Uptime Institute® INTELLIGENCE

PLANNING & STRATEGY

UI Intelligence report 24

Capacity planning in a complex, hybrid world

Capacity planning has become more difficult because of a range of technological developments, including virtualization, containers, converged infrastructure, and cloud services.

Authors
Rhonda Ascierto, Vice President of Research, Uptime Institute
Andy Lawrence, Executive Director of Research, Uptime Institute

UII-24 v1.1 published February 25, 2019 last updated April 2, 2019

This Uptime Institute Intelligence report covers:

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 https://uptimeinstitute.com/ui-research. For details about the primary Uptime Institute Intelligence survey data in this report, please see the Appendix.

Summary

Hybrid cloud and hybrid IT environments are bringing new complexities to capacity management and demand forecasting. Organizations are adopting various strategies, but the impact on demand/capacity can vary greatly. Managers and strategists seek to understand the capacity impact of choices available to them. Certain key technologies, such as public cloud services, virtualization, and new IT hardware, can play a crucial role.

While there is no one-size-fits-all approach, Uptime Institute Intelligence research suggests there are select areas worth integrating into any capacity management and forecasting strategy. Specific recommendations are listed at the end of this report.

KEY FINDINGS

• Public cloud is slowing demand growth for privately owned enterprise data centers, with nearly a fifth of enterprises saying it has had a major effect and 40% saying it had a lesser impact. Even so, the net impact on overall demand is currently minimal.

• Nearly half of enterprises report that overall demand for data center capacity is growing—both in their own sites and in colocation facilities. This stands in contrast to the widely held view that demand for non-public cloud capacity is shrinking rapidly.

• Nine of 10 enterprises (89%) say demand for data storage in their facilities is growing.

• In terms of technologies, virtualization has had the largest impact in reducing IT demand in the data center, followed by the use of public cloud computing and the deployment of more powerful servers.

• Most enterprises say they have no problem with virtual server sprawl and identify other benefits from the technology.

- **Nearly half of enterprises expect their next refresh of IT will enable them to reduce their numbers of physical servers. But there are mixed views on how this will affect demand for power and cooling.**
- **Roughly one-third of enterprises say their choice of IT hardware is causing power capacity concerns.**

Introduction

At a macro level, it is quite easy to forecast the demand for directly owned and managed data center capacity: Growth is slowing or flat and will likely go down, maybe slowly for a while, but sooner or later, much faster. Some analyst forecasts predict the eventual closure of most non-commercial sector data centers as the overwhelming majority of demand goes to public and private cloud services.

Uptime Institute's position (based on data from 451 Research) is somewhat nuanced (see below), but it does support the widely accepted view that there is a long-term, fundamental, and structural tilt toward ever-greater use of outsourced services in multi-tenant data centers, such as colocation facilities and outsourced cloud services, as well as a relative decline in enterprise data capacity demand and, therefore, investment. This is shown in Figure 1.

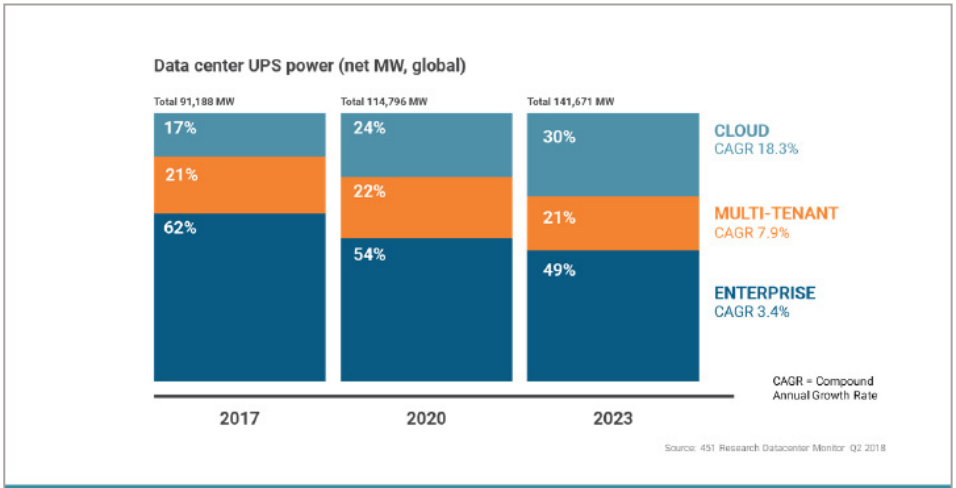


Figure 1. **Global data center growth projections**

Where our analysis differs from most, and what our data suggests, is this: We think it overwhelmingly likely that many large enterprises will be operating significant amounts of their own infrastructure (wherever it is sited) for at least another 15 years. Beyond this, technology and market shifts turn forecasting into guesswork, but there are good arguments for expecting that the infrastructure landscape will be large, diverse, fragmented, and complicated, with multiple types of ownership.

