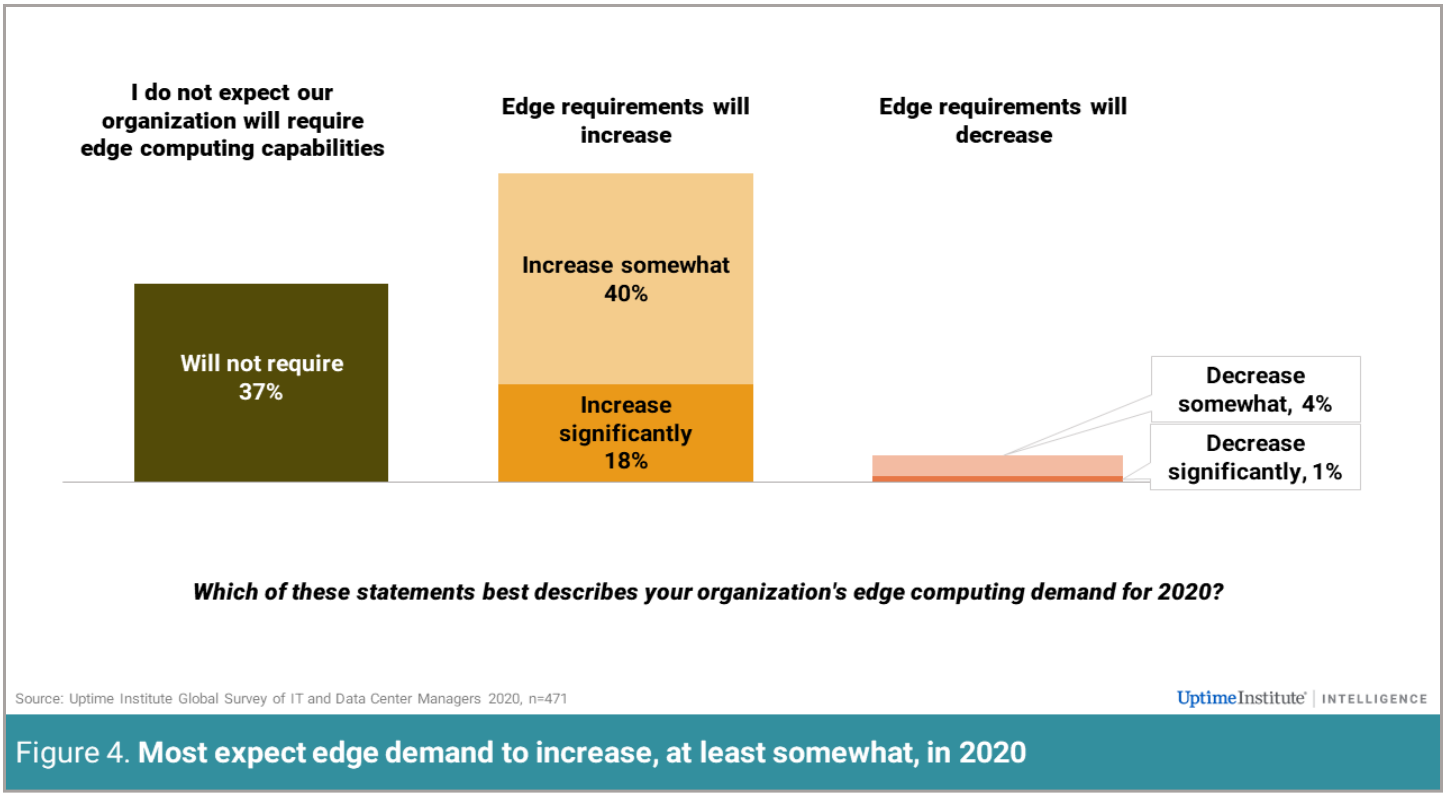


Figure 4. **Most expect edge demand to increase, at least somewhat, in 2020**

There is a deep discussion about the extent of data center capacity needed at the local edge — about just how many applications and services really need local edge processing, and about the type and size of IT equipment needed. While the technical answers to most of these questions are largely understood, questions remain about the economics, the ownership, and the scale and pace of deployment of new technologies and services — all of which critically affect edge development.

Power usage effectiveness flattens out

In our 2019 survey, our question on average power usage effectiveness (PUE) values yielded a surprising result: data centers in the survey had become marginally less efficient in the preceding year (1.67 in 2019, compared with 1.58 in 2018). In 2020, the average PUE for a data center in the global Uptime survey sample is 1.59, a slight improvement from 2019.

The new data (shown in Figure 5) conforms to a consistent pattern: Big improvements in energy efficiency were achieved between 2007 and 2013 using mostly inexpensive or easy methods (such as simple air containment) — moving beyond that involved more difficult or expensive changes. Since 2013, improvements in PUE have been marginal.

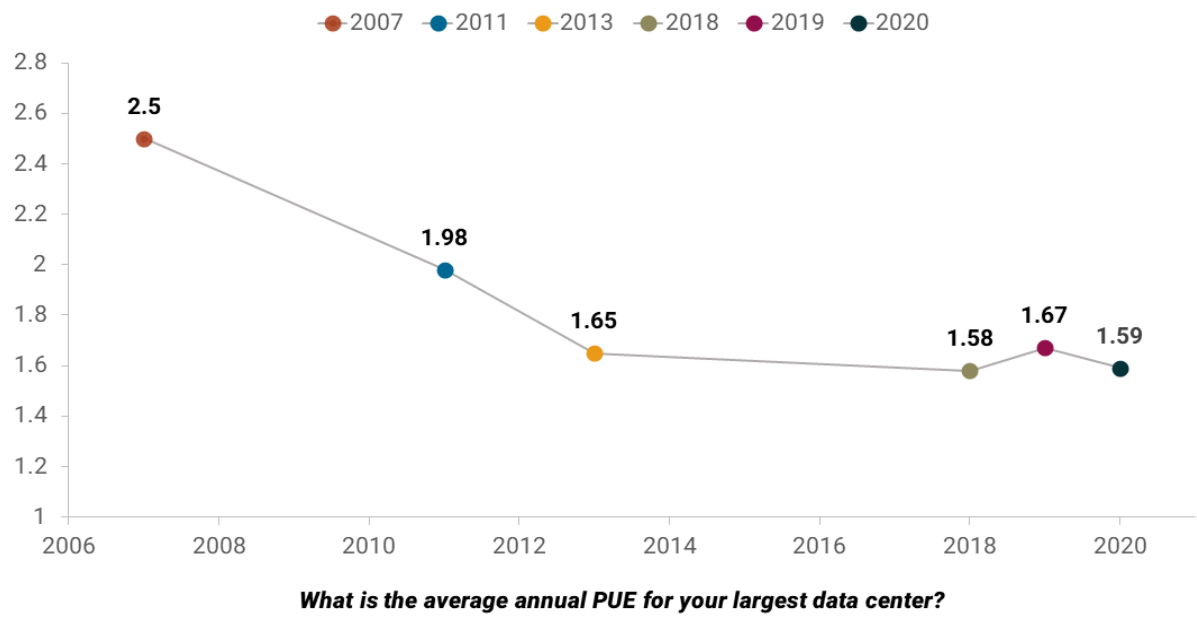


Figure 5. **Data center energy efficiency gains have flattened out**

Almost all operators strive to get their PUE ratio as near 1.0 as possible. Using the latest technology and practices, most new builds fall between 1.2 and 1.4. But there are still thousands of older data centers that cannot be economically or safely upgraded to become that efficient, especially if high availability is required. As stated, the 2019 average PUE increased slightly, with a number of possible explanations (see **Is PUE actually going up?**).

A single number, however, cannot tell a complete story. This data is based on the average PUE per site, regardless of size or age. Newer data centers, usually built by hyperscale or colocation companies, tend to be much more efficient — and larger. Therefore, a growing proportion of work is being done in these larger, more efficient data centers. Uptime Institute data in 2019 shows data centers larger than 20 MW have lower PUEs. Data released by Google shows almost exactly the same curve shape, with improvements flattening out — but at much lower (more efficient) values.

Operators that cannot improve their site PUE can still do a lot to reduce energy use and/or decarbonize operations. First, they can improve their IT utilization and refresh their servers to ensure IT energy optimization. Second, they can re-use the heat generated by the data center; and third, they can buy renewable energy or invest in renewable energy generation.

Density is rising

The power density per rack (kilowatts [kW] per cabinet) is a critical number in data center design, capacity planning, and cooling and power provisioning. There have been industry warnings about a meteoric rise in IT equipment rack power density for the past decade (at least).

One reason for this prediction is the proliferation of compute-intensive workloads (e.g., AI, IoT, cryptocurrencies, and augmented and virtual reality), all of which drive the need for high-density racks.

Our 2018 and 2019 surveys found that racks with densities of 20 kW and higher are becoming a reality for many data centers (we asked about highest rack density) — but not to the degree forewarned. Year-over-year, most respondents said their highest density racks were in 10-19 kW range, which is not enough to merit wholesale technical changes. When rack densities are higher than 20-25 kW, direct liquid cooling and precision air cooling becomes more economical and efficient. According to what we see in the field, such high densities are not pervasive enough to have an impact on most data centers.

This does not mean that the trend should be ignored. It is clear from our latest research that average mean rack density in data centers is rising steadily, as Figure 6 shows. Eliminating respondents with above 30 kW as high-performance outliers, the mean average density in our 2020 survey sample was 8.4 kW/rack. This is consistent with other industry estimates and safely within the provisioned range of most facilities.

