RISK & RESILIENCY

UI INTELLIGENCE REPORT 53

Resilient edge data centers

Unstaffed, affordable and flexible

Edge data centers need to be resilient to failures. This is commonly achieved by using redundant on-site infrastructure, possibly combined with software-based, distributed resiliency. Generators remain important, but batteries, solar panels and resilient remote monitoring systems have a growing role as well.

Uptime Institute | INTELLIGENCE

AUTHORS

Tomas Rahkonen, Research Director of Distributed Data Centers, Uptime Institute

Andy Lawrence, Executive Director of Research, Uptime Institute

UII-53 V1.0M PUBLISHED OCTOBER 1, 2021 LAST UPDATED OCTOBER 1, 2021

35-45 MINUTES TO READ

Synopsis

While physical (site) infrastructure redundancy remains the most common approach to edge data center resiliency, there is growing interest in distributed (multisite) resiliency techniques. Several potential benefits are driving this, such as falling costs of edge capacity; the ability to optimize service delivery by shifting workloads between sites; and the need to keep IT services running even when one or multiple sites are lost due to, for example, a regional issue or natural disaster. Another set of objectives revolves around the need for remote operations, low maintenance and environmental sustainability. This report explores the technical developments that move edge data centers closer to these goals.

- Nine of 10 organizations foresee using N+1 (or higher) physical edge site infrastructure redundancy, indicating distributed resiliency is an enhancement, not a replacement.
- Engine generators will stay unchallenged as a power supply for remote/harsh locations and backup of larger, critical IT loads. Operators of smaller edge sites will gradually favor alternative sources, chiefly batteries.
- Some edge sites use photovoltaic power as the primary energy source. By 2023 or 2024, using solar panels should be common.
- Remote monitoring systems are key to running unstaffed sites and for distributed resiliency. These systems will require dedicated power sources and out-of-band network management.

Contents

Illustrations

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 research@uptimeinstitute.com.

Introduction

Uptime Institute's research shows that demand for edge data centers is starting to grow across the world.

Demand is particularly strong for shared (colocation) edge facilities in North America. Organizations and suppliers alike expect growth (from low numbers) to continue across multiple industry verticals. (See our recent report **Demand and speculation fuel edge buildout** for further discussion of the demand picture.) As the number of edge data centers grows to complement larger sites, operators see a need for new thinking around infrastructure resiliency.

Edge resiliency requirements vary significantly by business case — arguably more so than for regional/departmental and core data center facilities. For example, for organizations running critical workloads at cloud providers, on-premises edge capability plays a critical role in maintaining the availability (including acceptable performance) of business-critical applications, via diverse connectivity paths and local instances of data and software.

In other cases, IT and network services may not be deemed critical, either because the type of application it supports is not business critical, or because the loss of the edge capability doesn't lead to data or application downtime (it is transparent to end users).

In certain scenarios, sustained availability is probably not even useful during a power failure — for example, a local node on the content delivery network that serves businesses and households that are all similarly offline due to a blackout.

This report presents several approaches to edge resiliency, including both single-site and multisite techniques, followed by a discussion of on-site power and remote management. Please see **Appendix A** for a list of key companies currently active in edge development and **Appendix B** for our research methodology.

> "Edge resiliency requirements vary significantly by business case — arguably more so than for regional/departmental and core data center facilities."

Edge resiliency approaches

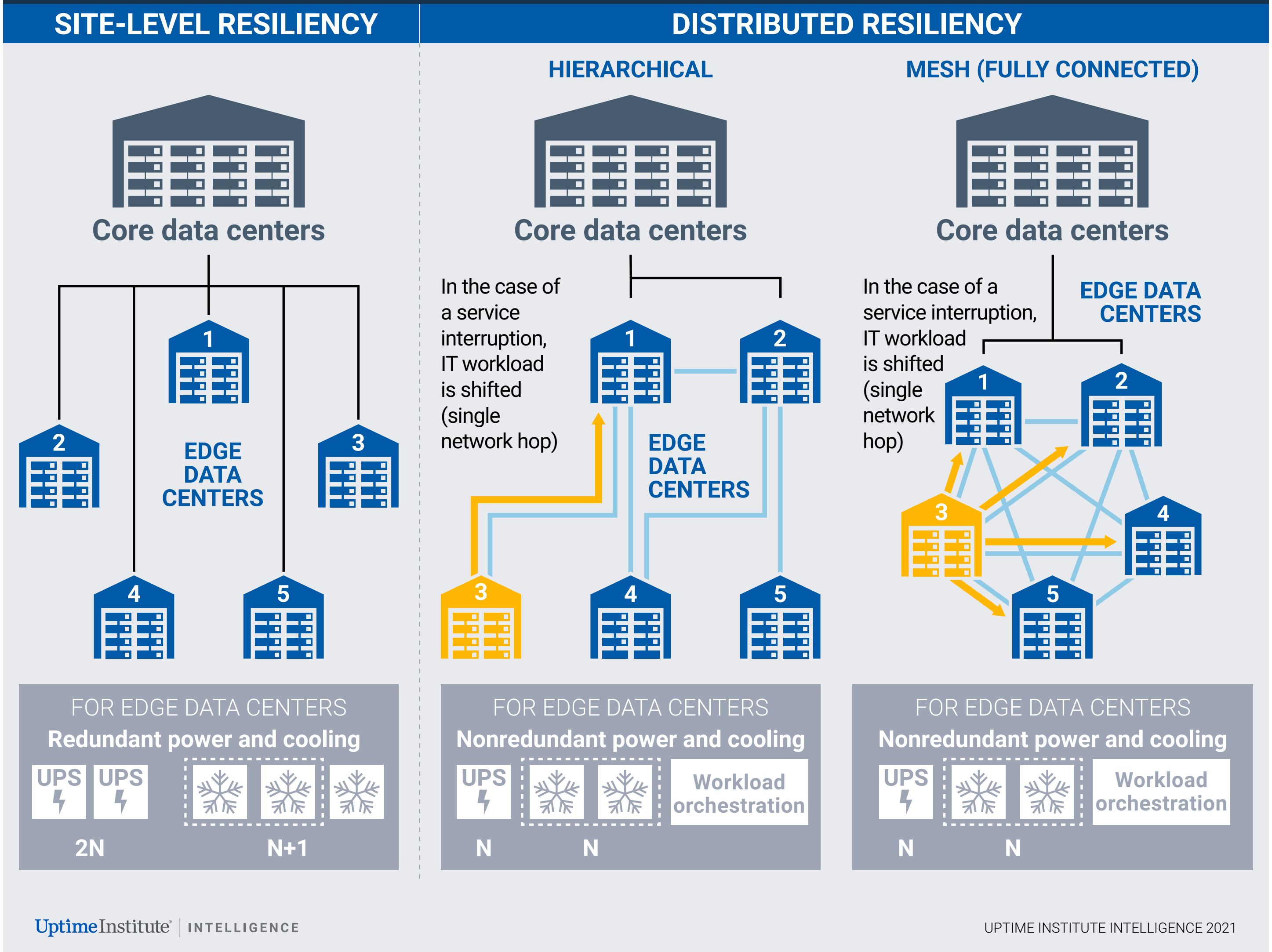
Resiliency in a technology context refers to an organization's ability to maintain acceptable service levels during failures and other disruptions to normal operations. Uptime Institute's research shows that data center outages have many causes: IT software/configuration changes; IT hardware failure; power and cooling system failures; networking failures; and human error are the most common. (For more on data center outages, see our recent report **Annual outage analysis 2021: The causes and impacts of data center outages**.)