Feedback from data center operators around the world has been mixed for several years, with many reporting growing demand for capacity but

others dealing with low levels of facility and equipment utilization. This confused picture is partly because there are so many different drivers, business models, and types of data centers, and partly because of the changing technologies. At a macro level, some of the trends are:

- demand for all IT (and therefore data center capacity) is up, but cloud services are either driving or appropriating a big part of that;
- data center consolidation is cutting the enterprise footprint, but remaining data centers may consequently be larger and denser, and will need more power;
- technologies such as virtualization and, more recently, application containers, can push demand for power up for some, but down for others; and
- public cloud services may even have a counter-intuitive impact, attracting the newest applications but requiring local integration and back up, which drives local storage or non-cloud or private cloud requirements.

Senior business managers have frequently added to the confusion. Many companies have announced plans for a rapid transition away from data center ownership but later, and more quietly, realized the impracticality of their vision, at least in the near term, and reversed or postponed the move.

In Uptime Institute's 2018 annual survey of more than 250 data center designers, vendors, and consultants, capacity forecasting was identified as the number one challenge facing operators (see Figure 2).

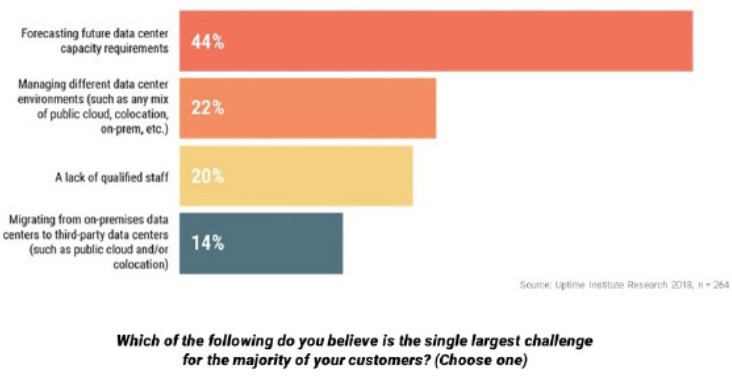


Figure 2. **Capacity forecasting is the top challenge operators face**

In this report, Uptime Institute Intelligence discusses recent research findings and the impact of some of the newer technologies on demand for power/space and offers some advice on the strategies that may reduce exposure. For more about our methodology and Uptime Institute Intelligence, see the Appendix.

Capacity Trends: Snapshot

While total global data center demand is shifting from enterprise-owned data centers to outsourced venues, it is clear that enterprise-owned facilities will not become obsolete and that leased colocation data centers are pivotal to many IT capacity strategies.

In our recent survey of more than 250 C-level executives and data center and IT managers at enterprises globally, over 75% said capacity for their owned data centers was either growing or flat—just 23% reported shrinking demand. The results were similar for capacity in their leased colocation data centers, an environment that enables enterprises to maintain ownership and control over their own IT (see Figure 3).

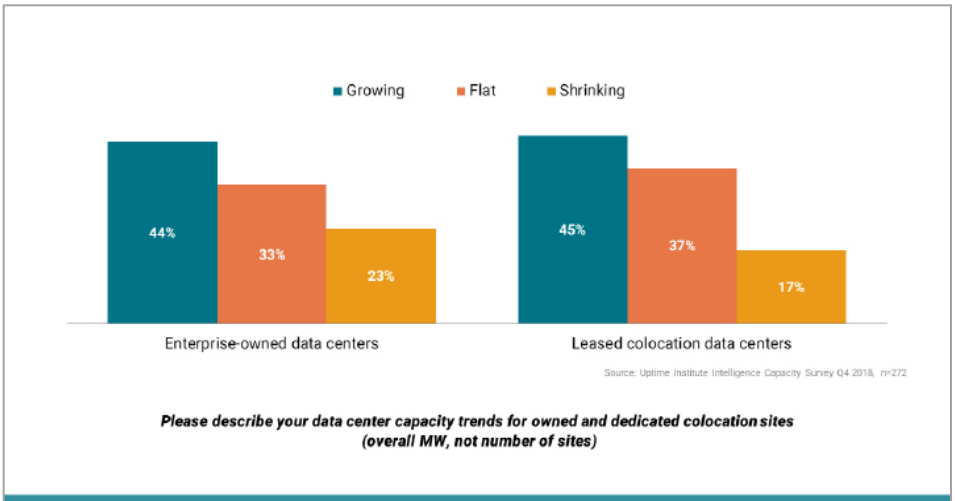


Figure 3. **Capacity at enterprise-owned and leased colocation data centers is growing**

We believe this is an important finding that is not widely understood across the cloud-focused IT industry. This finding should not be interpreted to mean there will be more enterprise data centers—it is likely consolidation will continue, and remaining data centers will be more efficient and dense.