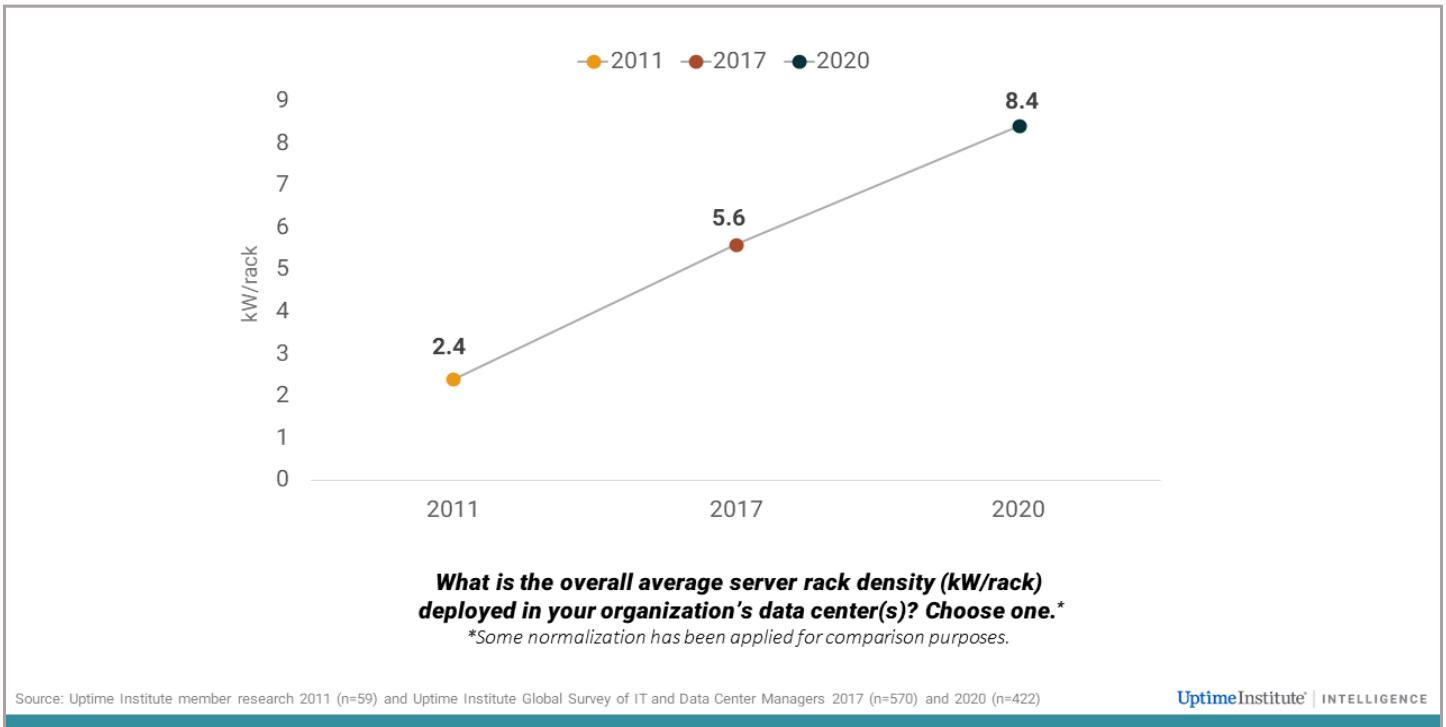
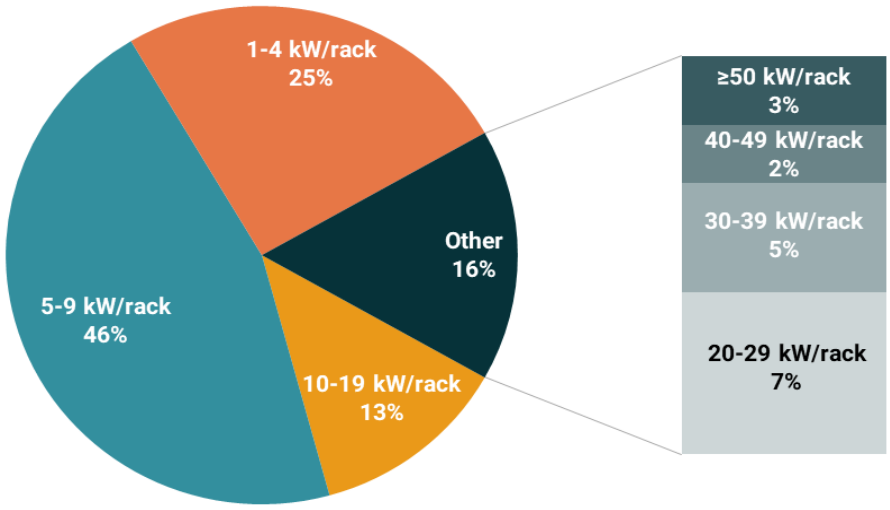


Figure 6. **Average density per rack is rising**

In our 2020 survey, we asked about the most common (modal average) server rack density, which is perhaps a better metric than overall average density. More than two-thirds (71%) reported a modal average of below 10 kW/rack, with just 16% widely deploying 20 kW or higher rack densities (Figure 7). The most common density was 5-9 kW/rack. Overprovisioning of power/cooling is probably a more common issue than underprovisioning due to rising rack densities.



What is the MOST COMMON (modal average) server rack density deployed in your organization's data center(s)? Choose one.*

**All figures rounded*

Source: Uptime Institute Global Survey of IT and Data Center Managers 2020, n=422

Figure 7. **Low rack densities (below 10 kW/rack) remain most common**

Assuming a trend that rack units will increasingly be filled with higher-powered servers that are well utilized, we anticipate that the modal average kW/rack will increase over time. Figure 8 shows that for most organizations — roughly half of those surveyed — average density is increasing, albeit only slowly.

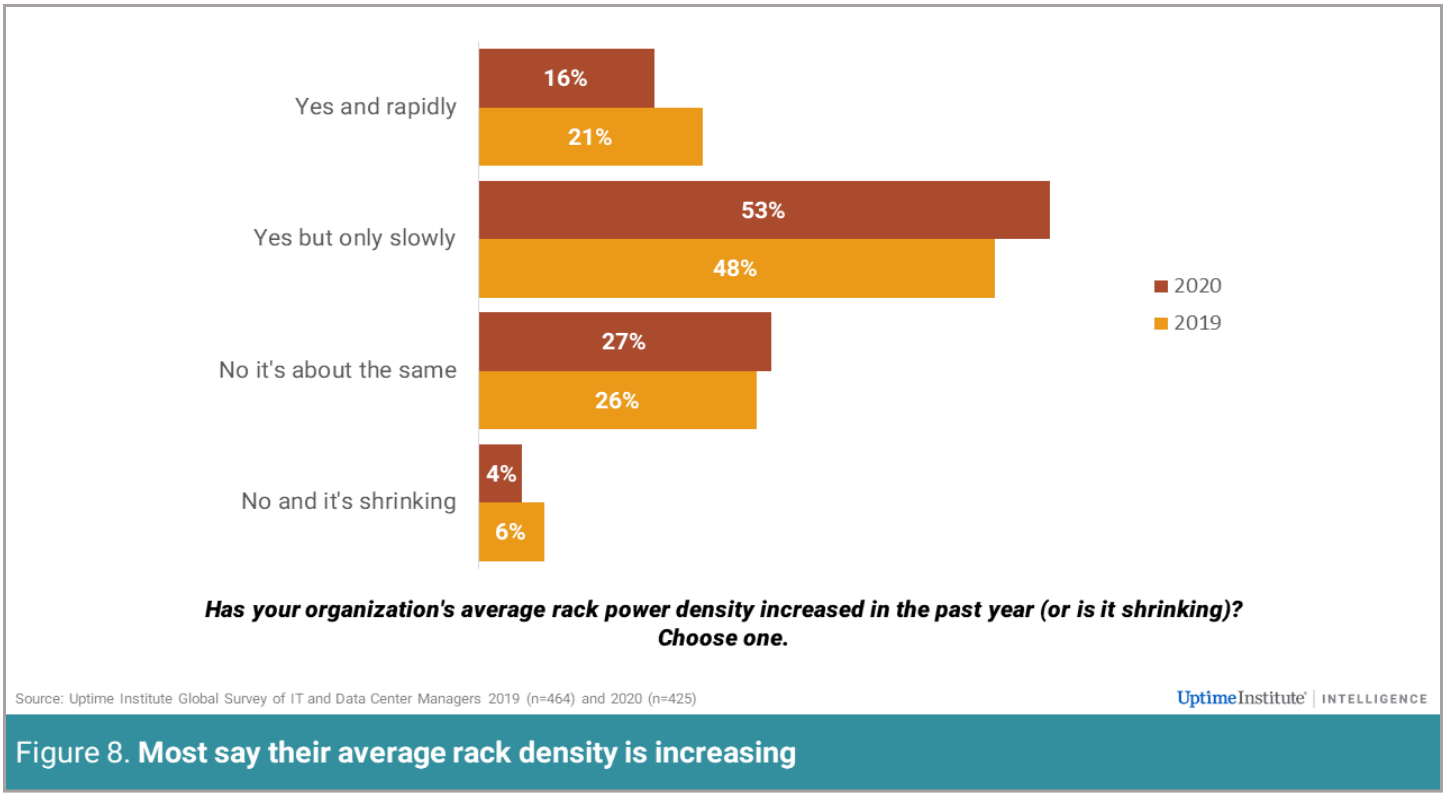


Figure 8. **Most say their average rack density is increasing**

We expect density to keep rising. Our research shows that the use of virtualization and software containers pushes IT utilization up, in turn requiring more power and cooling. With Moore's law slowing down (see **Hardware refresh cycles are prolonged**), improvements in IT can require more multi-core processors and, consequently, more power consumption per operation, especially if utilization is low. Even setting aside new workloads, increases in density can be regarded a long-term trend.

