Uptime Institute® INTELLIGENCE

PLANNING & STRATEGY

UI Intelligence report 42

Five data center trends for 2021

Sector dynamics, market developments, innovations, challenges and opportunities

Authors

Rhonda Ascierto, Vice President of Research, Uptime Institute

Andy Lawrence, Executive Director of Research, Uptime Institute

Uptime Institute's examination of some of the top trends in data centers in 2021 reveals a vibrant sector that is growing, especially around the edge, and increasingly embracing new innovations. It is also entering a phase of re-assessment; of infrastructure and service accountability, in terms of resiliency; and of material outcomes toward environmental sustainability.

30-45 minutes to read

UII-42 v1.1 published December 1, 2020 last updated December 22, 2020

This Uptime Institute Intelligence report includes:

Introduction

Heading into 2021, during a macroeconomic downturn, the critical digital infrastructure sector continues to expand and to attract enviable levels of new investment. The ongoing build-out of new data centers and networks is largely being driven by cloud, hosted, and other as-a-service workloads, as more enterprises seek to outsource more of their IT. However, for many managers, the COVID-19 pandemic has forced a reassessment — of strategies and, in particular, of risk.

The global economy's dependence on IT is growing, which means more digital services and the digital infrastructure supporting them are becoming more critical. This is catching the attention of an increasing number of customers, governments and watchdogs. Resiliency concerns and the impact of climate change will be among the hazards that IT and data center managers must successfully navigate.

The coming year (and beyond) also holds new opportunities. Edge computing, artificial intelligence (AI) and new innovations in hardware and software technologies promise greater efficiencies and agility.

Uptime Institute Intelligence's examination of five major digital infrastructure trends in 2021 shows a sector that is confidently expanding, in new and creative ways. It also points to an emerging next phase of sector maturity that will result in more complexity and that will require more responsibility.

FIVE DATA CENTER TRENDS FOR 2021

1. **Accountability — the "new" imperative**

   Enterprises want more cloud and greater agility, but they can't outsource responsibility — for incidents, outages, security breaches or even, in the years ahead, carbon emissions. In 2021, hybrid IT, with workloads running in both on- and off-premises data centers, will continue to dominate, but investments will increasingly be constrained and shaped by the need for more transparency, oversight and accountability. More will be spent on cloud and other services, as well as in on-premises data centers.

2. **Smarter, darker data centers**

   Following a scramble to effectively staff data centers during a pandemic, many wary managers are beginning to see remote monitoring and automation systems in a more positive light, including those driven by AI. An adoption cycle that has been slow and cautious will accelerate. But it will take more than just investment in software and services before the technology reduces staffing requirements.

3. **Edge — the next frontier**

   Significant new demand for edge computing, fueled by technologies such as 5G, the internet of things (IoT) and AI, is likely to build slowly but the infrastructure preparation is underway. Expect new alliances and investments across enterprise, mobile and wireline networks, and for a wide range of edge data centers, small and large. Smart and automated software-defined networks and interconnections will become as important as the physical infrastructure.

4. **Sustainability: More challenging, more transparent**

   For years, operators could claim environmental advances based on small, incremental and relatively inexpensive steps — or by adopting new technologies that would pay for themselves anyway. But the time of easy wins and greenwashing is ending: Regulators, watchdogs, customers and others will increasingly expect operators of digital infrastructure to provide hard and detailed evidence of carbon reductions, water savings and significant power savings — all while maintaining, if not improving, resiliency.

5. **A surge of innovation**

   Data center operators (and enterprise IT) are mostly cautious, if not late, adopters of new technologies. Few beyond hyperscale operators can claim to have gained a competitive advantage through technology. However, several new technologies are maturing at the same time, promising advances in the performance and manageability of data centers and IT. Storage-class memory, silicon photonics, ARM servers and software-defined power are ready for greater adoption.

TREND ONE

Accountability — the "new" imperative

Enterprises want more cloud, more agility, and to use more outsourcing. But to move forward, they will need more transparency, oversight and accountability.

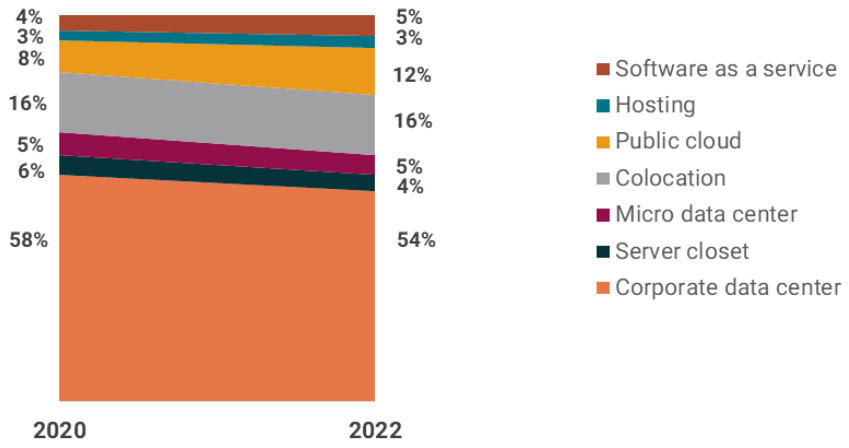
Outsourcing the requirement to own and operate data center capacity is the cornerstone of many digital transformation strategies, with almost every large enterprise spreading their workloads across their own data centers, colocation sites and public cloud. But ask any regulator, any chief executive, any customer: You can't outsource responsibility — for incidents, outages, security breaches or even, in the years ahead, carbon emissions.

Chief information officers, chief technology officers and other operational heads knew this three or four decades ago (and many have learned the hard way since). That is why data centers became physical and logical fortresses, and why almost every component and electrical circuit has some level of redundancy.

In 2021, senior executives will grapple with a new iteration of the accountability imperative. Even the most cautious enterprises now want to make more use of the public cloud, while the use of private clouds is enabling greater choices of third-party venue and IT architecture. But this creates a problem: cloud service operators, software-as-a-service (SaaS) providers and even some colos are rarely fully accountable or transparent about their shortcomings — and they certainly do not expect to be held financially accountable for consequences of failures. Investors, regulators, customers and partners, meanwhile, want more oversight, more transparency and, where possible, more accountability.