Public cloud's limited impact

Only about one in five respondents reported that public cloud adoption has had a major impact in reducing demand at their owned 'on-premises' data centers. For most, public cloud adoption has resulted in a minor reduction in demand, with 30% saying it has had no impact whatsoever (see Figure 4).

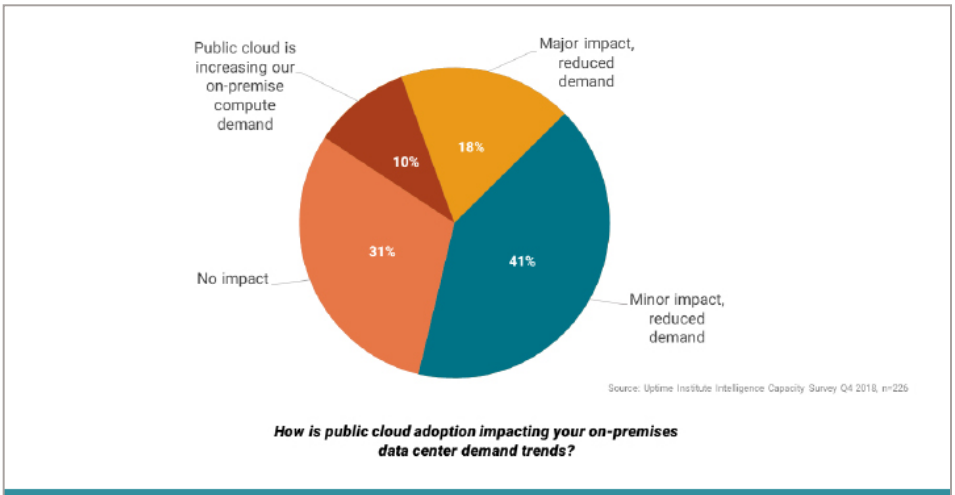


Figure 4. **Public cloud's net impact on enterprise data center demand is minimal**

Interestingly, for a small minority (10%) public cloud adoption is increasing their on-premises compute demand. One likely explanation is the widespread practice of using public cloud services for application test and development and, once complete, deploying those applications at scale ('in production') in an enterprise-owned or colocation data center for reasons of cost, control, data governance, security, and other factors. Another explanation is that some cloud applications can result in a need for more functions and data management by closely associated local, non-cloud or private cloud applications.

A mix of factors contribute to the limited impact of public cloud adoption on enterprise-owned and leased colocation data centers, including:

- the pace of demand growth for IT is exceeding outsourcing efforts
- the repatriation of workloads from public clouds to on-premises venues due to performance or availability issues, data sovereignty regulatory changes, higher-than-expected costs, and other reasons
- the immaturity of organizations' public cloud strategy; deployments to public clouds are currently limited because of organizational barriers, unsuitable workload characteristics, costs, etc.
- the creation of demand, both locally and in a public cloud, by conjoined public and non-cloud or private cloud applications, offsetting the movement out of enterprise's on-premises sites.

Most organizations are taking a hybrid approach by integrating off-premises and on-premises cloud environments and/or running different workloads in a mix of different venues based on governance, cost, and other requirements.

Public cloud providers are responding to the hybrid trend with offerings designed to extend the public cloud to on-premises data centers (enterprise-owned or leased).

Amazon Web Services' (AWS) forthcoming Outposts product is an example of this. It is a fully managed hardware rack of AWS-branded IT that can be deployed in customers' on-premises or leased data centers (running either AWS native or VMware Cloud on AWS—outside of an organization's firewall). Microsoft's Azure Stack and Google's Cloud Services Platform (GKE On-Prem) are similar, except they are designed to run on industry standard third-party IT hardware.

All extend public cloud to an on-premises data center (be it privately owned or leased) by simplifying the management of public-cloud workloads running in different venues. Enterprises can access and move public-cloud workloads to and from service provider environments and their own, managed by centralized software. These products are also expected to better position public cloud services running in distributed, edge environments, a market development that is still in the early stages.

Storage growth

One of the most acute and common capacity challenges at many data centers, and one that bears out in our research, is the significant and growing demand for data storage.