But, as our 2020 survey findings demonstrate, the expectation for 20 kW racks throughout the industry has not manifested. We believe that many compute-intensive workloads — those that will significantly push up power use, rack density and heat — currently reside across a relatively small group of hyperscale cloud data centers and are consumed by organizations as a service.

Hardware refresh cycles are prolonged

Hardware refresh is the process of replacing older, less efficient servers with newer, more efficient ones with more compute capacity. However, there is a complication to the refresh cycle that is relatively recent: the slowing down of Moore's law.

Moore's law refers to the observation made by Intel co-founder Gordon Moore that the transistor count on microchips would double every two years. This implied that transistors would become smaller and faster, while drawing less energy. Over time, the doubling in performance per watt was observed to happen around every year and a half.

It is this doubling in performance per watt that underpins the major opportunity for increasing compute capacity while increasing efficiency through hardware refresh. But in the past five years, it has been harder for Intel (and immediate rival AMD) to maintain the pace of improvement. This raises the question: Are we still seeing these gains from recent and forthcoming generations of central processing units? If not, the hardware refresh case will be undermined.

Our longitudinal survey data suggests that a small majority of data center managers think the case for more frequent hardware refreshes has diminished. As shown in Figure 9, the most common period for a hardware refresh in 2015 was three years; in 2020, it had extended to five years.

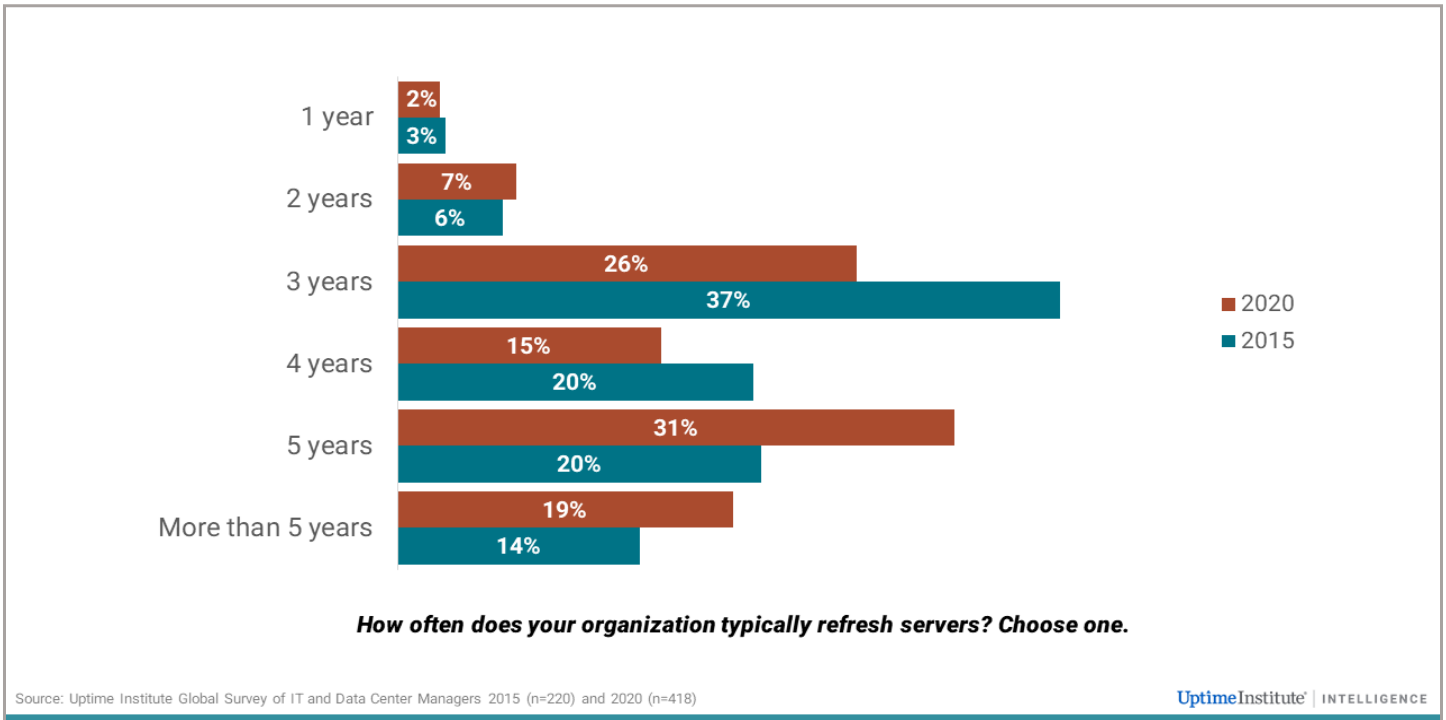


Figure 9. **Data center server refresh cycles, 2015 versus 2020**

While there may be other factors at play (such as budget or a move to cloud), this makes sense from an energy point of view. As detailed in our report **Beyond PUE: Tackling IT's wasted terawatts**, there remains a strong case for energy savings when replacing the oldest servers, but less so for refreshing more recent servers — say, those up to three years old.

Outages: More disruptive, more common

For the past several years, Uptime Institute has been surveying operators on their experiences of outages, as well as closely tracking publicly recorded incidents reported in the media. Such information is difficult to collect and assess; organizations' public relations teams don't like to discuss their failings, public reporting is unreliable, and the definitions of what constitutes an outage (and, beyond that, one that is worth recording) have become vaguer. Practical and shareable insights, based on causes, responses and impacts, are yet harder to extract.

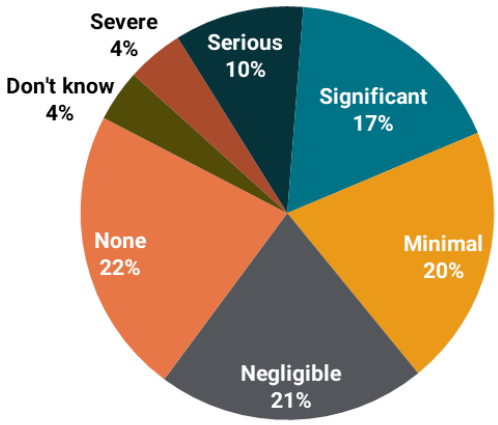
Even so, the trends from our research ring like an alarm bell: In surveys from 2018 and 2019, and now supported by our 2020 survey, it is clear that outages occur with disturbing frequency, that bigger outages are becoming more damaging and expensive, and that what has been gained in improved processes and engineering has been partially offset by the challenges of maintaining ever more complex systems. Avoiding downtime remains a top technical and management challenge for all owners and operators.

Frequency and severity of outages

How prevalent — how common — are outages? And when they do happen, how serious are they? In our 2018 and 2019 Uptime Institute surveys, we asked the same question each time, producing almost identical answers for those years: A third of organizations told us they experienced a major outage in the previous 12 months, and about half in the previous three years.

In 2020, we wanted to know more about the impact of an outage. Uptime classifies outages based on a severity scale we introduced in 2018. In particular, Uptime Institute was receiving feedback in the market that many organizations were suffering a worrying number of small service disruptions (Category 1 – negligible) that were not being recorded officially/picked up in our surveys.

The responses confirmed this. As shown in Figure 10, 78% of organizations say they had an IT service outage in the past three years — a higher percentage than in previous years — and 41% classified it as minimal or negligible. Outages in these categories signal bigger problems and are troubling more for their frequency than for their singular impact. When asked about significant, serious or severe outages — which can cause substantial financial and reputational damage — 31% have been affected.



Category 1 Negligible	Recordable outage, but little or no obvious impact on services.
Category 2 Minimal	Services disrupted. Minimal effect on users, customers and/or reputation.
Category 3 Significant	Customer/user service disruptions, mostly of limited scope, duration or effect. Minimal or no financial effect. Some reputational or compliance impact(s).
Category 4 Serious	Disruption of service and/or operation. Ramifications include some financial losses, compliance breaches, reputation damages, possibly safety concerns. Customer losses possible.
Category 5 Severe	Major and damaging disruption of services and/or operations with ramifications including large financial losses, possible safety issues, compliance breaches, customer losses, reputational damage.