The approaches to resiliency for edge data centers can be broadly divided into two types: site-level and distributed resiliency, as shown in **Figure 1.**

Site-level resiliency relies on redundant capacity components (including major

FIGURE 1

Different resiliency approaches are used for edge data centers


SITE-LEVEL RESILIENCY
DISTRIBUTED RESILIENCY
HIERARCHICAL
MESH (FULLY CONNECTED)
Core data centers
Core data centers
Core data centers
1
2
3
4
5
EDGE DATA CENTERS
In the case of a service interruption, IT workload is shifted (single network hop)
EDGE DATA CENTERS
In the case of a service interruption, IT workload is shifted (single network hop)
EDGE DATA CENTERS
FOR EDGE DATA CENTERS
Redundant power and cooling
UPS
UPS
2N
N+1
FOR EDGE DATA CENTERS
Nonredundant power and cooling
UPS
Workload orchestration
N
N
FOR EDGE DATA CENTERS
Nonredundant power and cooling
UPS
Workload orchestration
N
N
Uptime Institute | INTELLIGENCE
UPTIME INSTITUTE INTELLIGENCE 2021

equipment) for critical power, cooling and network connectivity — the approach widely adopted by almost all data centers of any size. Typically, this allows mission-critical facilities to continue operating at full capacity in the event of planned maintenance or component failures. The more resilient a data center is, the more (and wider in scope) concurrent failures it can survive with no degradation in service. Edge data centers using only site-level resiliency tend to run their own IT workloads independently from other edge data centers.

> “The more resilient a data center is, the more (and wider in scope) concurrent failures it can survive with no degradation in service.”

Edge data centers making use of distributed resiliency are connected and operated in a coordinated manner, commonly using either a hierarchical topology or a mesh topology to deliver multisite resiliency, as illustrated in **Figure 1**. The unit of redundancy in this schema is an edge data center — not a component or system. The two approaches are not mutually exclusive, although distributed resiliency creates opportunities to reduce component redundancy at individual edge sites without risking service continuity. **Figure 1** shows that a distributed architecture can provide resiliency despite nonredundant capacity components.

The key design feature that underpins distributed resiliency is the ability to shift most or all IT workloads and/or service requests from one data center to one or more edge sites. Although this capability is, at least in concept, not limited to edge sites, it is often difficult for even moderately large sites to adopt this approach, due to huge volumes of data traffic and the vast number of software applications involved, many of which do not easily support such transfer mechanisms. Edge sites not only handle relatively smaller amounts of data but also run a narrower set of applications that are designed using more recent software development methods for distribution and scalability.

Unstaffed centers require remote systems

Because edge data centers are typically unstaffed, resilient remote monitoring and good network management/IT monitoring are key, allowing off-site managers to detect disruptions and capacity limitations. (See **Remote monitoring and control** for a more detailed discussion.)

In a hierarchical edge topology, sites are directly connected, predominantly in an upstream-downstream relationship (“north-south” in networking jargon), where some nodes may act as a distribution hub and connect to a higher number of downstream sites (e.g., 3, 4 and 5 in **Figure 1** – Hierarchical). This architecture is the simpler of the two distributed resiliency approaches illustrated. It is easier to fit into existing network topologies, but its setup can reduce the flexibility to lift and shift IT workloads — due to longer latencies and network capacity limitations — than the alternative mesh approach. Online content distribution and multipathing for accessing critical services are two major examples that fit such an edge organization.

A smaller mesh topology can be fully connected (see **Figure 1** – Mesh [fully connected]), where all sites have direct network links to each other without intermediate hops. A mesh network can also be partially connected where many but not all sites are directly linked; this also allows for larger meshes. A fully connected mesh network maximizes the flexibility to shift IT workloads, as there are many fallback edge data centers within a single network hop — but this flexibility typically involves higher costs for networking and greater operational complexity, due to the higher number of network links. The architecture is also likely to involve more complex load balancing and shifting, with differing and managed levels of replication according to the demands of the workloads.

Resiliency through mesh topology

A mesh topology provides a more robust distributed edge resiliency schema, where the pooling of multiple sites creates a resilient "virtual" data center spanning a relatively small area, such as an industrial estate or a city. End users may prefer distributed local resiliency (potentially on top of physical site infrastructure redundancy) for mission-critical applications (such as industrial data and control systems), latency-sensitive commercial services (such as online gaming), or emergency and public security services.

Operators can benefit from using a mix of site-level resiliency and distributed resiliency (as many operators of large, core data centers do now). For an edge data center running highly critical IT workloads, comprehensive site-level resiliency can be enhanced using distributed resiliency approaches to cope with serious failures, adverse natural events or physical attacks.

In either of these major approaches, edge nodes in any resilience architecture need not be — and probably won't be — identical in their resilience, by design. In a hierarchical edge topology, hub nodes will typically have more site-level resiliency, including generators, to maintain services even in the event of a grid failure.

The same goes for a mesh edge architecture, where some nodes are not equal to others but are backed by a generator in case of a power loss to the entire area. For this reason, one or a few nodes in an edge mesh may reside in a larger data center, such as a multi-megawatt colocation facility. The decision to increase resiliency at a particular site may depend on IT service revenue and/or if the hub is more important for overall service availability.

Distributed resiliency depends on reliable fiber networks for IT traffic. To guarantee multiple paths between edge nodes, redundant paths must be truly independent. Remote monitoring should extend to the power supply and performance of the communication system and should include port-level controls/thresholds for latency and congestion. 5G radio is expected to become the de facto standard for last-mile connectivity between edge data centers and client devices.

The following sections take a deeper look at single-site resiliency and distributed resiliency.

Site-level resiliency

Uptime Institute's research shows that site-level resiliency techniques, in terms of redundant power and cooling systems, are commonly used for edge data centers. The use of redundant infrastructure for edge data centers is based in part on tradition, re-using design principles known to work for larger data centers, and in part on the expectations of IT tenants (whether internal or external).