More organizations are using public cloud storage services, including data backup and recovery, for improved capacity scalability and flexible consumption. Yet the growth of storage continues to plague many. Nearly half of respondents in our survey said they had experienced a major uptick in storage demand at their enterprise-owned or colocation data center and that this uptick was expected by their department. Storage increases in these environments are widespread. Just 11% of respondents said there had been no change or a decrease in storage workload, as shown in Figure 5 (below).

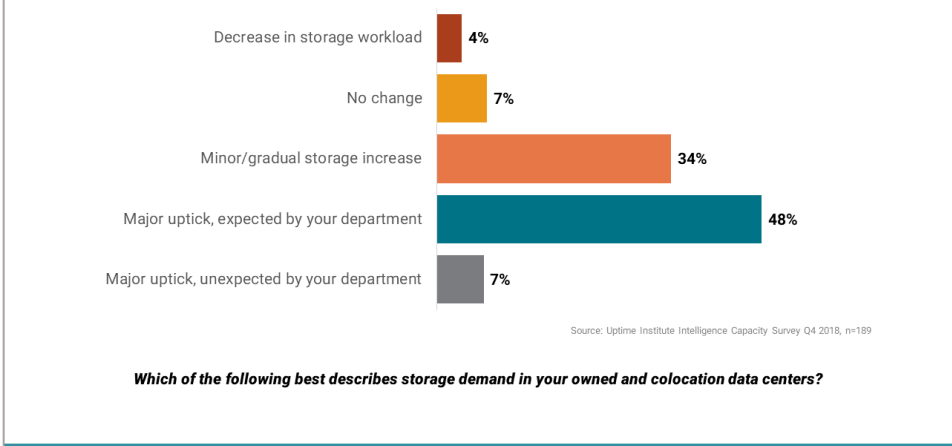


Figure 5. **Storage demand is growing in both enterprise-owned and colocation data centers**

Capacity management and DCIM

The most common tool for tracking, managing, and forecasting data center capacity is data center infrastructure management (DCIM) software. After many years of market hype and under-developed products, the technology has matured and reached mainstream levels of adoption. In our 2018 annual survey of IT and data center managers globally, a majority of respondents said they had deployed some type of DCIM, and typically their implementation had been successful. The most commonly reported motivation for deploying DCIM was capacity planning (75% of respondents), followed closely by power monitoring (74%).

Leading DCIM tools can model and forecast capacity usage, as well as changes to resource utilization versus total available capacity. They can also model changes in utilization of power, space, cooling, and port connectivity if, say, a certain number of servers are added. Importantly, they can predict when and how a data center will reach its capacity limitations (network ports or power or space, for example).

Capacity growth is typically modeled on historic data for power, power panels, floor space, enclosure positions, cooling, copper and fiber ports, and other resources. Capacity thresholds can be applied, as can costs (actual and forecasted). Users can blend various data points to determine thresholds and to help balance capacity across power, cooling, connectivity, and space. (These capabilities are also possible for those that do not have historical data by manually entering expected spikes and troughs, although the accuracy of this approach can be highly variable.)

DCIM is considered the most powerful tool to gain visibility into data center capacity trends (at the rack, row, room, data center, and portfolio level) and to forecast future capacity requirements. It can also identify stranded capacity and provide recommendations to drive up utilization rates and to better manage capacity demand. However, these capabilities are typically possible only in a mature implementation of leading

DCIM software. While most data centers have deployed some type of DCIM, relatively few have reached a level of deployment maturity that enables them to effectively manage—or even reduce—demand for physical resources.

Reducing capacity demand

If public cloud adoption is having a limited impact on reducing on-premises data center capacity, including storage—one of the fastest-growing areas of IT—what other factors are at play?

We asked managers about three of the most common factors: public cloud computing, server virtualization, and advancing server processor capabilities. The single most-important approach, according to our survey, was server virtualization (see Figure 6).