How would you classify the most significant outage that your organization has had in the past three years, either in your own site or a third-party provider site? Choose one.*

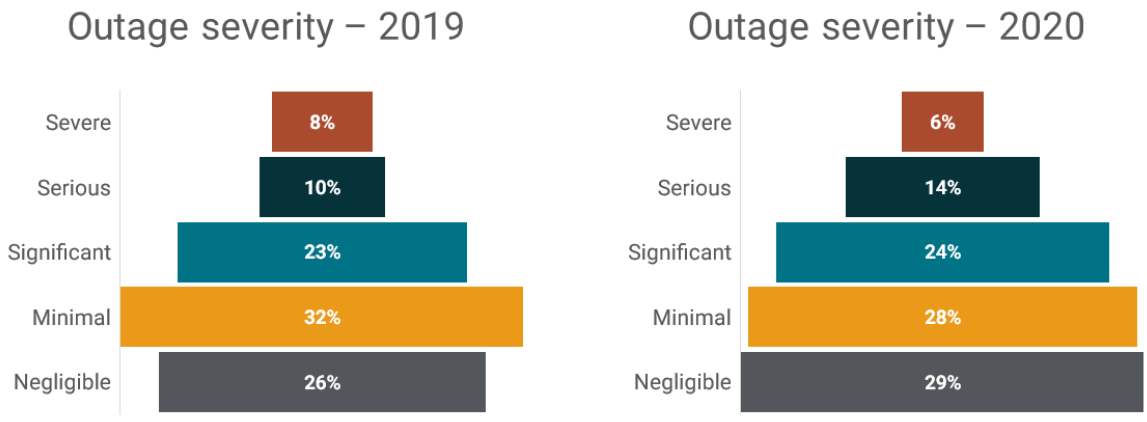
**All figures rounded*

Source: Uptime Institute Global Survey of IT and Data Center Managers 2020, n=494

Uptime Institute | INTELLIGENCE

Figure 10. **A third experienced a major outage in the past three years**

If we eliminate those that did not have an outage, we see a pyramid in outage severity (see Figure 11), with minor outages being more common. In both 2019 and 2020, almost 60% were minimal/negligible and about 40% were considered major (i.e., significant, serious or severe). About 20% of organizations had a serious or severe outage in the past three years — that is, an outage that was costly, caused reputational damage and, in some cases, had major other implications. At a rough-cut level, about a third of all outages cause financial/reputational damage.



How would you classify the most significant outage that your organization has had in the past three years, either in your own site or a third-party provider site? Choose one.*

**All figures rounded*

Source: Uptime Institute Global Survey of IT and Data Center Managers 2020, n=494

Uptime Institute | INTELLIGENCE

Figure 11. **About a fifth reported a serious or severe outage in the past three years**

Preventability and measurability

It is almost a truism that training and good operational procedures improve reliability and reduce accidents/incidents. Most managers and operators whose data centers had an outage clearly think they could have done better (see Figure 12): only a quarter think responsibility for their most recent outage lay outside their hands.

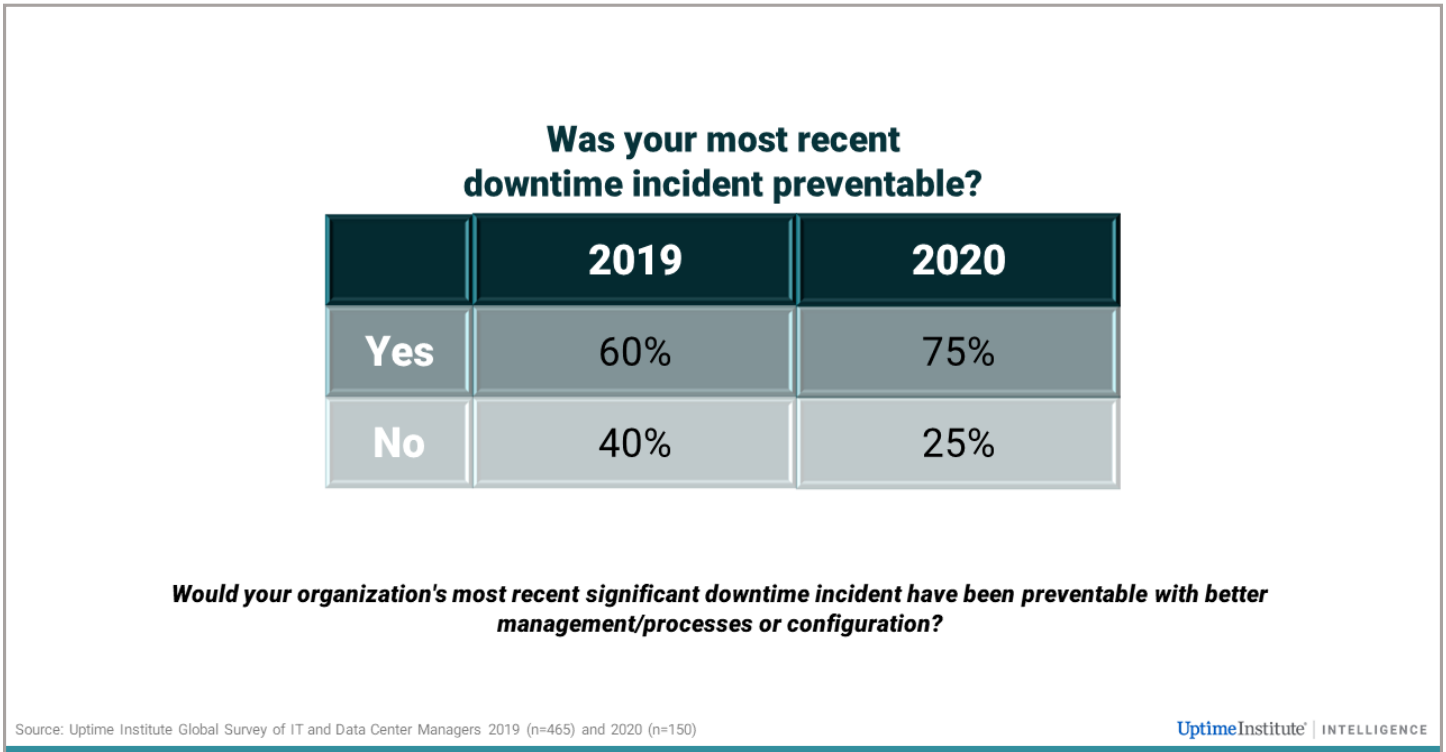


Figure 12. **Most outages are preventable**

This sounds like both a damning self-criticism and an honest self-assessment. However, it is not clear if operators are openly learning from process problems or blaming their managers. It's also possible managers are blaming the operators — or all could be blaming executives for underinvestment. Regardless, the findings point to the clear opportunity: With more investment in management, process and training, outage frequency would almost certainly fall significantly.

Cost of outages

As in our 2019 survey, we asked operators who reported an outage to estimate the total cost of their most recent significant one. In 2020, a greater percentage of outages cost more than $1 million (now nearly one in six rather than one in 10, as in 2019), and a greater percentage cost between $100,000 and $1 million (40% vs. 28%; see Figure 13).

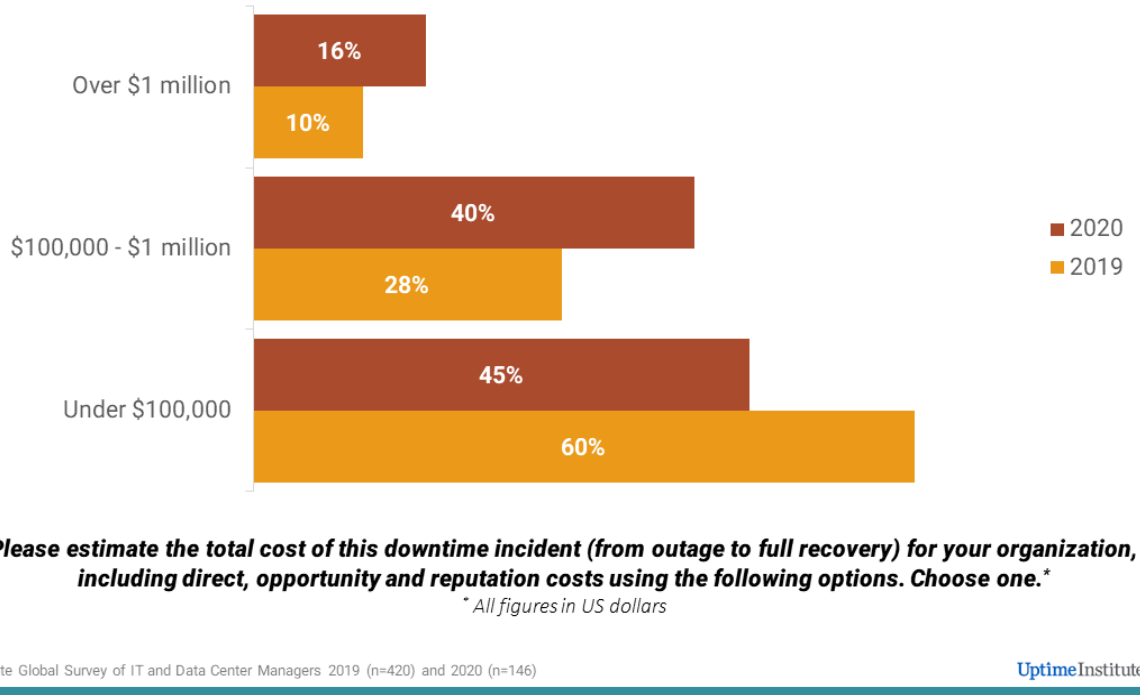


Figure 13. **Outage costs are rising**

This is part of a trend. While there are many small incidents that worry managers, especially in their frequency and potential, it is clear that the biggest outages are becoming more expensive. This reflects growing dependency on IT by all businesses and consumers, increasing interdependency of a growing number of systems in real time, and the immediacy of the impact of downtime on customers.

Uptime Institute is frequently asked about the average cost of an outage. Although various averages can be calculated, the insights gained are not useful. One reason is that a few big outages can be enormously costly and distort the overall picture. In our 2019 survey, there were 10 incidents that led to losses of over $25 million; in 2020, there were three. In all years, two-thirds to three-quarters of all outages cost less than $250,000.