This is forcing many organizations to take a hard look at which workloads can be safely moved to the cloud and which cannot. For some, such as the European financial services sector, regulators will require an assessment of the criticality of workloads — a trend that is likely to spread and grow to other sectors over time. The most critical applications and services will either have to stay in-house, or enterprise executives will need to satisfy themselves and their regulators that these services are run well by a third-party provider, and that they have full visibility into the operational practices and technical infrastructure of their provider.

The data suggests this is a critical period in the development of IT governance. The shift of enterprise IT workloads from on-premises data center to cloud and hosted services is well underway, as shown in Figure 1. (Note that Figure 1 shows only enterprise workloads, which are run to support a business whose primary function is not IT; it does not represent the totality of IT. Internet workloads, mobile phone traffic, video running over content distribution networks [CDNs], etc. are not included.) But there is a long way to go, and some of the issues around transparency and accountability have arisen only recently as more critical and sensitive data and functionality is considered for migration to the cloud.



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)

UptimeInstitute | INTELLIGENCE

Figure 1. **Fewer workloads will run in enterprise (on-premises) data centers in the future**

The first tranche of workloads moving to third parties often did not include the most critical or sensitive services. For many organizations, a public cloud is (or was initially) the venue of choice for specific types of workloads, such as application test and development; big-data processing, such as AI; and new applications that are cloud-native. (The list in Figure 2 includes common drivers of IT venue decisions.) But as more IT departments become familiar with the tool sets from cloud providers, such as for application development and deployment orchestration, more types of workloads have moved into public clouds only recently, with more critical applications to follow (or perhaps not). As discussed in **Outages drive authorities and businesses to act**, high-profile, expensive public cloud outages, increased regulatory pressures and an increasingly uncertain macroeconomic outlook will force many enterprises to assess — or reassess — where workloads should actually be running (a process that has been called "The Big Sort").

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 too expensive 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

Figure 2. **Common drivers for IT venue decisions**

Uptime Institute believes that many mission-critical workloads are likely to remain in on-premises or colo data centers — at least for many years to come: More than 70% of IT and critical infrastructure operators we surveyed in 2020 do not put any critical workloads in a public cloud, with over a quarter of this group (21% of the total sample) saying the reason is a lack of visibility/accountability about resiliency. And over a third of those who do place critical applications in a public cloud also say they do not have enough visibility (see Figure 3). Clearly, providers' assurances of availability and of adherence to best practices are not enough for mission-critical workloads. (These results were almost identical when we asked the same question in our 2019 annual survey.)

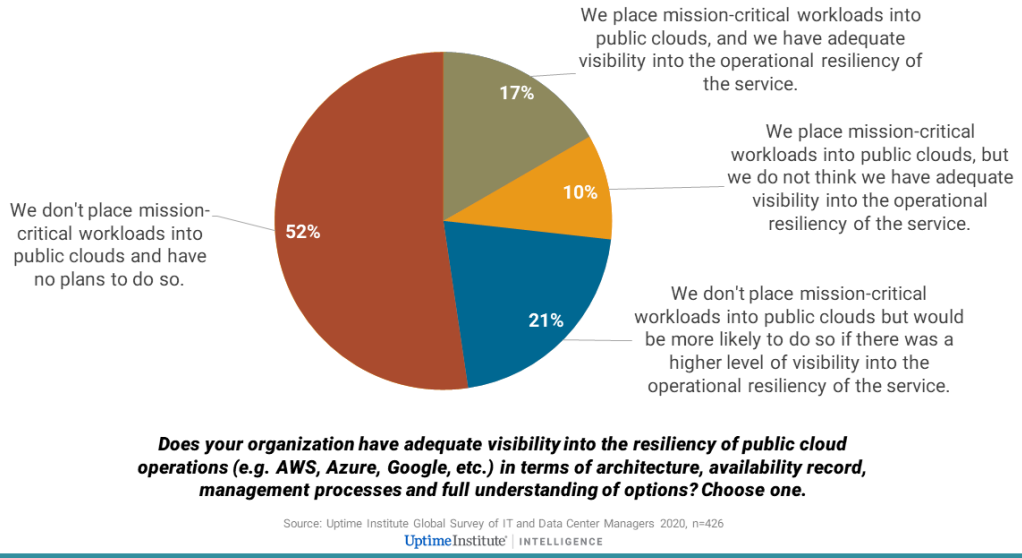


Figure 3. **More would use public cloud if providers gave better visibility into options and operations**

The issues of transparency, reporting and governance are likely to ripple through the cloud, SaaS and hosting industries, as customers seek assurances of excellence in operations — especially when financial penalties for failures by third parties are extremely light. While even the largest cloud and internet application providers operate mostly concurrently maintainable facilities, experience has shown that unaudited ("mark your own homework") assurances frequently lead to poor outcomes.

Creeping criticality

There is an added complication. While the definitions and requirements of criticality in IT are dictated by business requirements, they are not fixed in time. Demand patterns and growing IT dependency mean many workloads/services have become more critical — but the infrastructure and processes supporting them may not have been updated ("creeping criticality"). This is a particular concern for workloads subject to regulatory compliance ("compliance drift").

COVID-19 may have already caused a reassessment of the criticality or risk profile of IT; extreme weather may provide another. When Uptime Institute recently asked over 250 on-premises and colo data center managers how the pandemic would change their operations, two-thirds said they expect to increase the resiliency of their core data center(s) in the years ahead. Many said they expected their costs to increase as a result. One large public cloud company recently asked their leased data center providers to upgrade their facilities to N+1 redundancy, if they were not already.

But even before the pandemic, there was a trend toward higher levels of redundancy for on-premises data centers (see Figure 4).

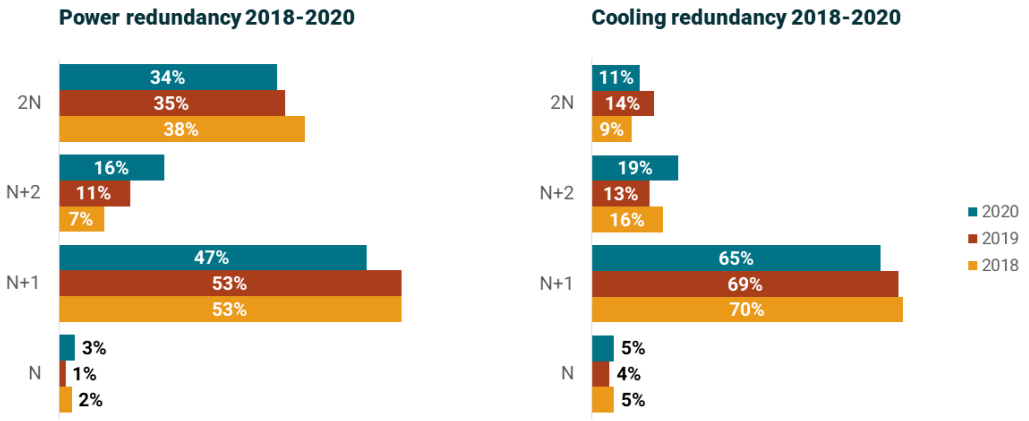


Figure 4. **Suppliers report higher levels of redundancy**

There is also an increase in the use of active-active availability zones, especially as more workloads are designed using cloud or microservices architectures. Workloads are more portable, and instances are more easily copied than in the past. (For a more detailed explanation, see the section "Use of availability zones is spreading" in **Uptime Institute global data center survey 2020**). But we see no signs that this is diminishing the need for site-level resiliency.