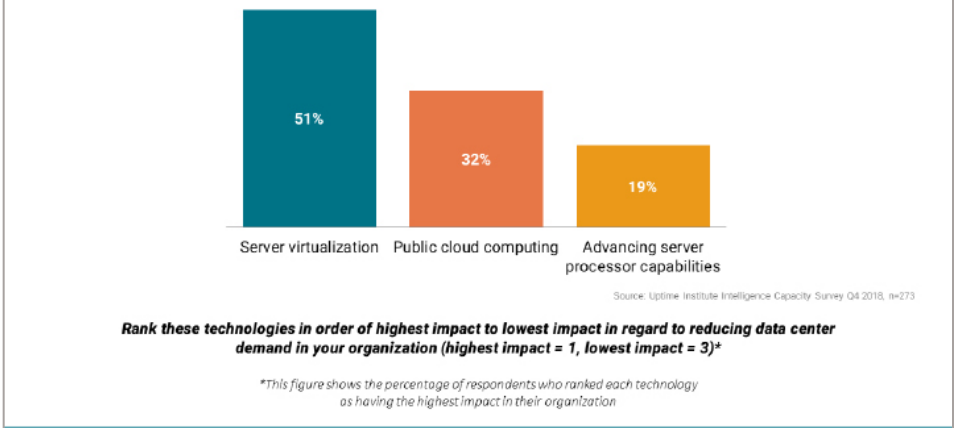


Figure 6. **Virtualization has the greatest impact on capacity demand**

This finding offers promise to managers struggling with a combination of capacity and even performance concerns. Virtualization, "cloud readiness," and hardware refresh are initiatives often taken together; each reaps benefits that, while supportive of each other, do not require a commitment to the public cloud until, or if, the organization is ready.

Server Virtualization Impacts

Server virtualization is keeping capacity growth demands in check in enterprise-owned and colocation data centers for half of the respondents to our survey, as shown in Figure 7. A further 40% said virtual machine (VM) compression, increasing the number of VMs per host server, is reducing capacity demand. Yet for some (one-third of respondents), capacity gains have been short-lived: virtualization helped initially but is no longer a factor in reducing capacity demand.

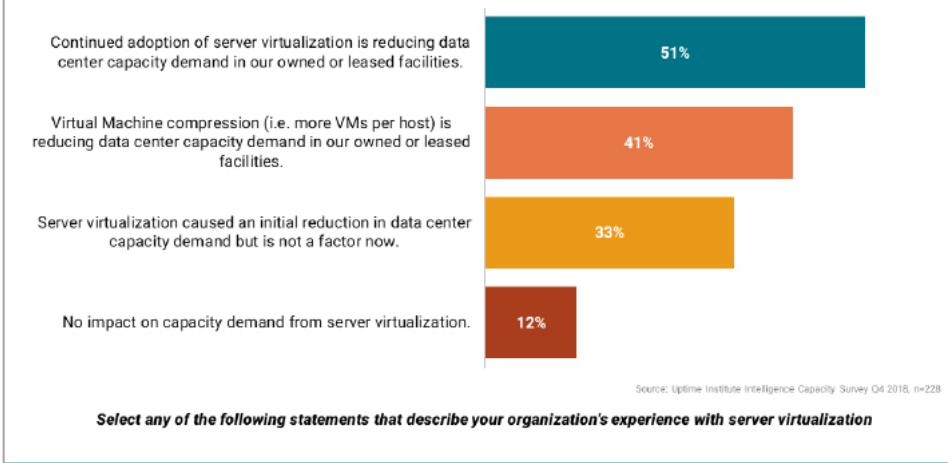


Figure 7. **Virtualization reduces data center capacity demand**

This is not to suggest that organizations are adopting virtualization with a primary goal of reducing capacity demand. They do so for greater workload agility and automation, including to simplify workload migrations and load shifting and to deploy private clouds layered on top of virtualized environments, among other reasons.

The benefits of virtualization are extensive even for those experiencing short-lived capacity gains. For example, 60% of respondents said virtualization has enabled them to implement self-service provisioning—and most (77%) have done so successfully without negative capacity impacts, such as VM sprawl and disorganization.

It is clear that virtualization, now a mature technology, is proving beneficial to many, including the capacity challenged.

The promise of application containers

More organizations are adopting a method of virtualization known as application containers ('containers'), the most common of which is Docker (although there are many others).

Unlike VMs, a container consists of an entire runtime environment: an application plus the supporting dependent software, libraries, and configuration files. In other words, containers do not require a dedicated, pre-provisioned support environment (notably, a dedicated operating system) and, therefore, will usually require less compute and memory capacity.

In our survey, just 23% of respondents are using containers, with an additional 18% in discovery/proof-of-concept trials. As shown in Figure 8, about one-quarter of those using or considering containers expect the technology to reduce their physical server footprint, while 35% expect that increases will be offset by capacity growth.