> "More than nine of 10 organizations foresee using redundant (N+1 or more) power and cooling systems."

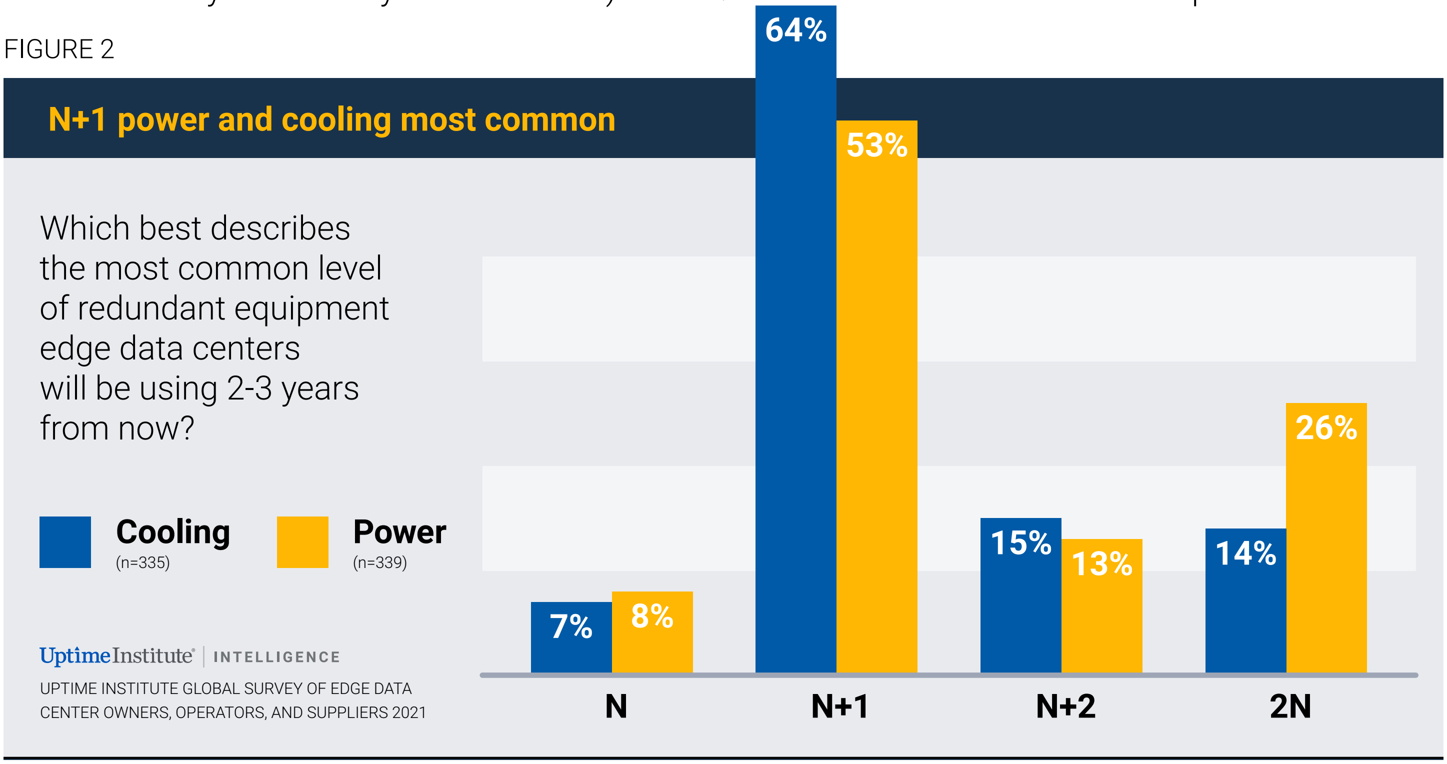
More than nine of 10 organizations foresee using redundant (N+1 or more) power and cooling systems for their edge data centers in the next two to three years (see **Figure 2**), according to Uptime Institute research. (For small sites, 2N redundancy can be very similar to N+1.)

Certain organizations use N+2 for their larger edge data centers. (In some recent cases, this is being additionally driven by COVID-19, which is forcing organizations to handle longer periods without site maintenance visits.) Other organizations believe that consistent use of N+1 is more appropriate, as it requires and drives implementation of better operational, staffing and spare parts handling procedures.

While Uptime Institute's survey foresees common use of site-level resiliency, our research also shows that there are organizations, including certain big cloud companies and edge colocation providers, that instead focus on simplifying the design of each edge data center to a bare minimum (N), with site-level resiliency limited to a few critical mechanical components (e.g., pumps). Such organizations typically deploy fleets of edge data centers that work in concert to optimize overall performance metrics such as cost-of-compute or service

FIGURE 2

N+1 power and cooling most common



UPTIME INSTITUTE GLOBAL SURVEY OF EDGE DATA CENTER OWNERS, OPERATORS, AND SUPPLIERS 2021 availability, while making use of distributed resiliency topologies and approaches to achieve greater uptime levels and lower physical infrastructure costs. In addition to reduced capital equipment costs, a simplified data center design may reduce maintenance requirements, eliminate some sources of technical errors, and decrease the likelihood of human errors.

Core data centers now commonly use sophisticated cooling systems, operated by specialist staff, to reduce power consumption and thereby decrease utility power costs and environmental impacts. For most edge sites, operational simplicity is imperative, and this may involve restricting the use of complex cooling systems. At suitable geographical locations, direct free air cooling may allow operators to combine power savings with operational simplicity — a practice made more possible where IT loads and power density are relatively low. Notably, this is common practice for telecom network sites in certain geographies.

The recent ASHRAE technical bulletin **Edge Computing: Considerations for Reliable Operation** warns that it is far more difficult (and possibly more expensive) to control the environment in edge data centers, particularly when field technicians are present. Key points of advice from ASHRAE include:

> "For most edge sites, operational simplicity is imperative."

- Where possible, service edge data centers only during moderate weather, and monitor humidity/condensation/temperature during servicing.
- Be aware of local pollutants.
- Monitor remotely for air quality, including corrosives and particulates.

The ASHRAE technical bulletin is further discussed in Intelligence note 75, **Why ASHRAE is concerned about edge data centers**.

In the context of both resiliency and maintenance, direct liquid cooling (DLC) is a promising option for small, energy-dense edge data centers. DLC techniques, whether cold-plate or immersion systems, offer several advantages, but the most valuable benefits for edge sites must be reduced failure rates (for both IT hardware and facilities equipment) and low maintenance needs due to fewer components (particularly fans and compressors). However, Uptime Institute's research shows that uptake of DLC for edge sites has been slow. There is not a single reason for this, but rather the friction created by the combined business and technical complexities of making the change in both supply chain (e.g., limited choice in IT systems or a lack of standards) and operational practices (e.g., how to service liquid-cooled servers or how to address a failure in the DLC systems). Initial costs may also be higher.