Colos are well-positioned to provide both site-level resiliency (which is transparent and auditable) and outsourced IT services, such as hosted private clouds. We expect more colos will offer a wider range of IT services, in addition to interconnections, to meet the risk (and visibility) requirements of more mission-critical workloads. The industry, it seems, has concluded that more resiliency at every level is the least risky approach — even if it means some extra expense and duplication of effort.

Uptime Institute expects that the number of enterprise (privately owned/on-premises) data centers will continue to dwindle but that enterprise investment in site-level resiliency will increase (as will investment in data-driven operations, see **Smarter, darker data centers**). Data centers that remain in enterprise ownership will likely receive more investment and continue to be run to the highest standards.

The world is moving to hybrid IT. But the choice of which workloads and services will run where, and which third-party partners will be used, will increasingly be constrained and shaped by the need for more transparency, oversight and accountability.

TREND TWO

Smarter, darker data centers

We expect a wave of new investment in remote monitoring and automation. It will take more than just new software and services, however, before the technology reduces staffing requirements.

Following a scramble to effectively staff data centers during a pandemic, many wary managers are beginning to see remote monitoring and automation systems in a more positive light, including those driven by AI.

The reasoning is clear. Against a background expectation that another pandemic will occur, and with many operators struggling to find candidates for open jobs, a goal is often to reduce staffing levels — or at least, to increase the ability to operate for sustained periods with fewer staff. Remote technologies promise to do just this, but significant investments will often be required, and the largest barrier is likely to be the organizational change and the monetary commitment such change requires (see Figure 5).

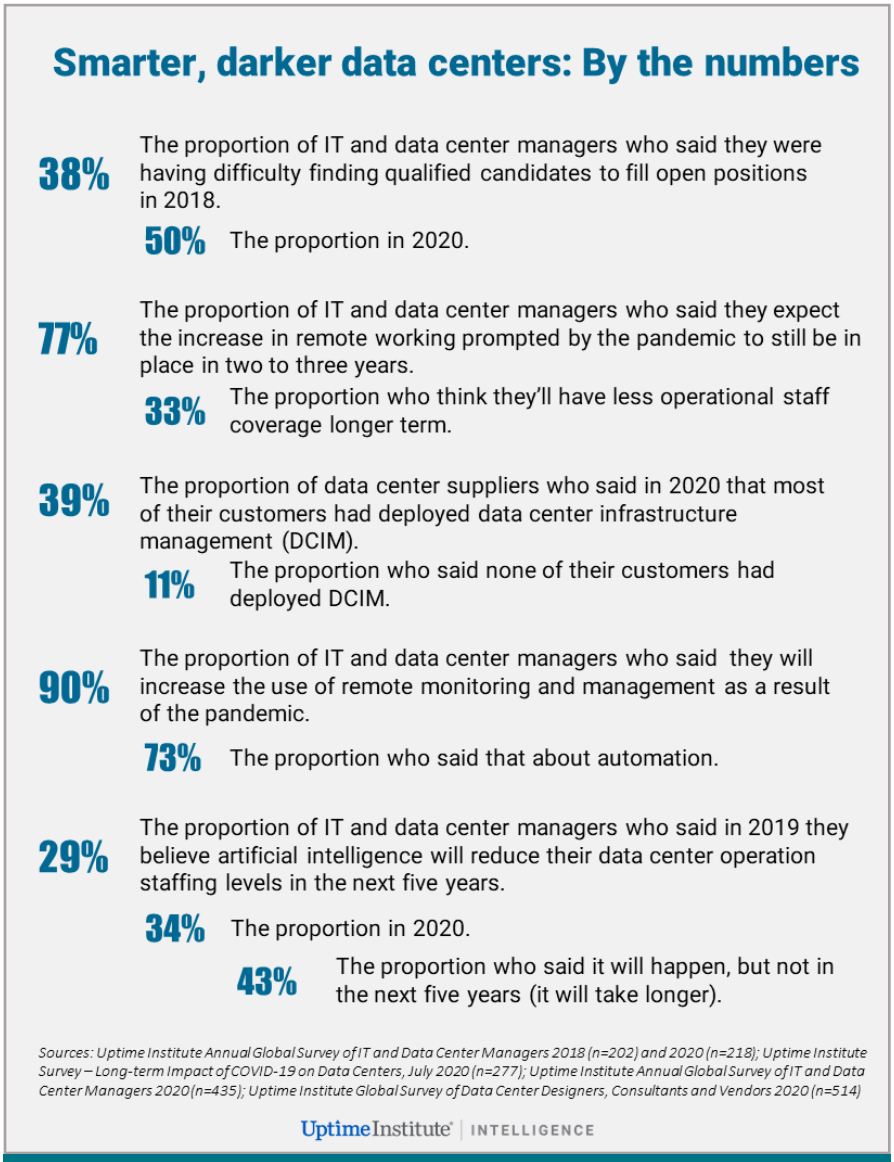


Figure 5. **Smarter, darker data centers: By the numbers**

Accurate, normalized data from numerous sources is a bedrock of remote management, automation, and AI-driven data center approaches. Most data centers already have some form of data center infrastructure management (DCIM) software, whether commercial or home-grown. DCIM analyzes data from both facility and IT equipment, and often also from building management and other control systems.

DCIM collects and normalizes data (to smooth variances in sample rates, units of measurement, etc.), then analyzes it in a way that enables users to reduce risk, increase efficiency and improve forecasting. Yet most DCIM deployments today are limited. It's common for just one of the two main DCIM components to be deployed: either DCIM monitoring (power and environmentals) or DCIM asset management, including for asset changes and configurations. Each of these components is valuable but represent activities that cannot be carried out in isolation if data is to be relied on for critical decision-making — particularly if those decisions are made remotely. For example, a power monitoring alert will need to be mapped to the assets that may be affected, along with information about those assets, such as redundancy, maintenance history, operating ranges, etc.