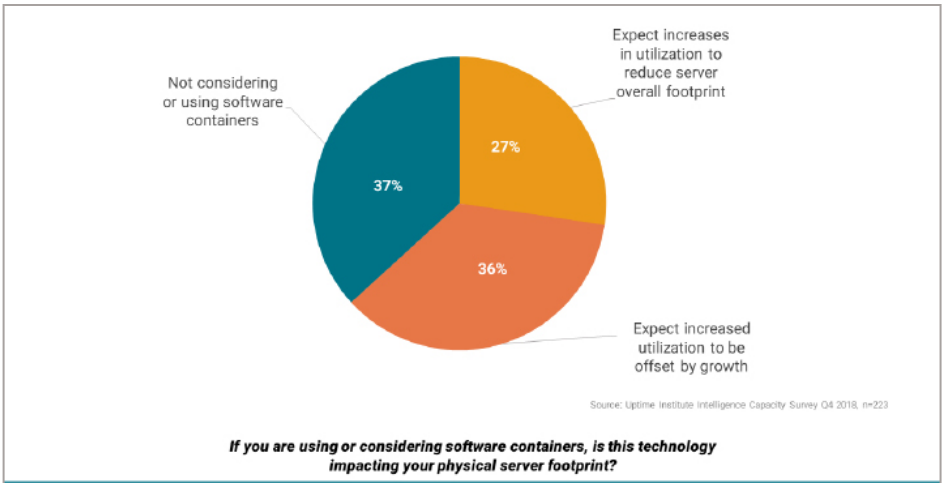


Figure 8. **Enterprises using or considering containers expect limited capacity benefits**

Containers can run on various infrastructure: bare-metal servers, traditional server environments, virtual environments, or any type of cloud (public, private, or hybrid). Today, most containers are deployed within VMs, which requires maintaining the same software infrastructure that was in place to run the VMs initially, thus mitigating the capacity gains containers promise.

However, there is growing momentum around bare-metal container deployments, an approach that enables organizations to exploit containers' lightweight footprint. As adoption of bare-metal containers grows, we believe containers will emerge as an important factor in managing growing on-premises capacity (and have heard anecdotal evidence of this from some operators running large bare-metal containerized cloud environments).

Power Density, Utilization, and Server Technologies

Available IT throughput and capacity can be increased by simply increasing server rack density and server utilization, or so the logic goes—assuming the power is available. Research by Uptime Institute and other organizations suggests that most servers are under-utilized, despite virtualization. While this can be a deliberate strategy (to deal with peaks and demand growth), it is often the simple result of Moore's Law—the hardware capabilities have outstripped the application's needs. Initiatives to drive up IT utilization (to reduce stranded compute) are underway at many organizations.

Power, cooling capacity, and other supporting infrastructure limitations can stymie efforts to increase rack densities, which remain low:

- The average rack density for more than two-thirds of respondents in our 2017 annual survey was less that 6 kW (kilowatt) per rack.
- In our 2018 annual survey we asked about the highest server density deployed; for about one-third of respondents it was less than 10 kW per rack, and for 30% it was 10-19 kW per rack.

IT hardware refreshes

Our research shows that the majority of data centers plan to keep pace with Moore's Law: 85% of survey respondents are expecting to upgrade to the latest generation of x86 processers in their next server refresh. Nearly half believe their next server refresh will enable them to reduce their numbers of physical servers, as shown in Figure 9.

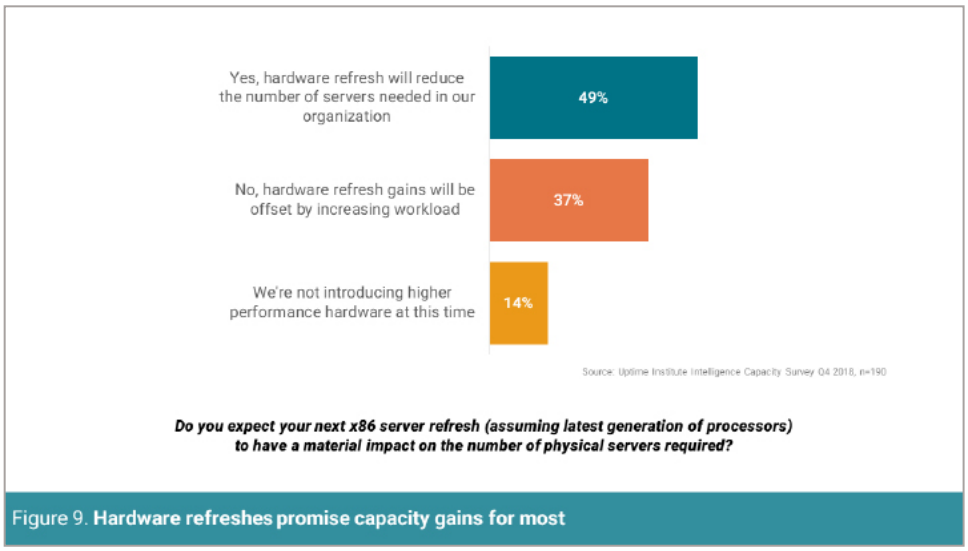


Figure 9. **Hardware refreshes promise capacity gains for most**

There are mixed views on how this will affect demand for power and cooling, as shown in Figure 10. A small majority (36%) are experiencing or expecting higher power and cooling density requirements, which they can accommodate, while a slightly smaller portion (31%) believe future server deployments will have no effect. This is highly dependent on the utilization of the servers, as well as the particular workloads.
For nearly one-quarter of respondents (23%), future IT hardware deployments are expected to draw less power and cooling capacity, providing a measure of control for managing capacity demand. However, capacity gains from x86 server refreshes may diminish over time, with the inevitable slowdown of Moore's Law.