One troubling fact revealed in our survey is that only about half of organizations actually calculate the cost of a downtime incidents. This number is trending up, probably as a result of the cost impact and publicity that results from service interruptions. Uptime institute recommends all incidents be logged and that the real and potential cost impacts be tracked or projected; only in this way is it possible to accurately understand the return of investments in increased availability.

Causes of outages

Understanding the causes of outages is critical to preventing them and to knowing where investment is necessary. Unfortunately, corporate secrecy, and sometimes a failure to carry out a proper analysis, make sharing of insights more difficult.

For our 2020 survey, we asked organizations a specific question: What was the primary cause of your most major (Category 3 and above) outage? Smaller outages with negligible/minimal impact were not included. The results show that on-site power problems remain the single biggest cause, followed by software/systems and network (see Figure 14). In spite of many concerns to the contrary, problems at cloud or SaaS providers cause only a small proportion of outages.

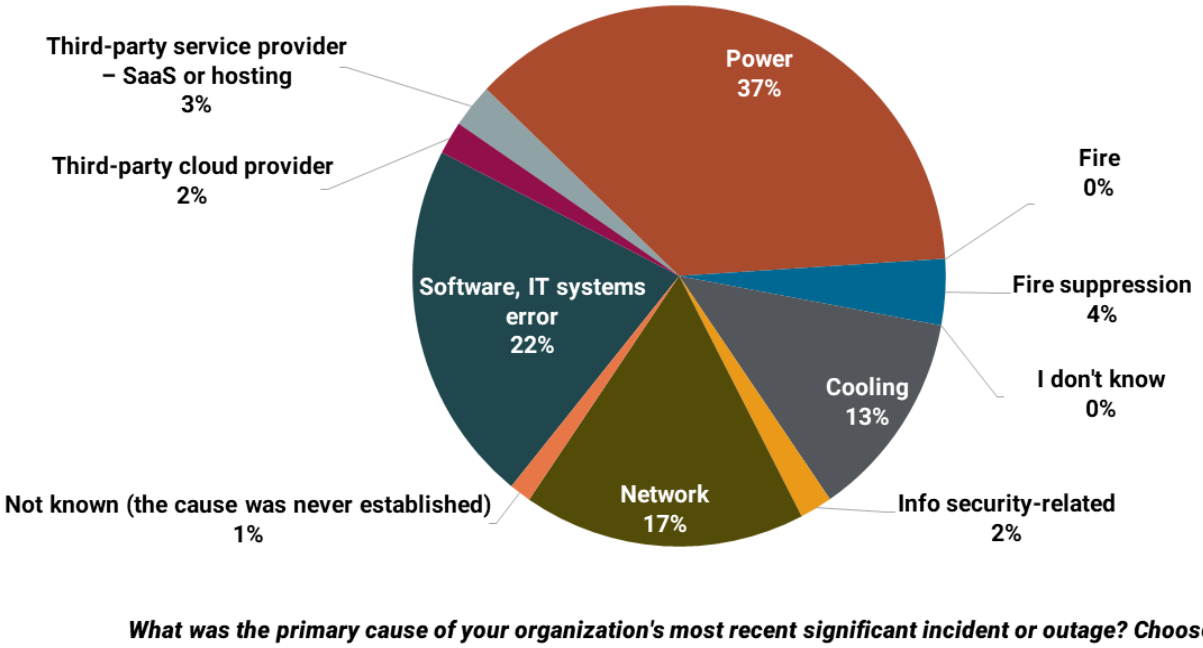


Figure 14. **On-site power issues cause over a third of noteworthy outages**

In recent years, our research has shown a growing proportion of outages are caused by software and network issues. While these incidents are more common and can be difficult to fix, many cause only minor problems. The impact of a power outage is wide and deep, and the knock-on effects can be long lasting — even if the initial failure is quickly fixed.

Testing of resiliency lags responsibility

During Uptime Institute's ongoing engagements around the issue of downtime and resiliency, advisors find that the management of resiliency varies widely. The most effective organizations — those most able to achieve high levels of availability and those able to recover quickly — have many common characteristics. One of these is having a single senior manager responsible for resiliency and/or a senior person responsible for regularly testing for resiliency.

Among the organizations surveyed in 2020, not quite half had a senior manager responsible for regular testing of resiliency, while a further 21% had a manager responsible for overall resiliency, but not testing (see Figure 15). While the exact roles require clarification, there is a trend towards greater executive management of resiliency.

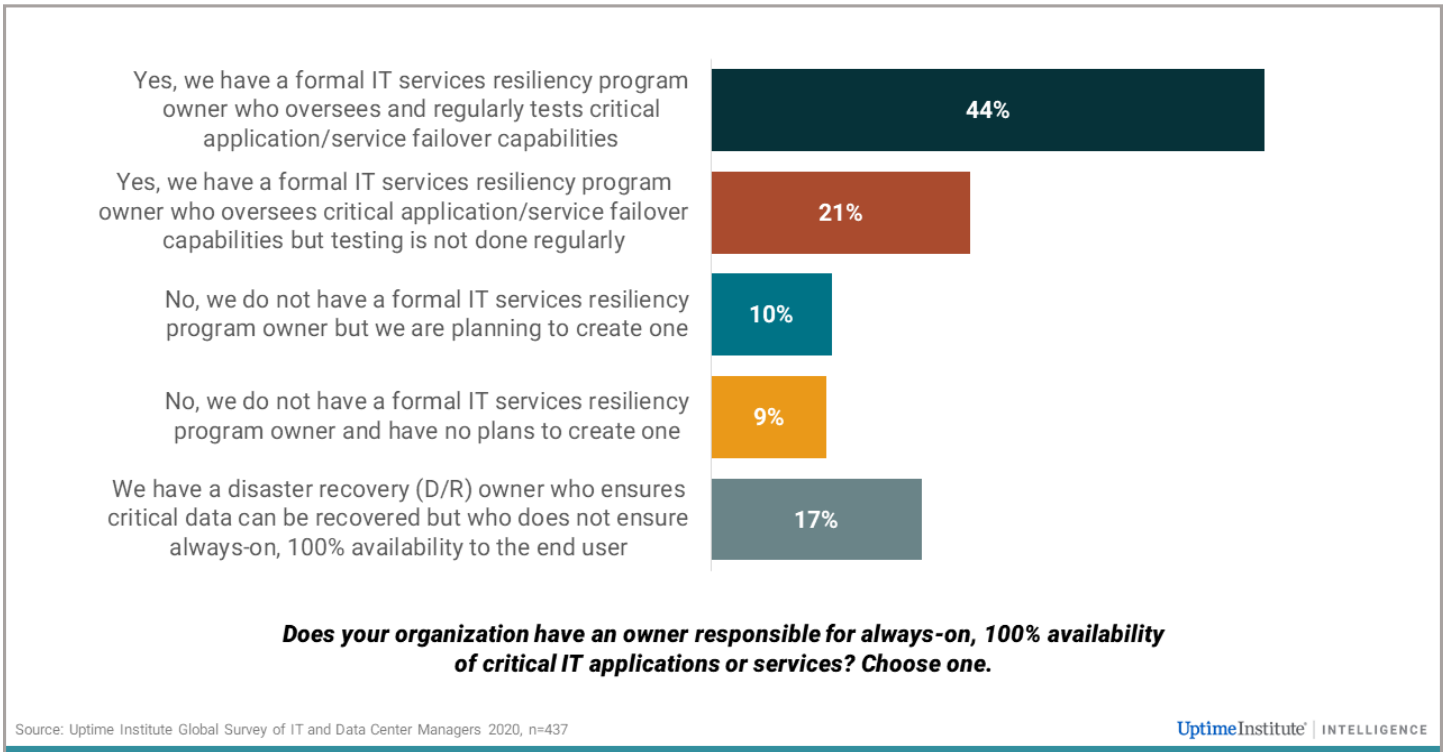


Figure 15. **Most have a formal IT services resiliency program owner but testing varies**

Use of availability zones is spreading

In the past decade, availability zones have become the de facto way for hyperscale operators to maintain always-on, at-scale service. Using the strategy, data centers are organized into three (or more) clusters, and data and processing distribute across all three. In the event of a failure of one site, the others (two or more) take up all the load synchronously. Each of the sites is fully active — there is no disaster recovery. As long as there are at least three data centers sufficiently near each other to ensure low latency, but far enough away to avoid a localized event affecting more than one data center, the approach is considered initially expensive but highly effective.

The approach is not limited to hyperscales. An enterprise, for example, can use racks at three colocation sites to achieve the same result. In this way, the need for fault tolerance at a single data center may be theoretically reduced (this is debatable – most operators still maintain at least a Tier III-level data center). The approach is highly effective for private cloud workloads that can move seamlessly between data centers.

As Figure 16 shows, this availability zone approach is spreading rapidly across enterprises, with half of operators now saying they employ this architecture, using their own data centers and/or with colocation partners. This is a significant increase on earlier surveys and supports the general trend. The growing use of availability zones may, in some cases, indicate that the applications and data are becoming “cloud ready” and could move onto a public cloud at some point. Of those that do not use availability zones, about a third expect to do so within three years.