In a situation where service availability is threatened, data about IT workloads may also need to be integrated, such as from virtual machine or IT service management software. For example, if workloads running on at-risk assets are critical, they may need to be moved to another location with an adequate risk profile, including for compliance and service levels.

All of these functions and (bi-directional) data integrations are available from commercial DCIM software packages, yet most data centers have not taken full advantage of their capabilities. Progress is typically slowed by the organizational change required to benefit from the software, such as new procedures to input data and to process and implement recommendations, as well as the requirement for close collaboration between facilities teams and IT departments (when IT and/or workload data is integrated).

In 2021, we expect more organizations will look closely at their available data/software capabilities and begin to address both the technical and the organizational requirements needed. A small but growing number of data centers, particularly smaller facilities, are likely to bypass DCIM software or to augment existing deployments by using newer cloud-based services, known as data center management as a service (DMaaS). These services aggregate and analyze telemetry and/or DCIM data (if available) over a wide area network (WAN), typically the internet. Customers receive alerts, analysis and recommendations that are tailored specifically to their individual data centers, based on the insights of the larger collective pool (typically several large data lakes).

DCIM software and cloud-delivered DMaaS are increasingly using AI and other types of big-data methods to spot anomalies and patterns, and to predict and forecast. Work in this area is developing; to date, it has focused mostly on efficiency and predictive maintenance

of uninterruptible power supplies (UPSs). We expect suppliers will increasingly focus on features to reduce risk and, over time, to help lower on-site staffing requirements.

However, managers who rely solely on a cloud-delivered service may face limitations: few, if any, will trust a WAN when it comes to the automation of critical systems or for critical alerts; installing some data and software components locally, on-premises, is advisable.

Whether installed locally or delivered as a service, most AI-driven data center alerts and automation will be an iterative, human-driven process before the system is trusted — and before lower staffing levels are possible without increasing risk. (See our report **Very smart data centers: How artificial intelligence will power operational decisions**.)

It is likely just a matter of time before more feature-rich, data-driven approaches are adopted and trusted, and data centers become smarter and require less operational staff. Over time, we also expect the use of novel remote-enabling technologies, such as:

- Virtual reality, including for training and crisis simulation.
- Augmented reality, including for non-skilled staff working on-site.
- AI-enabled video surveillance, ranging from physical security to detecting dislodged cables.
- Acoustic technology to monitor the status of equipment, such as transformers, generators, etc.
- Robotics, including with vision and gripping capabilities.

Meanwhile, data center managers are likely to invest more in DCIM and DMaaS and will develop a better understanding of what the technology can and cannot offer, and the level of investment required for value realization.

The move toward smarter, darker data centers is well underway but the vision will continue to lag the technology and operational practices in most data centers for some time yet.

TREND THREE

Edge — the next frontier

The edge is still nascent, but it is a time to make partnerships, deliver technology, and build infrastructure.

One of the most widely anticipated trends in IT and infrastructure is significant new demand for edge computing, fueled by technologies such as 5G, IoT and AI. To date, net new demand for edge computing — processing, storing and integrating data close to where it is generated — has built slowly (see Figure 6). As a result, some suppliers of micro data center and edge technologies have had to lower their investors' expectations.

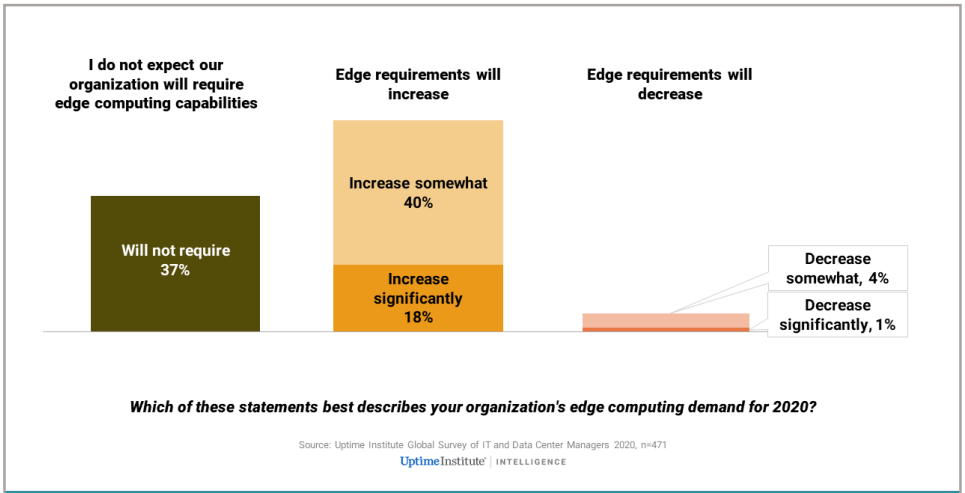


Figure 6. **Most expect edge demand to increase**

This slow build-out, however, does not mean that it will not happen. Demand for decentralized IT will certainly grow. There will be more workloads that need low latency, such as healthcare tech, high performance computing (notably more AI), critical IoT, and virtual and augmented reality, as well as more traffic from latency-sensitive internet companies (as Amazon famously said 10 years ago, every 100 milliseconds of latency costs them one percent in sales). There will also be more data generated by users and "things" at the edge, which will be too expensive to transport across long distances to large, centralized data centers (the "core").

For all these reasons, new edge data center and connectivity capacity will be needed, and we expect a wave of new partnerships and deals in 2021. Enterprises will connect to clouds via as-a-service (on-demand, software-driven) interconnections at the edge, and the internet will extend its reach with new exchange points. Just as the internet is a network of tens of thousands of individual networks connected together, the edge will require not just new capacity but also a new ecosystem of suppliers working together. The year 2021 will likely see intense activity — but the long-expected surge in demand may have to wait.

The edge build-out will be uneven, in part because the edge is not a monolith. Different edge workloads need different levels of latency,

bandwidth and resiliency, as shown in the data center schema in Figure 7. Requirements for data transit and exchanges will also vary. Edge infrastructure service providers will need to rely on many partners, including specialist vendors that will serve different customer requirements. Enterprise customers will become increasingly dependent on third-party connections to different services.