Distributed resiliency

Distributed resiliency, using software and networks to move IT workloads and redirect traffic between data centers to avoid/recover from disruption at a certain site, has until recently been predominantly used by large hyperscale operators, large IT-oriented enterprises, and service providers. Certain colocation providers now deploy critical elements of this approach to meet their customers' ambitions to build similar capabilities. These include providing application program interface (API) access to software-defined networks and allow interconnections between edge data centers, bare metal servers and storage; and providing rich, real-time monitoring data. Programmatic access allows software systems (applications and operational tools) to monitor and automate workload placement based on telemetric data about the quality of service and to make policy settings related to different sites and services. Certain edge colocation providers are working with software partners to provide operational tools, including orchestration systems, that support distributed resource scaling and, in turn, resiliency.

Software orchestration systems are key

Distributed resiliency depends on the use of remote monitoring systems (see **Remote monitoring and control**) to collect operational status and utilization data from edge data centers. The data is required (in near-real time) by software orchestration systems, which then use it as one input for any decision to shift workloads between edge data centers. Other information might include the availability of — and trending data on — capacity across sites and the performance of edge IT systems. A detailed discussion of orchestration systems is beyond the scope of this report.

Using distributed resiliency can bring several benefits to a network of edge data centers. These include:

- Improved IT service availability when there is a serious failure or some other disruption.
- Enhanced business agility resulting from the flexible placement and shifting of IT workloads, based on quality-of-service needs.
- The opportunity to decrease cost, physical size and complexity of edge data centers due to less need for site-level redundancy.
- The opportunity to use fewer and relatively less-skilled staff (generalist power/mechanical maintenance contractors), thanks to reduced site infrastructure complexity, improving technical support response times and lowering operational costs.

Barriers to using distributed resiliency include:

- Increased cost and complexity of networks and monitoring systems.
- Costs to rearchitect workloads not originally developed for distributed execution; for some legacy workloads, this can be an insurmountable obstacle.
- New risks that may be introduced — for example, resiliency becomes dependent on networks, remote monitoring systems, and IT orchestration tools to shift workloads as needed; risk of undocumented behavior of control systems, creating or exacerbating service problems.

When adopting a distributed resiliency approach, organizations should expect challenges throughout the initial validation and pilot stages, largely due to the increased software and networking complexity. Generally, it is considered best practice to start a distributed resiliency program with deployments for less-critical workloads and gradually expand the scope.

Uptime Institute's research shows that more than eight of 10 edge data center owners, operators and suppliers expect that distributed resiliency will be somewhat commonly or very commonly used two to three years from now (see **Figure 3**).

Notably, the same survey respondents mostly foresee using N+1 or higher site-level resiliency (see **Figure 2**). This suggests many organizations plan to use a mix of site-level resiliency and distributed resiliency for their edge data centers. It is possible other respondents would agree that distributed resiliency makes sense at the edge, but they have not yet been exposed to projects that involve a need to combine or select between site-level resiliency and distributed resiliency.

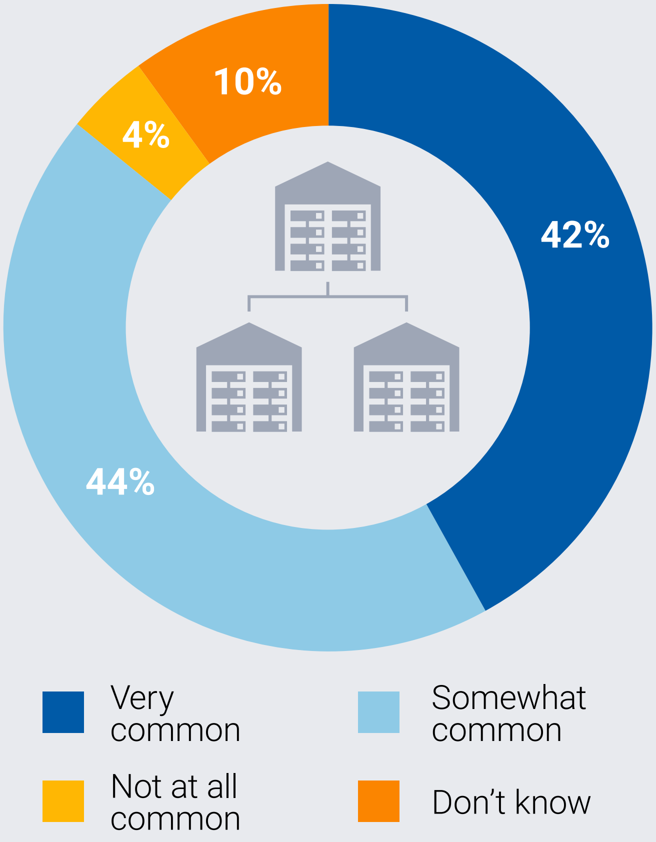
> "Start a distributed resiliency program with deployments for less-critical workloads and gradually expand the scope."

For distributed resiliency to work, each edge data center needs to have sufficient capacity to accommodate the IT workloads shifted from other edge data centers. This can be achieved in different ways, including configuring each edge data center with spare capacity or using load prioritization to run additional services (as discussed in **Site-level resiliency**).

FIGURE 3

Distributed resiliency common at the edge

Flexible shifting of IT loads and traffic between edge data centers is often referred to as distributed resiliency. How common do you think it will be for edge data centers to utilize distributed resiliency 2-3 years from now?



Uptime Institute | INTELLIGENCE

UPTIME INSTITUTE GLOBAL SURVEY OF EDGE DATA CENTER OWNERS, OPERATORS, AND SUPPLIERS 2021 (n=358)

For example, it may mean running these services with reduced functionality and/or delayed or reduced processing during incidents. It could even be acceptable to increase latency temporarily and run less-critical workloads from core data centers.

The requirements of the workload will dictate the architecture used, as well as the level and type of replication and synchronization between sites. Workloads that make use of distributed resiliency can be broadly divided into two categories: active-passive and active-active.

Active-passive architecture workloads are designed so that the primary site processes production loads at a single (active) edge data center (regardless of edge resiliency topology). However, such workloads also have passive (dormant) reserves at one or more other edge data centers. These passive deployments are typically smaller or downsized from the active site, to save scarce IT platform resources (primary storage and processing), but the secondary sites are still useful to avoid a lengthy cold start where the workload first needs to be installed from scratch. At failover to a (minimized) passive deployment, the edge data center creates all needed machine instances and conducts tests before the IT service goes live again.