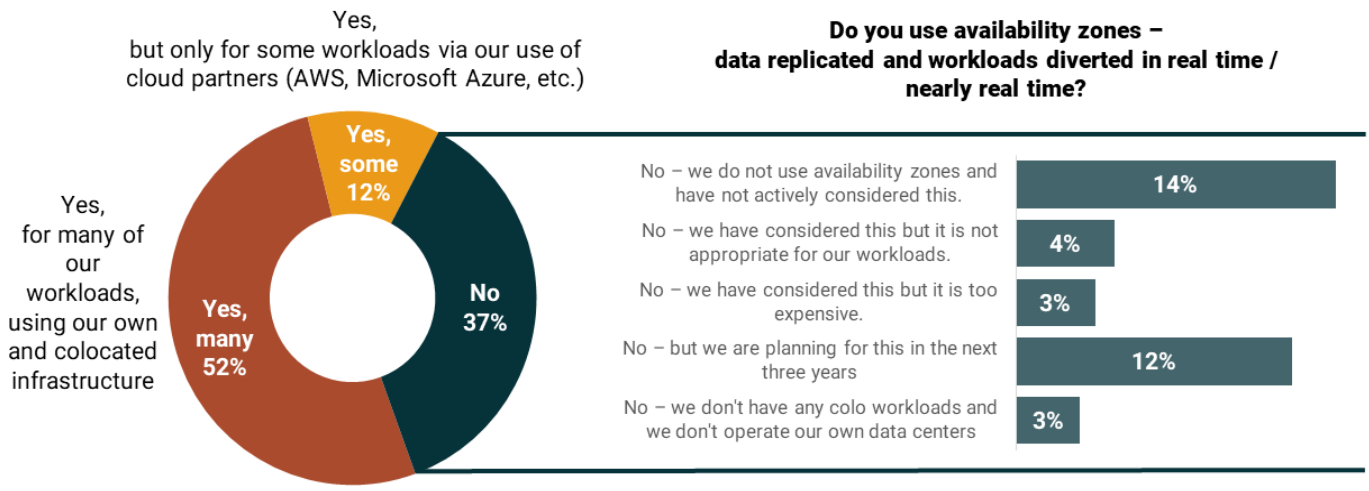


Figure 16. **Most use availability zones**

Most assess criticality, but not regularly

Uptime Institute's Tier Standards, widely adopted in the data center industry, set out four design levels — ranging from basic functionality to fault tolerant — for data centers. As a general (but not hard) rule, the cost and level of redundancy increases with each Tier.

The Tier Standards should enable data centers owners to build a data center that is appropriate to their business requirements. Those that build to a level beyond the business requirements of the IT clients may be paying for more than they need. Those building below may be taking unacceptable risks with IT service availability or, in the case of colocation, deterring potential clients requiring assurance that downtime is sufficiently low.

Identifying the appropriate level requires an understanding of the actual needs of the IT services and applications that will run in the data center and, often, a formal process for assessing the criticality needs of the service and the business impact of downtime.

Our survey suggests that about a third of organizations don't carry out this analysis at all; instead, they rely on the underlying IT infrastructure. The remainder do carry out this important analysis — but of these, only two-thirds regularly review it (see Figure 17).

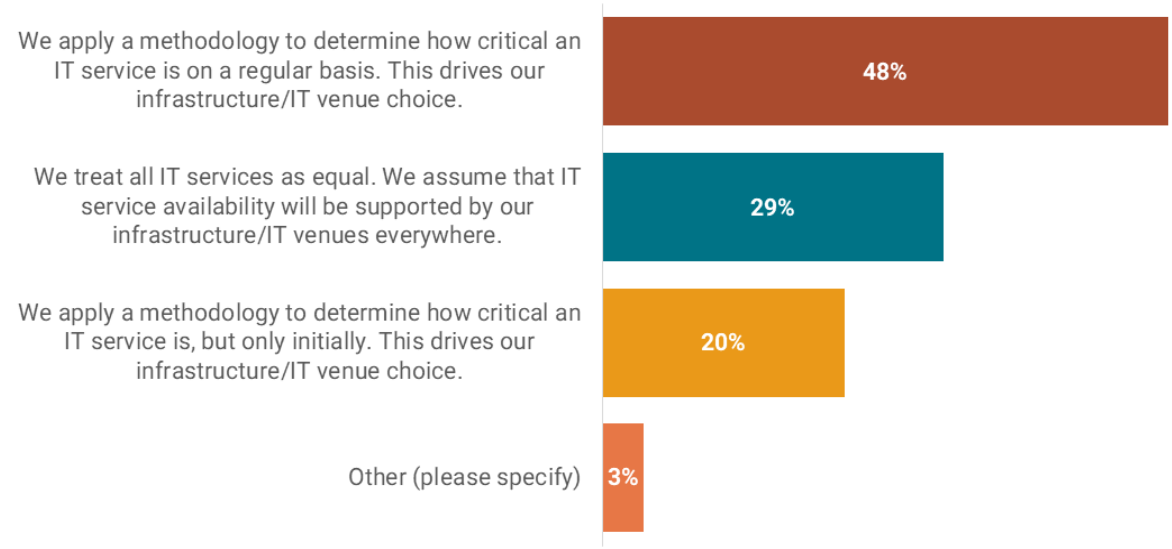


Figure 17. **Many regularly assess the criticality of IT services for best venue decisions**

Uptime Institute believes that regular criticality assessments for both the infrastructure and the IT/data is important. In earlier times, with simpler IT, this was rarely straightforward. It has become even more challenging in recent years, for several reasons:

- Applications and services that were not once deemed critical (say, issuing an airline boarding pass) are becoming increasingly important, if not essential, to doing business.
- Distributing workloads across availability zones changes the risk of failure at a single site.
- Distributed cloud applications and containers may have interdependencies with other distributed services and data that are difficult to track. The supply chain of applications and data has become a dynamic mesh.

Our research suggests that many outages may be due to “asymmetric resiliency” or “creeping criticality,” where the resiliency needs of the applications/service profile have changed (in almost all cases, increased) but the underlying operational and designs of the infrastructure have not been sufficiently upgraded. Similarly, within IT organizations, processes such as synchronization, replication, backup and disaster recovery may have fallen behind business needs.

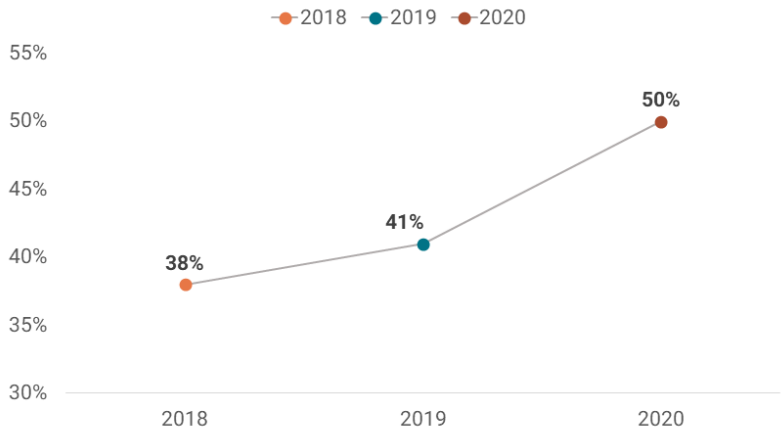
Workforce pipeline: Work in progress

Data center owners and operators have long been concerned about a lack of qualified, available staff. The number, size and job needs of facilities have grown very rapidly — faster than most recruitment practices have adapted. There is also the threat of a "silver tsunami," when a cohort of experienced professionals soon retire and leave behind numerous unfilled jobs and a vast experience gap.

For at least a decade, Uptime Institute Network members and others have said that recruiting staff is difficult. The reasons cited include:

- There is a general shortage of engineers in many countries.
- In many regions, demand for staff created by new builds has increased dramatically.
- The sector is largely invisible. Most students/members of the general public have a low level of familiarity about data centers or the sector's career paths, which causes them to overlook careers some would find rewarding.
- Gender and other demographic imbalances may have deterred some candidates from seeking a career in the field and reduced the size of the labor pool. For those working in the industry, a lack of diversity can create staff retention issues (among other negatives).
- Fundamental aspects of the job — such as being scheduled for second- or third-shift work (especially in 24/7 operations); limited pay; and high-skill requirements for even entry-level positions — can reduce the number of qualified applicants.

The result is a shortage of skilled candidates that is nearing a crisis. As Figure 18 shows, half of respondents said they had difficulty finding qualified candidates for open jobs — up from 41% a year earlier and 38% in 2018.



*Are you having difficulty finding qualified candidates for open jobs?**
*Yes answers only

Source: Uptime Institute Global Survey of IT and Data Center Managers 2018 (n=202), 2019 (n=190) and 2020 (n=218)

Figure 18. **The proportion having difficulty recruiting new staff is growing**

The entry of the hyperscales into the market has had a dramatic effect in some areas. Of those surveyed, 16% say they are having difficulty retaining staff, as they are being hired away competitors (i.e., doing data center work), while 11% said they are having difficulty retaining staff, as they are being hired by non-competitors (i.e., doing non-data center work).

Some operators are recruiting outside the industry and retraining staff much more actively. Many are turning to outsourced resources. While most (54%) say they primarily use an in-house staffing model, a significant portion (39%) use a mix of in-house and outsourced staff (see Figure 19). A small portion primarily rely on an outsourced staffing model.

Does your organizational primarily use an in-house or outsourced staffing model for data center operations?