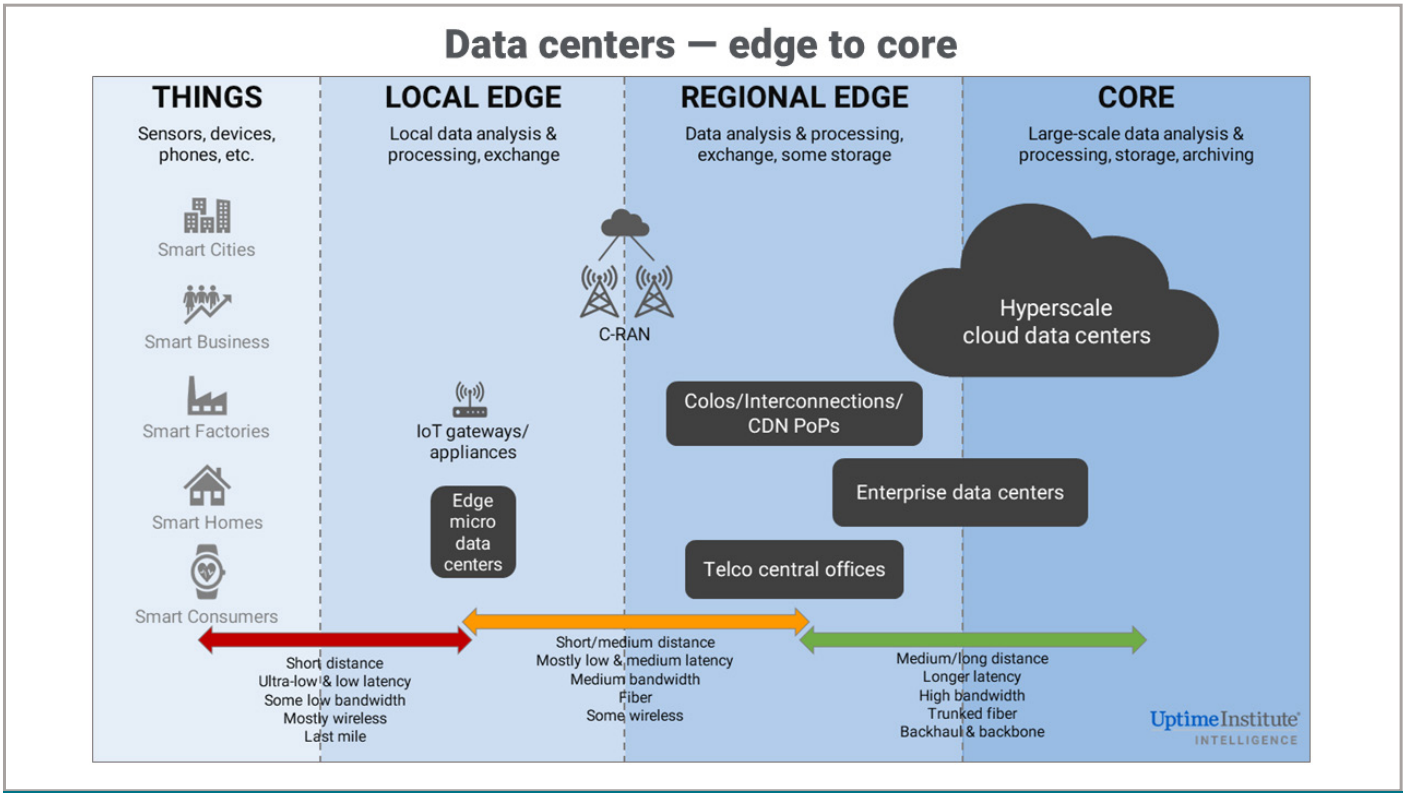


Figure 7. **Uptime Institute's edge-to-core data center schema**

So far, much attention has been focused on the local edge, where connectivity and IT capacity are sited within a kilometer or so from devices and users. In urban areas, where 5G is (generally) expected to flourish, and in places where a lot of IoT data is generated, such as factories and retail stores, we are slowly seeing more micro data centers being deployed. These small facilities can act either as private connections or internet exchange points (or both), handing off wireless data to a fiber connection and creating new "middle-mile" connections.

We expect that edge micro data centers will be installed both privately and as shared infrastructure, including for cloud providers, telcos and other edge platform providers, to reduce latency and keep transit costs in check. To get closer to users and "things," fiber providers will also partner with more wireless operators.

In 2021, most of the action is likely to be one step further back from the edge, in regional locations where telcos, cloud providers and enterprises are creating — or consuming — new

interconnections in carrier-neutral data centers such as colo and wholesale facilities. All major cloud providers are increasingly creating points of presence (PoPs) in more colos, creating software-defined WANs of public (internet) and private (enterprise) connections. Colo customers are then able to connect to various destinations, depending on their business needs, via software, hardware and networks that colos are increasingly providing. These interconnections are making large leased facilities a preferred venue for other suppliers to run edge infrastructure-as-a-service offerings, including for IoT workloads. For enterprises and suppliers alike, switching will become as important as power and space.

We expect more leased data centers will be built (and bought) in cities and suburbs in 2021 and beyond. Large and small colos alike will place more PoPs in third-party facilities. And more colos will provide more software-driven interconnection platforms, either via internal development, partnerships or acquisitions.

At the same time, CDNs that already have large edge footprints will further exploit their strong position by offering more edge services on their networks directly to enterprises. We're also seeing more colos selling "value-add" IT and infrastructure-as-a-service products — and we expect they will extend further up the IT stack with more compute and storage capabilities.

The edge build-out will clearly lead to increased operational complexity, whereby suppliers will have to manage hundreds of application program interfaces and multiple service level agreements. For these reasons, the edge will need to become increasingly software-defined and driven by AI. We expect investment and partnerships across all these areas.

How exactly it will play out remains unclear; it is simply too early. Already we have seen major telco and data center providers pivot their edge strategies, including moving from partnerships to acquisitions.

One segment we are watching particularly closely is the big internet and cloud companies. Having built significant backbone infrastructure, they have made little or only modest investments to date at the edge. With their huge workloads and deep pockets, their appetite for direct ownership of edge infrastructure is not yet known but could significantly shape the ecosystem around them.

New alliances and investments will be made across enterprise, mobile and wireline networks and for various types of data center capacity, small and large. As the edge develops, smart and automated software-defined networks and interconnections will become as important as the physical infrastructure.