The recovery time for active-passive workloads can be a few minutes. Failures impacting the active workload (active edge data center) can be detected using load balancing software, which at failure redirects traffic to the previously passive edge data center now taking over the workload. For distributed resiliency to work without severely negative impacts on IT services, there may be a need to replicate data at high speed from the primary site to secondary sites. The amount of data replicated and the update frequency can vary between workloads depending on their criticality, among other things.

Active-active architectures are designed so that production workloads are received and processed at multiple edge data centers. Load balancing software continuously spreads out the workload traffic among edge data centers to mitigate failures or to dynamically manage edge data center capacity (e.g., to prevent any single edge data center from reaching its capacity limit). A virtualized storage platform shared among sites is used to continuously mirror data while allowing concurrent write access from different edge data centers. Active-active workloads can require significant investment in software design, validation and repeat testing, but also bring the full benefits of distributed resiliency, including low and guaranteed recovery time (seconds), flexible use of infrastructure investments, and efficient use of capacity. This technology is used at scale by hyperscale operators, switching loads among their core data centers.

A major business challenge for engineering the infrastructure for distributed resiliency is that much of the distributed edge infrastructure will be deployed in colos — in other words, it will be operated by companies seeking to serve a diverse customer set. A mix of applications from a set of customers will likely mean diverse capabilities are required, ranging from no explicit edge resiliency design in software to the use of traditional snapshot backups to other sites or even scale-out software that creates instances across multiple edge nodes in the same area, based on network performance and end-user demand. This suggests that most edge operators will continue to build in, and pay for, infrastructural resiliency in their edge footprint, with N+1 or higher physical site infrastructure redundancy.

“Most edge operators will continue to build in, and pay for, infrastructural resiliency in their edge footprint”

On-site power sources

Diesel engine generators continue to be standard for primary on-site power in data centers, with the power grid used as a lower-cost alternative. While this remains the case for many (if not most) edge facilities as well, some edge operators have started taking advantage of the ongoing advances in battery and photovoltaic technologies. Avoiding the installation of diesel generators can reduce build costs and, perhaps more importantly, achieve better environmental credentials — which helps with permitting and attracts customers. **Figure 4** illustrates some of the options for on-site power. Other on-site power sources exist or are possible, including fuels cells and wind turbines, usually used in conjunction with some form of energy storage.

Engine generators

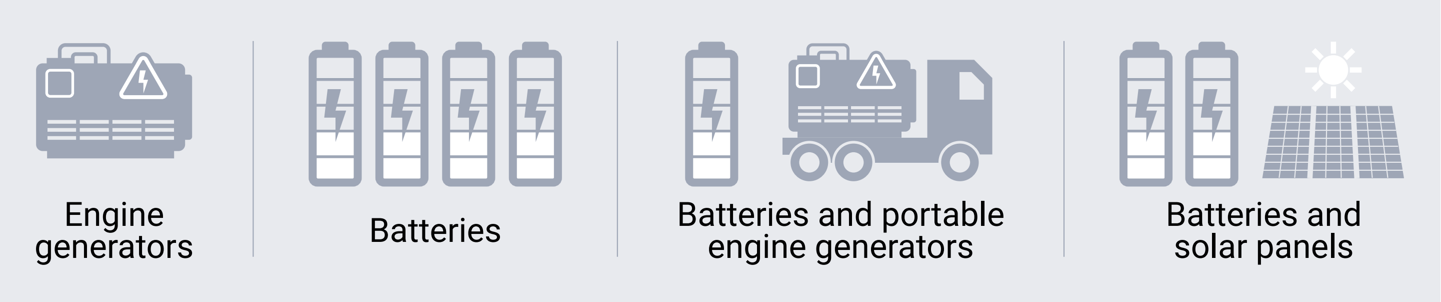
Uptime Institute's research shows that, despite their various environmental and maintenance challenges, engine generators are common at edge data center sites. Enterprises, colocation providers and telecommunication firms all greatly rely on generators to guarantee edge services availability — often in redundant configurations when no single point of failure is a requirement. Smaller edge data centers with an IT load of a few tens of kilowatts (kW), often for cost or space reasons, are less likely to use engine generators.

Compared with larger (regional and core) data centers, many edge sites must meet added demands on generator installations. For example, those in urban areas with tight spaces must comply with stringent regulations on noise, pollution and fire safety. At such locations, double-skinned fuel tanks are commonly needed, together with an anti-spillage system, while noise requirements in some cities require engine generators to be installed indoors or in special noise-isolating containers.

Engine generators require regular maintenance and test runs to ensure reliable operation should a grid outage occur. Some data center operators avoid more complex electrical arrangements (e.g., paralleling synchronized engine generators) in locations where there is limited local availability of skilled maintenance contractors. Having capable maintenance contractors travel to such

FIGURE 4

A range of power sources are used at edge sites


Engine generators
Batteries
Batteries and portable engine generators
Batteries and solar panels

Uptime Institute | INTELLIGENCE
UPTIME INSTITUTE INTELLIGENCE 2021

remote sites increases costs and introduces the risk of lengthy repair times. Even if the capital expenditure for a 2N generator setup (delivering two independent power feeds) can be higher than for a solution using synchronized generators, many edge operators will likely prefer a 2N design for maintenance reasons.

Some edge data centers operate in areas where power grid quality is not reliable, especially when compared with core data centers that make use of highly reliable direct power feeds. On-site switching between power grids and engine generators at such edge sites is commonly done using a mechanical automatic transfer switch (ATS). As an ATS uses moving parts, it can be a cause of power failures and must be monitored. Certain edge data center suppliers, including Zella DC, use semiconductor-based switching to replace the mechanical switch and to enable more intelligent engine generator controls (e.g., allowing the genset to warm up while running on batteries before taking on load). A semiconductor-based switch can also provide rich monitoring data on engine generator performance. This is useful to determine its operational status, without a need to install or maintain separate sensors. Such switches, however, can be expensive.

> “Another factor is balancing battery capacity with field engineer response time, which can vary from two hours in urban areas to eight hours or more in remote areas.”

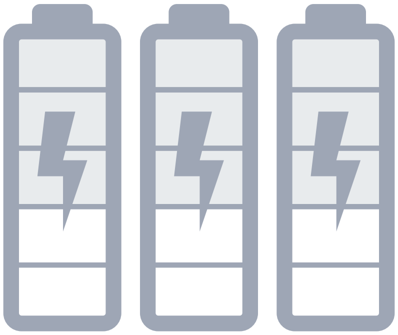
Very small data centers and some telecom sites (up to around 10-15 kW) in off-grid areas are commonly equipped with batteries to optimize the running of the engine generator (the primary power source). This is to reduce engine generator hours and fuel consumption by running the engine close to its efficiency optimum — similar in concept to the engine-battery operation in hybrid electrical cars. Such sites use engine generators with a capacity significantly higher than the site load so that they can supply the load and charge the batteries in parallel.