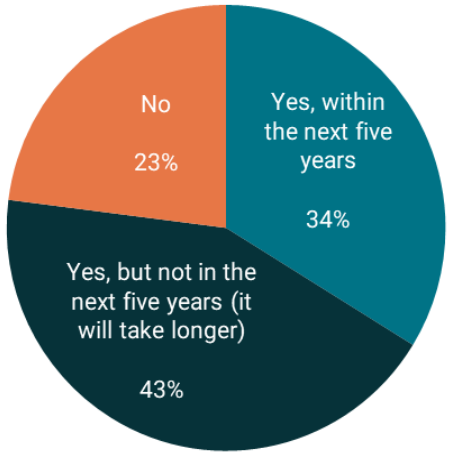
Source: Uptime Institute Global Survey of IT and Data Center Managers 2020 (n=429)

Figure 19. **Many use at least some outsourced staff**

These results appear to indicate an uptick in the use of outsourced staff in recent years. We asked a similar question in our 2017 survey, but the results are not directly comparable because the answer options differed. In 2017, more than 70% said they used an in-house staffing model (with the remainder using outsourcing). This suggests the use of third-party staff today is more prevalent than just a few years ago.

What about the use of technology and automation to reduce staffing requirements? One of the most promising approaches in this area is the greater use of AI for data center operations. Uptime's own research does suggest that certain tasks can be effectively automated (powered by AI) and that smaller data centers can operate without any staff at all. But in larger data centers, AI-driven approaches can have limits and, today, are unproven in many areas.

Still, most in the industry agree: Nearly 80% believe that AI will reduce their data center operations staffing levels. As shown in Figure 20, a third think this will be achieved within the next five years; most think it will take longer. These results are almost identical to a year ago when we asked the same question.



Do you believe artificial intelligence (AI) will reduce your data center operations staffing levels in the next five years?

Source: Uptime Institute Global Survey of IT and Data Center Managers 2020 (n=435)

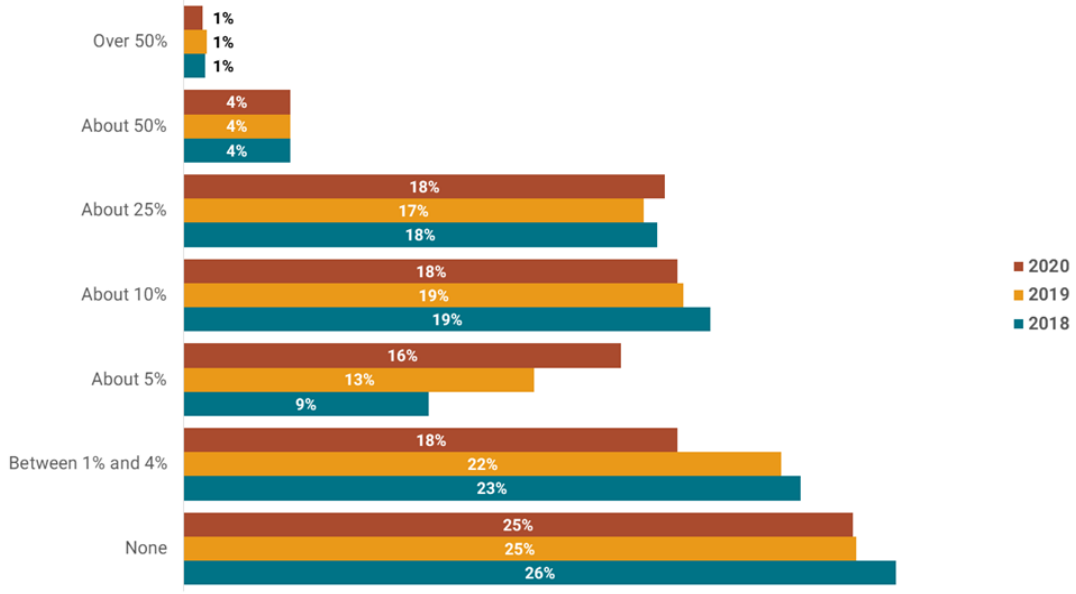
Uptime Institute | INTELLIGENCE

Figure 20. **Most say that AI will not reduce staffing levels within the next five years**

Gender imbalances persist

The lack of women working in most all roles in the data center industry has been widely discussed. Whatever the causes — and they are deep, wide and long-standing — there is a clearly a general, if not urgent, desire across the industry to redress the gender imbalance. For some, this is primarily a commercial issue: A more inclusive approach would bring more people, and more skills, into the sector.

How prevalent is the gender imbalance? We've been asking that question for a few years and, as shown in Figure 21, the answers are more or less consistent. A sizeable minority have no women in their organization's data center design, build or operations teams; for those that do, women most commonly represent between 1% and 4%. Although there are small signs of change, it is in single digits over three years.



***What portion of your organization's data center design, build or operations staff is women? Choose one.** **

**All figures rounded*

Source: Uptime Institute Global Survey of IT and Data Center Managers 2018 (n=508), 2019 (n=470) and 2020 (n=432)

Uptime Institute | INTELLIGENCE

Figure 21. **The proportion of women in the data center industry remains low**

As shown in Figure 22, the majority of those surveyed (71%) think the data center profession offers the same career opportunities for men and women — a slight decrease from a year ago.

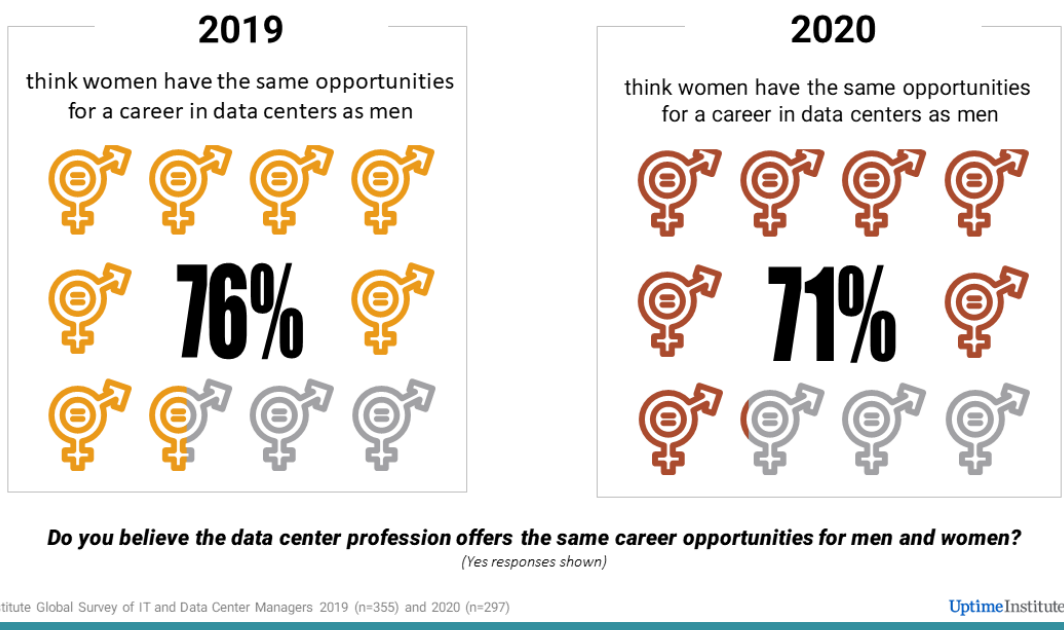


Figure 22. **Most think the industry offers the same career opportunities for women as for men**

In this year's survey, we also asked whether organizations would like to hire more women in its data centers; 40% of respondents say their organization would. However, there appears a clear disconnect between thoughts and action. As Figure 23 shows, only about a quarter have a plan or initiative in place to boost the hiring of women in their data center operations. As in previous years, most have no plans to do so.

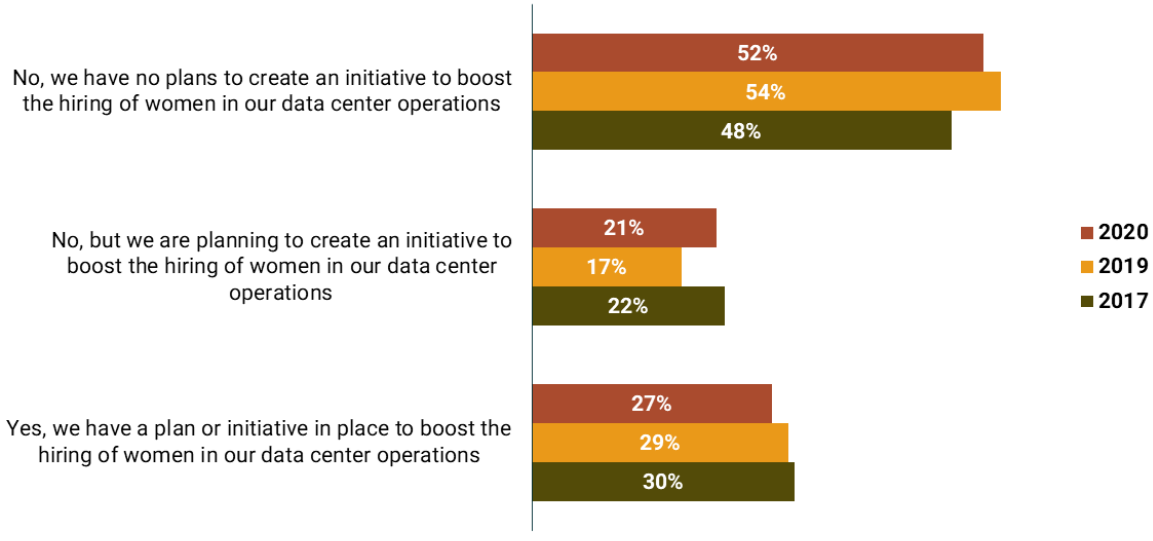


Figure 23. **Roughly a quarter have a plan to hire more women data center ops staff**

Water usage: Only half collect data

The threat of water scarcity is growing in a number of regions. Due to population growth and prolonged droughts, demand for fresh water globally will exceed available supplies by 40% by 2030, according to the United Nations. For data centers, which typically use up to 8 million gallons of water per year per MW, water scarcity threatens future growth and operational reliability.