TREND FOUR

Sustainability: More challenging, more transparent

The time of easy wins and greenwashing is ending. Sustainability is not a choice, and there will be few places for data center operators to hide.

Through 2021 and beyond, the world will begin to recover from its acute crisis — COVID-19 — and will turn its attention to other matters. Few if any of these issues will be as important as climate change, a chronic condition that will become more pressing and acute as each year passes.

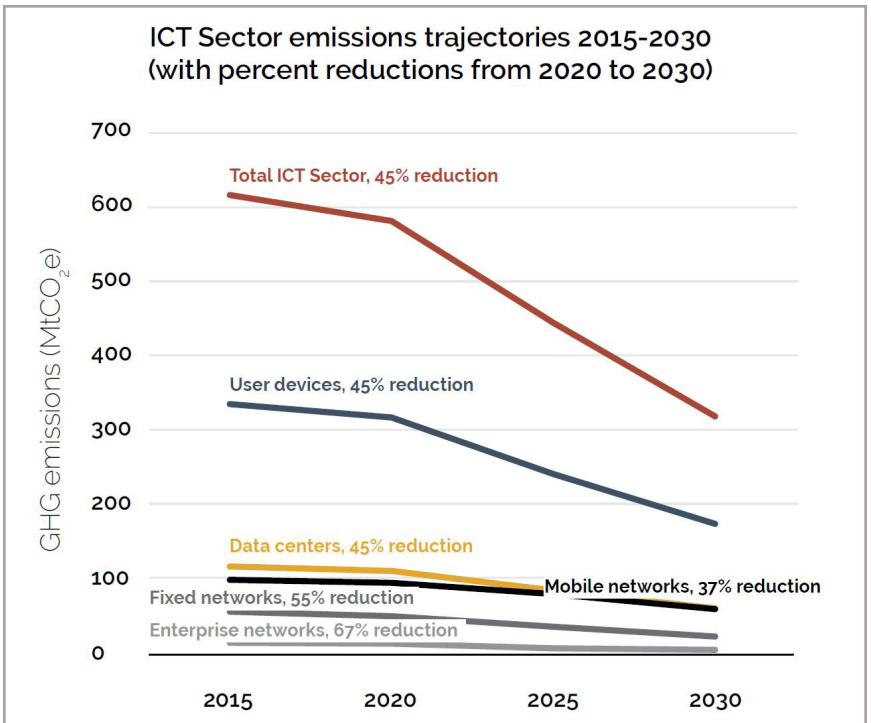
In the critical digital infrastructure sector, as in all businesses, issues arising directly or indirectly from climate change will play a significant role in strategic decision-making and technical operations in the years ahead. And this is regardless of the attitude or beliefs of senior executives; stakeholders, governments, customers, lobbyists and watchdogs all want and expect to see more action. The year 2021 will be critical, with governments expected to act with greater focus and unity as the new US government rejoins the global effort.

We can group the growing impact of climate change into four areas:

- **Extreme weather/climate impact** - As discussed in our report The gathering storm: Climate change and data center resiliency, extreme weather and climate change present an array of direct and indirect threats to data centers. For example, extreme heatwaves — which will challenge many data center cooling systems — are projected to occur once every three or four years, not once in 20.
- **Legislation and scrutiny** - Nearly 2,000 pieces of climate-related legislation have been passed globally to date (covering all areas). Many more, along with more standards and customer mandates, can be expected in the next several years.
- **Litigation and customer losses** - Many big companies are demanding rigorous standards through their supply chains — or their contracts will be terminated. Meanwhile, climate activists, often well-resourced, are filing lawsuits against technology companies and digital infrastructure operators to cover everything from battery choices to water consumption.
- **The need for new technologies** - Management will be under pressure to invest more, partly to protect against weather events, and partly to migrate to cleaner technologies such as software-defined power or direct liquid cooling.

In the IT sector generally — including in data centers — it has not all been bad news to date. Led by the biggest cloud and colo companies, and judged by several metrics, the data center sector has made good progress in curtailing carbon emissions and wasteful energy use. According to the Carbon Trust, a London-based body focused on reducing carbon emissions, the IT sector is on course to meet its

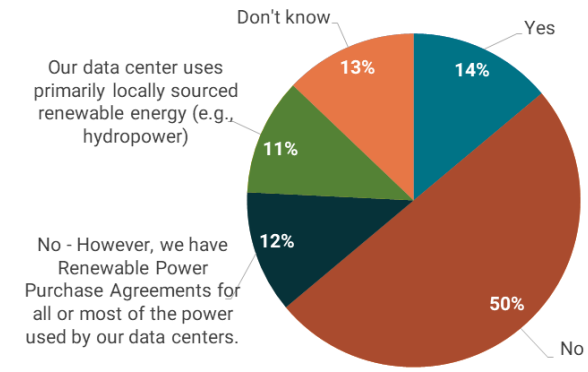
science-based target for 2030 — a target that will help keep the world to 1.5 degrees Celsius (34.7 degrees Fahrenheit) warming (but still enough warming to create huge global problems). As shown in Figure 8, its data shows IT sector carbon emissions from 2020 to 2030 are on a trajectory to fall significantly in five key areas - data centers, user devices, mobile networks, and fixed and enterprise networks. Overall, the IT sector needs to cut carbon emissions by 50% from 2020 to 2030.



Source: Guidance for ICT companies setting science based targets (Mobile networks operators, fixed networks operators, and data centres operators) (2020) by Science based targets: Driving ambitious corporate climate action. https://www.itu.int/en/mediacentre/Documents/Documents/GSMA_IP_SBT-report_WEB-SINGLE.pdf

Figure 8. **Summary of ICT sector and sub-sector trajectories including embodied emissions and operation**

Data centers are just a part of this, accounting for more carbon emissions than mobile, fixed or enterprise networks, but significantly less than all the billions of user devices. Data center energy efficiency has been greatly helped by facility efficiencies, such as economizer cooling, improvements in server energy use, and greater utilization through virtualization and other IT/software improvements. Use of renewables has also helped: According to Uptime Institute data (our 2020 Climate Change Survey) over a third of operators now largely power their data centers using renewable energy sources or offset their carbon use (see Figure 9). Increasing availability of renewable power in the grid will help to further reduce emissions.