Using batteries with generators not foolproof

There are a few caveats, however. Because the batteries are in constant use, the battery type used must be designed for continuous charging and discharging (also called deep cycling); the lead-acid batteries commonly used for uninterruptible power system applications are not appropriate.

Another factor is the balance in battery sizing and field engineer response time. A common battery charging pattern includes turning on the engine generator when batteries are at a 40% charge level. If the engine generator fails to start, an alarm is issued and a maintenance contractor needs to reach the site and restore power within the time left to run on the batteries. Practical times for maintenance contractors to reach a site can vary from two hours in urban areas to eight hours or more in remote areas.

Batteries



For edge data centers not using permanently installed engine generators, batteries are the most common source of on-site power. The major distinction here is that while engine generators are used for on-site power generation, batteries are used for on-site power storage. In the event of a grid failure, batteries cannot be quickly "refueled" like an engine generator. Without a generation source, the data center will be powered only as long as the batteries last.

Some edge operators complement battery storage with portable engine generators that are transported to sites when the grid goes down (or is anticipated to go down) for an extended period. A few portable engine generators located at a central location may be used to support a network of edge data centers — provided the edge data centers are all accessible within a reasonable time and (ideally) are not all affected at the same time. Some providers may consider using portable batteries (rather than portable engine generators) for the same purpose, if and when such battery swap systems and services become available.

Operators often use batteries as the only on-site source of backup power for smaller (typically tens of kilowatts of power) edge data centers and telecom sites in areas with reliable power grids. Battery-only backup times tend to be in the range of two to eight hours, depending on site criticality, grid outage patterns, and the level (if any) of distributed edge resiliency. Batteries can be partitioned into two or more independent power feeds (A, B, ...) to increase fault tolerance.

The battery types most used today are valve-regulated lead-acid (VRLA) batteries and lithium-ion (Li-ion) batteries. Li-ion batteries bring many benefits: they require less space, are lighter, require less cooling, can cycle faster, and last longer. These factors, together with supplier warranty periods of up to five years, are making Li-ion batteries an increasingly popular choice for edge data centers. As costs per unit of capacity keep coming down, they present an increasingly attractive option for multihour backup at edge locations.

There are some challenges to using Li-ion batteries. Notably, Li-ion batteries represent an inherent fire risk compared with VRLA, according to research by the US-based National Fire Protection Association. This may mean higher fire protection costs and insurance premiums. Using Li-ion batteries raises ethical and sustainability (recyclability) concerns too, as discussed in Intelligence note 66, **Lithium-ion batteries: An ethical dimension?** These concerns don't apply to VRLA.

Load prioritization
Load prioritization (a collection of techniques that tier IT systems/applications by criticality) promises to be a powerful tool to extend battery run time for edge sites. An operator may choose different strategies for IT load prioritization, depending on the expected duration of a specific grid outage; for shorter outages, prioritization may not be needed. As an example, during a severe grid outage, battery power can be reserved for higher priority or higher service level agreement (SLA) IT workloads, while noncritical IT systems suspend their operation and power down. Consolidating instances onto fewer systems and throttling servers that stay online would bring further power reduction gains via vastly improved IT energy efficiency.

Such concerted efforts from both IT and facilities operations can potentially extend battery run time multiple times over the design target (e.g., 20 hours of scaled-back operation, compared with four hours at full load). The single biggest barrier to this is organizational and business complexity: the operator needs application owners to agree to such procedures and, most likely, a modification of SLAs. Technologies supporting this capability within larger colocation settings have not received much uptake, due to customer reluctance, but that may slowly change for select edge applications.

Workload prioritization is commonly used in telecom radio networks during grid power outages to maximize network availability and/or service revenues. For example, the operator of a distribution hub site may turn off some of its own radio capacity, while keeping up communication with downstream sites.

Photovoltaic solar power

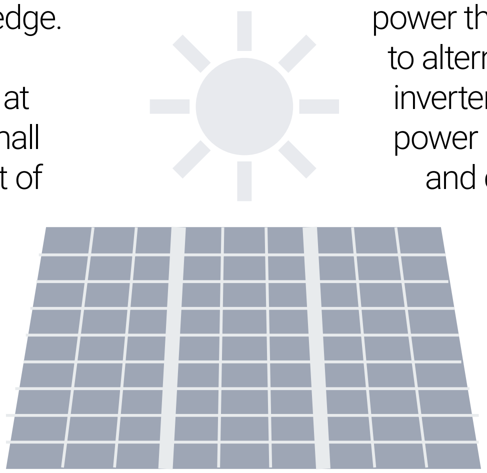
Photovoltaic (PV) solar has become the lowest-cost form of energy in most major countries, according to the International Energy Agency's estimations in their **World Energy Outlook 2020**. Although there are obvious limitations to using intermittent power as the primary energy source for data centers (discussed below), solar can be used to provide some (and sometimes all) of the energy needed — especially at the edge.

The same types of PV panels used at large solar farms can be used at small edge sites — with the same amount of solar power generated per panel. An edge data center network can thus be seen as an opportunity to deploy a distributed, efficient solar farm where power is produced at the points of consumption. Implementing tangible measures to reduce carbon emissions will be important in the years ahead because edge data center market growth comes with a risk of increasing, rather than decreasing, the environmental footprint of the data center sector.

In Uptime Institute's 2021 edge survey, almost half (48%) of the participating owners, operators and suppliers expect on-site solar panels to be one of the three most common edge data center power sources in two to three years. Respondents expect on-site solar panels to be roughly as common as batteries (see **Figure 5**).

In locations with a high number of sunshine hours, solar power is becoming viable as the primary power source for small edge sites. Australian colocation provider Edge Centres goes a step beyond: it uses on-site solar power and large batteries to power edge data centers with IT loads well over 100 kW.

Designing a pure solar-battery solution to independently handle many consecutive days with low solar radiation can, for many sites, involve very large battery arrays and/or solar power installations that become too large or expensive. Wherever connecting to the grid is not possible or practical, engine generators very often back up the site.



Solar panels generate direct current (DC) power that for regular servers is converted to alternating current (AC) power using inverters. Some edge data center power loads, like telecom equipment and certain Open Compute Project servers, use DC power and can make use of power from solar panels either directly or, when the DC voltage level needs to be adjusted, via a DC/DC converter. Combining solar panels with DC-powered equipment results in a system with less power loss than AC and at a lower cost (DC/DC converters, if needed, are cheaper and more efficient than inverters). Having fewer components in power distribution also means greater reliability.