Many large data center operators have stepped up their efforts to conserve water during the past decade, but progress across the industry has generally been slow. Some of the largest data center owners have only recently begun collecting comprehensive water usage data across their portfolios; others are still working to do so.

In our survey, only half of respondents say their organization collected water usage data for their IT/data center operations (Figure 24). Of those that do, four of five said they do so at the site level (i.e., for each data center), rather than at an aggregated regional or portfolio level.

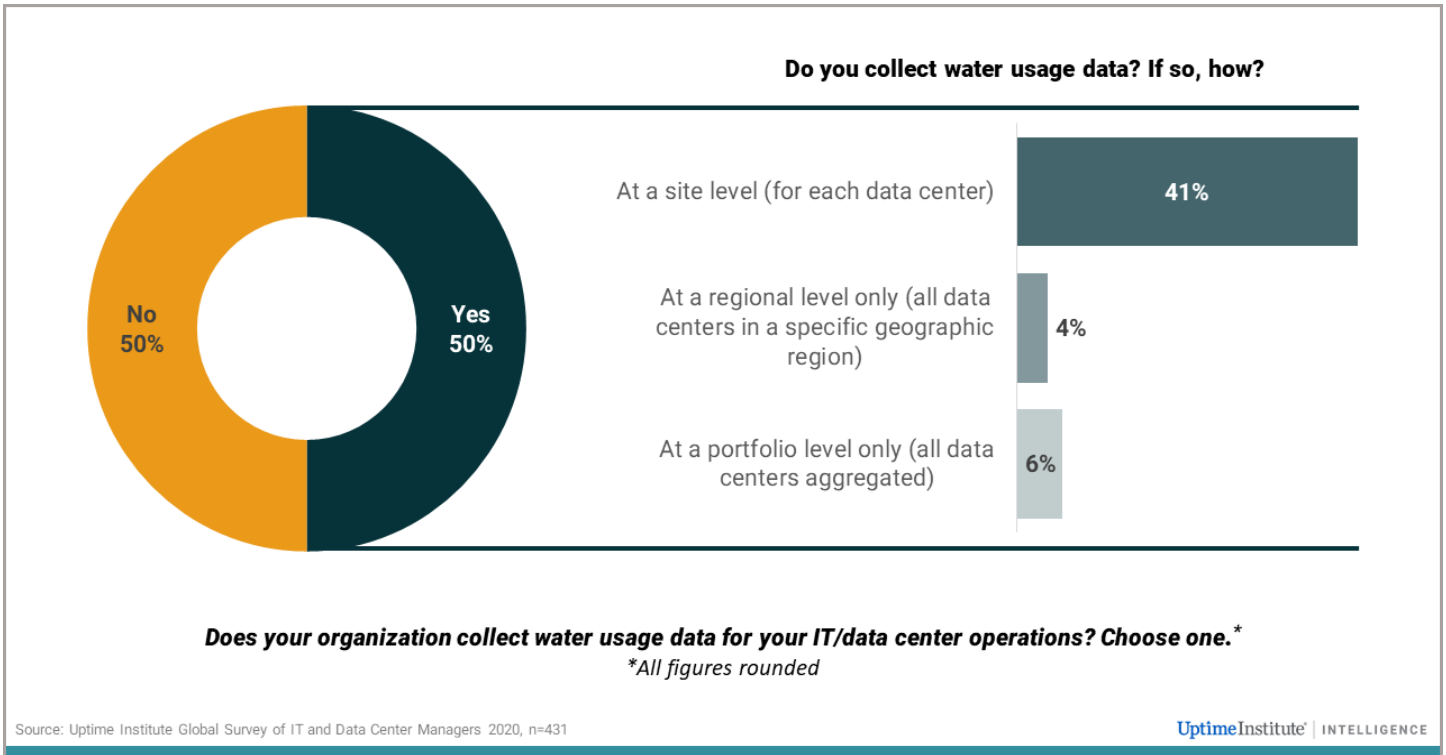


Figure 24. **Only half collect data center water usage**

Measurement is, of course, just the first step. Operators must assess their water risks (and prices) and develop a water resiliency plan. In a water-scarce future, it is not enough to just move or site data centers in regions with adequate water supplies: changes in climate and population growth can cause those regions to become water stressed in coming decades — certainly during a facility's lifespan.

We expect there will be increased focus and investment in alternatives to water-based cooling systems in the coming years, with more efficient air-cooled chilling and an increase in the use of "near-zero" water consumption designs.

Appendix

2020 Annual survey demographics

Uptime Institute's Global Survey of IT and Data Center Managers, now in its tenth year, is conducted annually online and by email. The 2020 survey was conducted in March and April 2020.

Respondents are separated into two groups: owner/operators of data centers; and suppliers, designers and advisors. This report focuses on the findings from the owner/operator survey — people responsible for managing infrastructure at the world's largest IT organizations. Job titles include senior executive, IT management, critical facilities management and design engineer.

As Figure A1 shows, the participants represent a wide range of industries in multiple countries, with just over half coming from North America and Europe. About a third of the respondents work for professional IT/data center service providers (i.e., staff with operational or executive responsibilities for a third-party data center, such as those offering colocation, wholesale, software or cloud computing services).

Job Function

Critical Facilities Management	34%
IT Management	23%
Senior Executive (VP or C-level)	24%
Design Engineer	19%

Verticals

Colo/Multi-tenant Data Centers	27%	Manufacturing	4%
Telecommunications	18%	Healthcare/ Pharmaceuticals	4%
Financial	13%	Retail/Wholesale/ Distribution	4%
Software and/or Cloud Services	11%	Education	3%
Government	7%	Transportation	3%
Utilities/Energy	5%	Media	1%
		Insurance	1%

Location

U.S. and Canada	28%
Europe	25%
Asia-Pacific (not China)	16%
Africa and Middle East	14%
Latin America	13%
Russia and CIS	3%
China	1%

Uptime Institute | INTELLIGENCE

Source: Uptime Institute Global Survey of IT and Data Center Managers 2020, n=846

Figure A1. **Respondent demographics: Uptime Institute Global Survey of IT and Data Center Managers 2020**

A total of 846 end users registered for the survey — this means they answered at least one question. Because respondents were not required to answer all questions, the number of respondents for individual questions ("n") varies widely. Previous survey findings are available on the Uptime Institute Network member website, Inside Track.

The results of the designer/supplier/advisor survey will be available in September 2020. We will also publish some particular slices ("cuts") of findings (e.g., by region, by industry) in the form of Uptime Institute Intelligence Notes in the coming weeks.

If you have queries, comments or seeking further insights, please contact intel@uptimeinstitute.com.

2020 Pandemic-related surveys

During April and July, Uptime Institute conducted additional industry surveys on the impact of the COVID-19 pandemic. The results of the first research survey are discussed in the following posts:

- **COVID-19: Critical impact and legacy**
- **COVID-19: What worries data center management most?**
- **Pandemic is causing some outages and slowdowns**

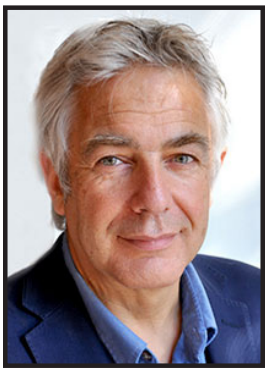
We will publish additional findings in the coming weeks and months.

For more information on our surveys or Uptime Intelligence, contact Rhonda Ascierto, Vice President, Research (rascierto@uptimeinstitute.com) or Brenda South, Vice President, Communications (bsouth@uptimeinstitute.com).

Uptime Institute® | INTELLIGENCE

ABOUT THE AUTHORS

Rhonda Ascierto is Uptime Institute's Vice President of Research. She has spent two decades at the crossroads of IT and business as an analyst, speaker, adviser, and editor covering the technology and competitive forces that shape the global IT industry. Contact: rascierto@uptimeinstitute.com



Andy Lawrence is Uptime Institute's Executive Director of Research. Mr. Lawrence has built his career focusing on innovative new solutions, emerging technologies, and opportunities found at the intersection of IT and infrastructure. Contact: alawrence@uptimeinstitute.com

ABOUT UPTIME INSTITUTE INTELLIGENCE

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 intel@uptimeinstitute.com.

ABOUT UPTIME INSTITUTE

Uptime Institute is an advisory organization focused on improving the performance, efficiency and reliability of business critical infrastructure through innovation, collaboration and independent certifications. Uptime Institute serves all stakeholders responsible for IT service availability through industry leading standards, education, peer-to-peer networking, consulting and award programs delivered to enterprise organizations and third-party operators, manufacturers and providers. Uptime Institute is recognized globally for the creation and administration of the Tier Standards and Certifications for Data Center Design, Construction and Operations, along with its Management & Operations (M&O) Stamp of Approval, FORCSS® methodology and Efficient IT Stamp of Approval.

Uptime Institute – The Global Data Center Authority®, a division of The 451 Group, has office locations in the US, Mexico, Costa Rica, Brazil, UK, Spain, UAE, Russia, Taiwan, Singapore and Malaysia. Visit uptimeinstitute.com for more information.

All general queries:
Uptime Institute
5470 Shilshole Avenue NW, Suite 500
Seattle, WA 98107 USA
+1 206 783 0510
info@uptimeinstitute.com