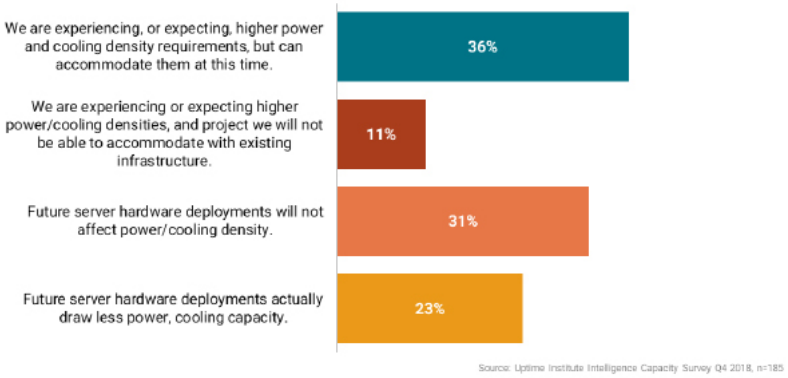


Figure 10. **The impact of server refreshes on power and cooling: more for some, less for others**

Hardware deployment choices

The prevalence of pre-integrated IT hardware is catching up to that of traditional IT deployment approaches. In our survey, we asked respondents for all the ways they are deploying hardware (which means they could choose multiple approaches); 62% of respondents are installing servers, storage, and networking components individually (traditional approach) while 70% are installing some form of pre-integrated IT:

- 32% are installing converged infrastructure (pre-integrated cabinets); and
- 38% are installing hyper-converged infrastructure (pre-integrated clustered virtualized servers, storage, and network as a single IT appliance).

Enterprises typically deploy converged infrastructure for ease and speed of procurement and deployment. A converged approach is becoming de facto for small and distributed IT environments, such as remote/branch offices and closets, where on-site IT expertise is often lacking.

Hyper-converged infrastructure (HCI) offers similar advantages, yet, even in large organizations, their deployments are mostly experimental today. HCI is often deployed beside traditional standalone servers and storage in various types of data centers. (HCI is also a prime candidate for on-premises cloud deployments, mostly for private cloud [versus hybrid cloud].) Many organizations experimenting with HCI will likely expand their footprints over time.

What impact are pre-integrated hardware choices having on capacity?

The configuration of components in pre-integrated hardware are pre-set and this can potentially prohibit very high utilization rates of the physical hardware. In contrast, the traditional approach of installing and configuring server, storage, and networking components individually can, when done well, enable higher utilization of all components (compared with pre-integrated). However, higher utilization outcomes with traditional approaches depend on the availability of skilled staff who can rack and optimally configure IT environments. In other words, pre-integrated infrastructure has a pre-determined limit on the level of utilization possible while the limits of traditional approaches depend on how the hardware is manually deployed.

In our survey, nearly one-third of respondents said their choice of IT hardware is causing power capacity concerns. More than one-quarter said cooling capacity was strained because of hardware deployment choices, and one in five are experiencing issues with networking (see Figure 11).

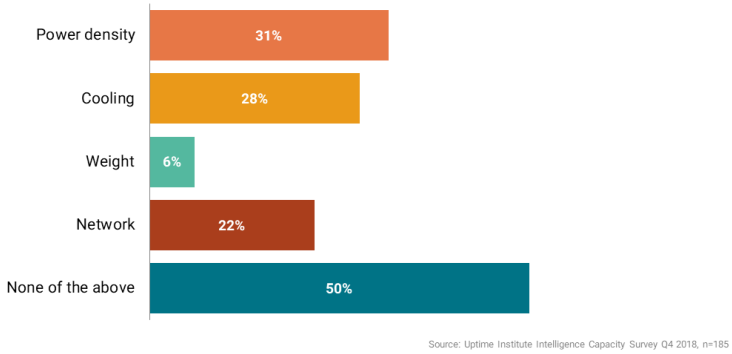


Figure 11. **Hardware choices push capacity limits for some**

When taken together, these data suggest that many enterprises are using a mix of traditional and pre-integrated infrastructures and that the effect on capacity demand is mixed. For those who choose less pre-integration, active management is an important key to higher utilization and better capacity management.