However, most edge data centers will mainly use solar power to supplement the grid, due to high build costs and space constraints. Installing solar panels to produce a peak power of 50 kW at ideal conditions (midday, with sunny, clear weather) will require around 300 square meters (3,230 square feet) of sun-exposed site space. Installing solar panels to produce the yearly energy consumed by a 50 kW edge data center (operated 24/7) in a sunny region like California (US) requires around 1,600 square meters (17,220 square feet) of sun-exposed site space. Taking day-to-day solar power variations into account, as well as the added capacity needed to recharge batteries, will further increase the area of the solar farm. This is well beyond what is available at most edge data center sites.

A particularly strong business case can be made for solar panels in off-grid areas, where engine generators are commonly used together with batteries, as they reduce engine generator costs (prolonging maintenance intervals and replacement cycle, as well as reducing fuel consumption) and increase battery life (due to a reduction of battery charge-discharge cycles).

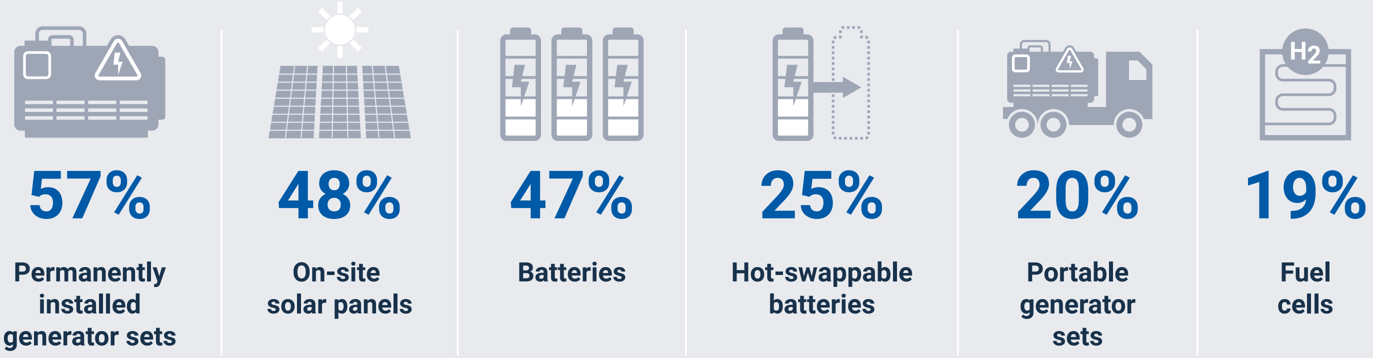
From a resiliency perspective, solar panels can, in theory, help to extend daytime availability of on-site power during a grid power outage (assuming it occurs during daylight hours, with good weather conditions). This is highly probabilistic, however, and cannot be relied on due to the time and weather dependency of solar power generation. In general, solar power production at the time of an outage should not be assumed.

Solar panels will need regular cleaning to maintain efficiency. The required cleaning interval can vary depending on the levels of dust, air contaminants, bird spillage, snowfall, etc. in the area. Remote monitoring systems measuring power production, combined with suitably placed cameras, are useful to determine the need for panel cleaning at unstaffed edge data centers.

FIGURE 5

Many expect edge data center sites will use solar power

In addition to utility electricity grids, what power sources do you think edge data centers will be most commonly using 2-3 years from now? (Choose no more than three)



UptimeInstitute® | INTELLIGENCE

UPTIME INSTITUTE GLOBAL SURVEY OF EDGE DATA CENTER OWNERS, OPERATORS, AND SUPPLIERS 2021 (n=358)

Remote monitoring and control

Remote operations will, for reasons of cost and staffing challenges, likely become the standard model for most edge sites. As is the case for all critical systems, scheduled and unscheduled maintenance and outage recovery will need on-site presence, but the baseline mode of operation and repair will be more similar to that of telecommunication equipment sites than to a typical data center. Data center infrastructure at the edge, in most cases, will need to operate primarily without local staff and will require software-mediated control.

> “Data center infrastructure at the edge, in most cases, will need to operate primarily without local staff and will require software-mediated control.”

Minimizing site visits

Due to a large number of geographically dispersed locations, minimizing the number and duration of site visits for edge data centers will be essential. Edge operators are seeking ways to use condition-based maintenance tools and regimes; comprehensive monitoring; automation; data center infrastructure management software; and/or artificial intelligence to predict performance and failures. We expect edge data center suppliers (particularly vendors offering prefabricated modular edge data centers) and colocation operators to integrate more of such functionality into their edge products.

Minor alarms at unstaffed edge data centers will mostly be handled at the next scheduled maintenance visit. A major alarm, like an outage or one redundant system failing, should automatically generate a work order to the maintenance contractor, which may be a local generalist firm (i.e., not a data center specialist). Basic maintenance activities can be handled by local generalist staff, if available, or by IT staff supported by remote specialist staff at network operation centers as needed. This means specialist mechanical and electrical staff will increasingly need to be trained experts in real-time monitoring and management software.

Remote monitoring and analysis required
Remote operations and distributed resiliency rely on remote monitoring (and analysis) to determine the operational status and utilization of edge data centers. (Workloads cannot be moved to edge data centers without first establishing their operational status, the performance level of IT systems, and available capacity.) Some edge colocation providers offer an API as a convenient way to access operational status and utilization data. The Synse API, developed and published by Vapor IO, is one example.

If remote monitoring is lost, an edge data center could revert to default settings and continue to operate at an elevated risk level. To avoid this situation, most edge data center operators take measures to design remote monitoring systems for increased resiliency, including:

- Physical separation of remote monitoring systems from other equipment in an edge data center.
- Use of one or more dedicated on-site power sources (feasible due to low power draw).
- Use of redundant monitoring systems and sensors in each edge data center.
- Use of one or more out-of-band communication links that are separated from the communication links handling regular IT traffic. These may need to be separated from the internet for security reasons; some operators use wireless 3G/4G LTE networks at remote locations as a fallback to check edge data center operational status but not to synchronize all data.

Organizations using colocation providers that operate unstaffed edge data centers will need to evaluate the provider's ability to guarantee resilient remote operations. In areas prone to flooding or earthquakes, additional sensors can be added to trigger failure-mode procedures and optionally shut down an edge data center in case of such adverse natural events.

Summary

Organizations deploying edge data centers can benefit from the combined use of site-level resiliency and distributed resiliency. While the former is used for improving site resiliency against equipment failures, the latter can provide services resiliency against loss of a site, enable lower infrastructure costs, and provide increased technical agility by flexible placement and shifting of IT workloads.

Organizations should expect teething problems with distributed edge resiliency architectures, as they involve prolonged test and validation stages because of the increased software and network complexity. IT workload prioritization techniques can further be used to extend on-site power run time for critical workloads during power grid outages, particularly at sites without engine generators.

A range of power sources will be used at edge centers, with generators continuing to play a role at large sites or at remote locations where the grid may be unreliable. A combination of power generation and storage systems, together with software for managing workloads and power use, may prove the most effective approach.