Does your organization buy carbon offsets (including RECS) to balance out carbon emissions associated with your data center?*

**All figures rounded*

Source: Uptime Institute Climate Change Survey 2020 (n=194)

Uptime Institute | INTELLIGENCE

Figure 9. **Many take steps to minimize or offset their carbon emissions**

But there are some caveats to the data center sector's fairly good performance. First, the reduction in carbon emissions achieved to date is contested by many who think the impact of overall industry growth on energy use and carbon emissions has been understated (i.e., energy use/carbon emissions are actually quite a lot higher than widely accepted models suggest — a debatable issue that Uptime Institute continues to review). Second, at an individual company or data center level, it may become harder to achieve carbon emissions reductions in the next decade than it has been in the past decade — just as the level of scrutiny and oversight, and the penalty for not doing enough, ratchets up. Why? There are several reasons, including the following:

- Many of the facilities-level improvements in energy use at data centers have been achieved already — indeed, average industry power usage effectiveness values show only marginal improvements over the last five years. Some of these efficiencies may even go into reverse if other priorities, such as water use or resiliency, take precedence (economizers may have to be supplemented with mechanical chillers to reduce water use, for example).
- Improvements in IT energy efficiency have also slowed — partly due to the slowing or even ending of Moore's Law (i.e., IT performance doubling every two years) — and because the easiest gains in IT utilization have already been achieved.
- Some of the improvements in carbon emissions over the next decade require looking beyond immediate on-site emissions, or those from energy supplies. Increasingly, operators of critical digital infrastructure — very often under external pressure and executive mandate — must start to record the embedded carbon emissions (known as Scope 3 emissions) in the products and services they use. This requires skills, tools and considerable administrative effort.

The biggest operators of digital infrastructure — among them Amazon, Digital Realty, Equinix, Facebook, Google and Microsoft — have made ambitious and specific commitments to achieve carbon neutrality in line with science-based targets within the next two decades. That means, first, they are setting standards that will be difficult for many others to match, giving them a competitive advantage; and second, these companies will put pressure on their supply chains — including data center partners — to minimize emissions.

For those organizations that lack the will to reduce their digital infrastructure carbon footprints, or that are lagging, there will be nowhere to hide – and there will be fewer opportunities to avoid close and meaningful scrutiny. The age of greenwashing is coming to an end.

TREND FIVE

A surge of innovation

Several new technologies are maturing at the same time, promising some significant advances in the performance and manageability of data centers and IT.

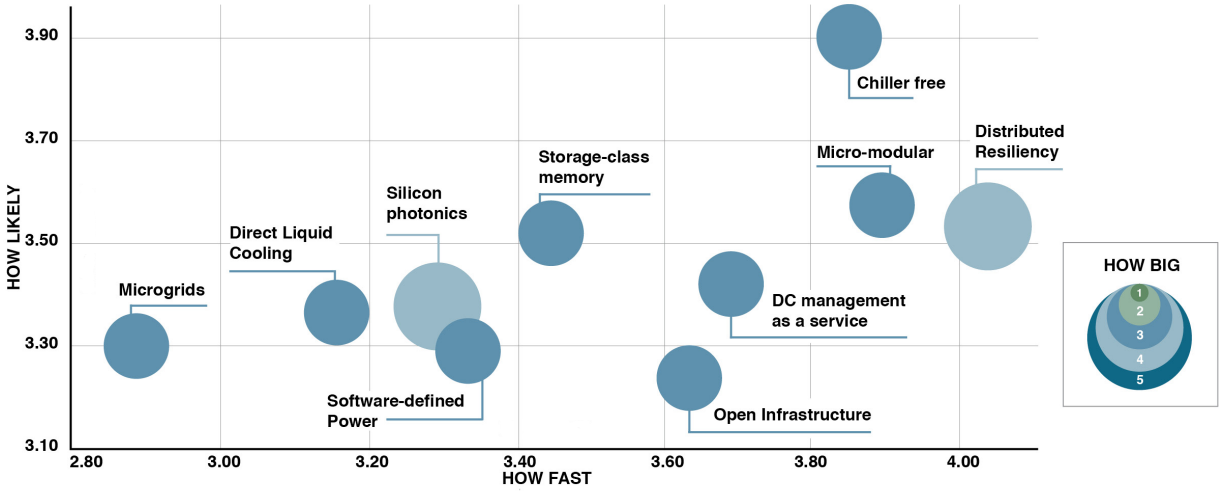
Data center operators (and enterprise IT) are generally cautious adopters of new technologies. Only a few (beyond hyperscale operators) try to gain a competitive advantage through their early use of technology. Rather, they have a strong preference toward technologies that are proven, reliable and well-supported. This reduces risks and costs, even if it means opportunities to jump ahead in efficiency, agility or functionality are missed.

But innovation does occur, and sometimes it comes in waves, perhaps triggered by the opportunity for a significant leap forward in efficiency, the sudden maturing of a technology, or some external catalyst. The threat of having to close critical data centers to move workloads to the public cloud may be one such driver; the need to operate a facility without staff during a weather event, or a pandemic crisis, may be another; the need to operate with far fewer carbon emissions may be yet another. Sometimes one new technology needs another to make it more economic.

The year 2021 may be one of those standouts in which a number of emerging technologies begin to gain traction. Among the technologies on the edge of wider adoption are:

- **Storage-class memory** - A long-awaited class of semiconductors with ramifications for server performance, storage strategies and power management.
- **Silicon photonics** - A way of connecting microchips that may revolutionize server and data center design.
- **ARM servers** - Low-powered compute engines that, after a decade of stuttering adoption, are now attracting attention.
- **Software-defined power** - A way to unleash and virtualize power assets in the data center.

All of these technologies are complementary; all have been much discussed, sampled and tested for several years, but so far with limited adoption. Three of these four were identified as highly promising technologies in the Uptime Institute/451 Research Disrupted Data Center research project summarized in the report **Disruptive Technologies in the Datacenter: 10 Technologies Driving a Wave of Change**, published in 2017. As the disruption profile in Figure 10 shows, these technologies were clustered to the left of the timeline, meaning they were, at that time, not yet ready for widespread adoption.



In 2017, Uptime Institute and 451 Research identified 10 technologies that had the potential to make a big, disruptive impact on the data center. Those to the right of the chart would do so sooner than those on the left, while those nearest the top would be most likely to have a disruptive impact. The size of the circle indicates how large an impact the technology would likely make.

Figure 10. **Disruption profile of emerging technologies**

Now the time may be coming, with hyperscale operators particularly interested in storage-class memory and silicon photonics. But small operators, too, are trying to solve new problems — to match the efficiency of their larger counterparts, and, in some cases, to deploy highly efficient, reliable, and powerful edge data centers.