Conclusions and Recommendations

Capacity management is becoming more complex as more organizations adopt hybrid cloud and hybrid IT approaches. Public cloud computing is helping, but server virtualization is proving the most effective in reducing capacity demand for many (and the adoption of application containers promises further gains in the future). Server refresh cycles are providing capacity demand relief for some, although hardware deployment choices are having no obvious or clear impact.

For effective capacity management and forecasting, Uptime Institute Intelligence recommends that managers/operators do the following:

- Create standardized capacity management and operational processes and ongoing assessment mechanisms per IT venue/location.
- When considering new outsourcing approaches to off-premises capacity, plan to do so in incremental steps. Understand that public cloud can both reduce and drive demand in existing data centers.
- Create capacity KPIs (key performance indicators) that focus on business and financial—rather than just infrastructure—metrics. This should help keep IT expenditure aligned with overall business requirements and enable greater leverage/proactive education during discussions with upper management.
- Develop/execute a DCIM software strategy that includes both monitoring (power and environmental) and asset management, with a goal of modelling and forecasting capacity usage on an ongoing basis, as well as changes to resource utilization versus total available capacity. Capacity management planning should integrate IT and data center capacity models.
- Investigate a cross-disciplinary plan (involving both data center facilities and IT departments) for the evaluation/testing of bare-metal application containers.
- Evaluate the return on investment of server refreshes to include the value from additional throughput per kilowatt with the latest generation of processor technology.
- Incorporate the potential capacity and utilization benefits and limitations when assessing hardware deployment strategies, including that of pre-integrated infrastructure (such as converged and hyperconverged).

Appendix

Most of the data in this report is from Uptime Institute Intelligence's November 2018 Capacity Survey of data center and IT managers globally. The survey respondents are end-users—people responsible for managing IT and IT infrastructure—as well as C-level executives, at some of the world's largest IT organizations. As shown in Figure A1, the participants represent a wide range of industries and different geographical regions.

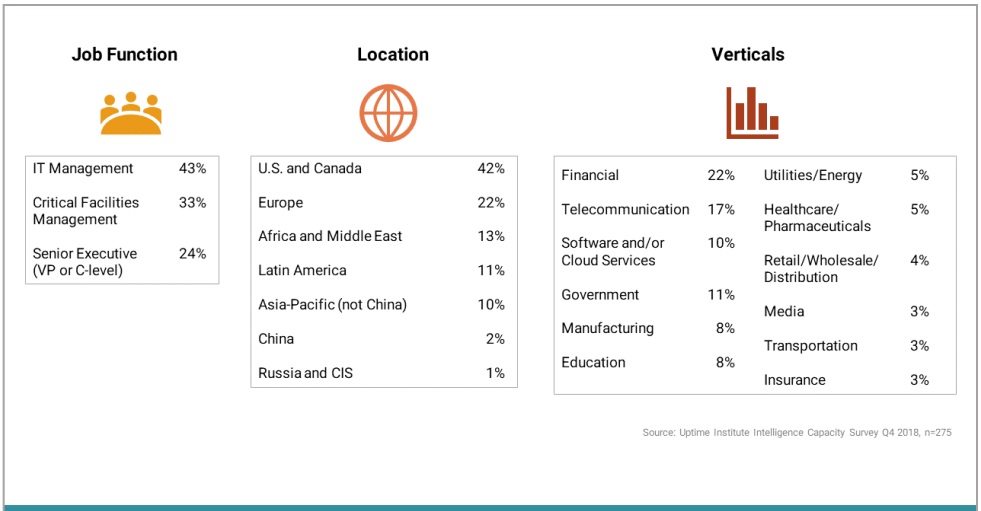
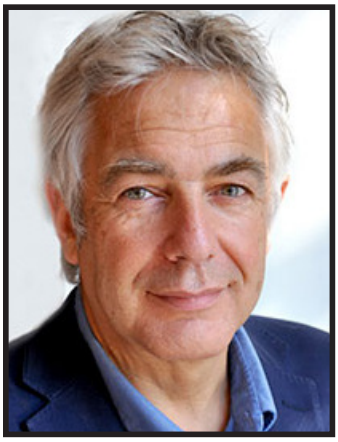


Figure A1. **Respondent demographics, Uptime Institute Capacity Survey, Q4 2018**

Uptime Institute® | INTELLIGENCE

ABOUT THE AUTHORS

Rhonda Ascierto is VP of Research at the Uptime Institute. She has spent nearly 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

Uptime Institute is an unbiased 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 U.S., Mexico, Costa Rica, Brazil, U.K., Spain, U.A.E., Russia, Taiwan, Singapore, and Malaysia.
Visit www.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