Resilient remote monitoring systems that include independent power supplies and out-of-band networks are critical for distributed resiliency and unstaffed operations. The optimal mix of resiliency approaches will depend on IT workloads and business priorities.

Appendix A: Key companies

The following companies — all notable edge data center suppliers, owners, operators, and innovators — are among the players currently defining and implementing edge data center resiliency:

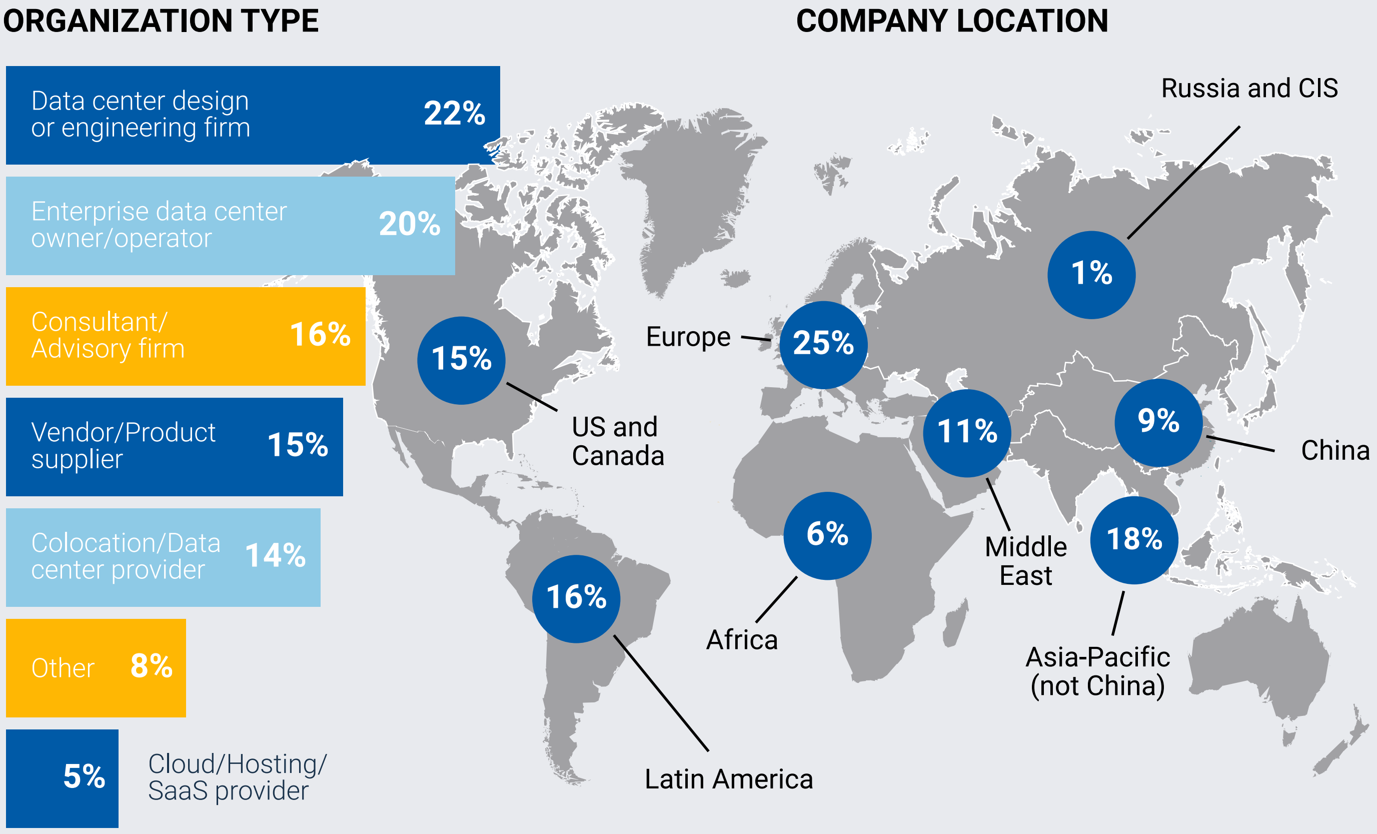
- Amazon Web Services
- Cannon Technologies
- CloudFrame
- Compass Datacenters
- Delta Electronics
- DCI Data Centers
- Edge Centres
- EdgeMicro
- Google
- Huawei
- Leading Edge Data Centres
- Microsoft
- Schneider Electric
- Silent-Aire (Johnson Controls)
- Vapor IO
- Vertiv
- Zella DC

Appendix B: Methodology

This report draws on roughly two dozen in-depth interviews with selected companies that either use, supply or provide supporting infrastructure for edge data centers, including cloud providers, colocation providers, telecoms and edge data center suppliers. The report also cites findings from Uptime Institute's Global Survey of Edge Data Center Owners, Operators and Suppliers 2021, which was conducted during the first quarter of 2021. More than 430 decision makers participated in the survey. **Figure B1** provides respondent demographics.

FIGURE B1

Survey demographics



Uptime Institute | INTELLIGENCE

UPTIME INSTITUTE GLOBAL SURVEY OF EDGE DATA CENTER OWNERS, OPERATORS AND SUPPLIERS 2021 (n=434)

ABOUT THE AUTHORS

Tomas Rahkonen
Research Director of
Distributed Data Centers
Uptime Institute

Dr. Rahkonen has spent over 25 years in global positions in the telecommunications, mobile communications and data center sectors. He most recently served over 10 years as CTO of Flexenclosure, where he managed the design and delivery of prefab data centers to four continents.
Contact: trahkonen@uptimeinstitute.com

Andy Lawrence
Executive Director
of Research
Uptime Institute

A founding member of Uptime Institute Intelligence, Mr. Lawrence has spent three decades analyzing developments in IT, emerging technologies, data centers and infrastructure; and advising companies on their technical and business strategies.
Contact: alawrence@uptimeinstitute.com

ABOUT UPTIME INSTITUTE

Uptime Institute and the company's Tier Standard is the globally recognized digital infrastructure authority known for the creation and administration of the world's most adopted standards for data center performance and resilience. For over 25 years, Uptime Institute has been providing customers with the assurance that their digital infrastructure can perform at a level consistent with their business needs, across a wide array of operating conditions. With its data center Tier Standard & Certifications, Management & Operations reviews, Digital Infrastructure Resiliency Assessments and other services, along with our accredited educational curriculum for data center professionals, Uptime Institute helps organizations optimize critical IT assets while managing costs, resources, and efficiency. Today, thousands of companies rely on Uptime Institute to enable their digital-centric business success.

Uptime Institute - The Global Digital Infrastructure Authority®, a division of The 451 Group, has office locations throughout the world, including North and South America, Europe, UAE, Russia, and Asia. 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