Storage-class memory

Storage-class memory (SCM) is a generic label for emerging types of solid-state media that offer the same or similar performance as dynamic random access memory or static random access memory, but at lower cost and with far greater data capacities. By allowing servers to be fitted with larger memories, SCM promises to heavily boost processing speeds. SCM is also nonvolatile or persistent — it retains data even if power to the device is lost and promises greater application availability by allowing far faster restarts of servers after reboots and crashes.

SCM can be used not just as memory, but also as an alternative to flash for high-speed data storage. For data center operators, the

(widespread) use of SCM could reduce the need for redundant facility infrastructure, as well as promote higher-density server designs and more dynamic power management (see **Software-defined power**).

However, the continuing efforts to develop commercially viable SCM have faced major technical challenges. Currently only one SCM exists with the potential to be used widely in servers. That memory was jointly developed by Intel and Micron Technology, and is now called Optane by Intel, and 3D XPoint by Micron. Since 2017, it has powered storage drives made by Intel that, although far faster than flash equivalents, have enjoyed limited sales because of their high cost. More promisingly, Intel last year launched the first memory modules powered by Optane.

Software suppliers such as Oracle and SAP are changing the architecture of their databases to maximize the benefits of the SCM devices, and major cloud providers are offering services based on Optane used as memory. Meanwhile a second generation of Optane/3D XPoint is expected to ship soon, and by reducing prices is expected to be more widely used in storage drives.

Silicon photonics

Silicon photonics enables optical switching functions to be fabricated on silicon substrates. This means electronic and optical devices can be combined into a single connectivity/processing package, reducing transceiver/switching latency, costs, size and power consumption (by up to 40%). While this innovation has uses across the electronics world, data centers are expected to be the biggest market for the next decade.

In the data center, silicon photonics allows components (such as processors, memory, input/output [I/O]) that are traditionally packaged on one motherboard or within one server to be optically interconnected, and then spread across a data hall — or even far beyond. Effectively, it has the potential to turn a data center into one big computer, or for data centers to be built out in a less structured way, using software to interconnect disaggregated parts without loss of performance. The technology will support the development of more powerful supercomputers and may be used to support the creation of new local area networks at the edge. Networking switches using the technology can also save 40% on power and cooling (this adds up in large facilities, which can have up to 50,000 switches).

Acquisitions by Intel (Barefoot Networks), Cisco (Luxtera, Acacia Communications) and Nvidia (Mellanox Networking) signal a much closer integration between network switching and processors in the future. Hyperscale data center operators are the initial target market because the technology can combine with other innovations (as well as with Open Compute Project

rack and networking designs). As a result, we expect to see the construction of flexible, large-scale networks of devices in a more horizontal, disaggregated way.

ARM servers

The Intel x86 processor family is one of the building blocks of the internet age, of data centers and of cloud computing. Whether provided by Intel or a competitor such as Advanced Micro Devices, almost every server in every data center is built around this processor architecture. With its powerful (and power-hungry) cores, its use defines the motherboard and the server design and is the foundation of the software stack. Its use dictates technical standards, how workloads are processed and allocated, and how data centers are designed, powered and organized.

This hegemony may be about to break down. Servers based on the ARM processor design — the processors used in billions of mobile phones and other devices (and soon, in Apple MacBooks) — are now being used by Amazon Web Services (AWS) in its proprietary designs. Commercially available ARM systems offer dramatic price, performance and energy consumption improvements over current Intel x86 designs. When Nvidia announced its (proposed) $40 billion acquisition of ARM in early 2020, it identified the data center market as its main opportunity. The server market is currently worth $67 billion a year, according to market research company IDC (International Data Corporation).

Skeptics may point out that there have been many servers developed and offered using alternative, low-power and smaller processors, but none have been widely adopted to date. Hewlett Packard Enterprise's Moonshot server system, initially launched using low-powered Intel Atom processors, is the best known but, due to A variety of factors, market adoption has been low.

Will that change? The commitment to use ARM chips by Apple (currently for MacBooks) and AWS (for cloud servers) will make a big difference, as will the fact that even the world's most powerful supercomputer (as of mid-2020) uses an ARM Fujitsu microprocessor. But innovation may make the biggest difference. The UK-based company Bamboo Systems, for example, designed its system to support ARM servers from the ground up, with extra memory, connectivity and I/O processors at each core. It claims to save around 60% of the costs, 60% of the energy and 40% of the space when compared with a Dell x86 server configured for the same workload.

Software-defined power

In spite of its intuitive appeal and the apparent importance of the problems it addresses, the technology that has come to be known as "software-defined power" has to date received little uptake among operators. Software-defined power, also known as "smart

energy," is not one system or single technology but a broad umbrella term for technologies and systems that can be used to intelligently manage and allocate power and energy in the data center.

Software-defined power systems promise greater efficiency and use of capacity, more granular and dynamic control of power availability and redundancy, and greater real-time management of resource use. In some instances, it may reduce the amount of power that needs to be provisioned, and it may allow some energy storage to be sold back to the grid, safely and easily.

Software-defined power adopts some of the architectural designs and goals of software-defined networks, in that it virtualizes power switches as if they were network switches. The technology has three components: energy storage, usually lithium-ion (Li-ion) batteries; intelligently managed power switches or breakers; and, most importantly, management software that has been designed to automatically reconfigure and allocate power according to policies and conditions. (For a more detailed description, see our report **Smart energy in the data center**).

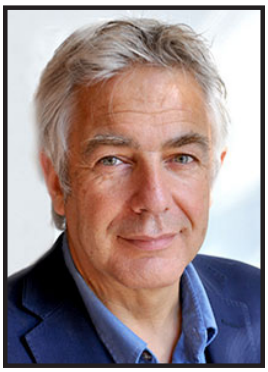
Software-defined power has taken a long time to break into the mainstream — and even 2021 is unlikely to be the breakthrough year. But a few factors are swinging in its favor. These include the widespread adoption of Li-ion batteries for UPSs, an important precondition; growing interest from the largest operators and the biggest suppliers (which have so far assessed technology, but viewed the market as unready); and, perhaps most importantly, an increasing understanding by application owners that they need to assess and categorize their workloads and services for differing resiliency levels. Once they have done that, software-defined power (and related smart energy technologies) will enable power availability to be applied more dynamically to the applications that need it, when they need it.

Innovation in the data center sector is experiencing a resurgence. With significant gains in efficiency and agility possible, more data center operators will be likely to adopt newer technologies.

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.

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

